Histograms

An educational eye

LONNEKE BOELS

Histograms

An educational eye

Lonneke Boels

L. B. M. M. Boels

Histograms - An educational eye

Histograms
An educational eye

Histogrammen
Een educatieve kijk

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 20 september 2023 des middags te 4.15 uur

door

Léone Bernadette Martha Maria Boels

geboren op 30 april 1966

te Heerlen

Promotoren:

Prof. dr. P.H.M. Drijvers

Prof. dr. W. Van Dooren

Copromotor:

Dr. A. Bakker

Beoordelingscommissie:

Dr. G.F. Burrill

Prof. dr. J.T. Jeuring

Prof. dr. S.E. McKenney

Prof. dr. M.F. van der Schaaf

Prof. dr. W.R. van Joolingen

Table of Contents

1

The role of histograms in developing statistical literacy

"Ignite the mind's spark to rise the sun in you." [1]
Attributed to Florence Nightingale

"Alone we can do so little; together we can do so much." [2]
Helen Keller

---

[1] 50 Florence Nightingale Quotes, NURSING.com, https://blog.nursing.com/florence-nightingale-quotes
[2] Garson O'Toole (2014). Quote Investigator, *Alone we can do so little; together we can do so much*. https://quoteinvestigator.com/2014/04/21/together/

Chapter 1

1.1   The histogram as a spider in a web of knowledge

1.1.1   The relevance of graphs in statistics

A correct use of statistics can literally save people's lives. A famous example is from the first female statistician Florence Nightingale who saved many lives with her polar graph (Martineau, 1859), which shows that more soldiers died from preventable diseases—caused by bad hygienic circumstances in hospitals as well as a lack of beds and blankets—than from the wounds caused by the Crimean War (1853–1856). The graph on the right (Figure 1.1) shows the data when Nightingale started her data collection with each circle section indicating one month. As can be seen from the graph on the left, providing beds, blankets, and clean pottery dramatically reduced the number of preventable deaths indicated by the blue areas.

Figure 1.1 Nightingale's famous polar graph (1858) with causes of deaths in the British Army
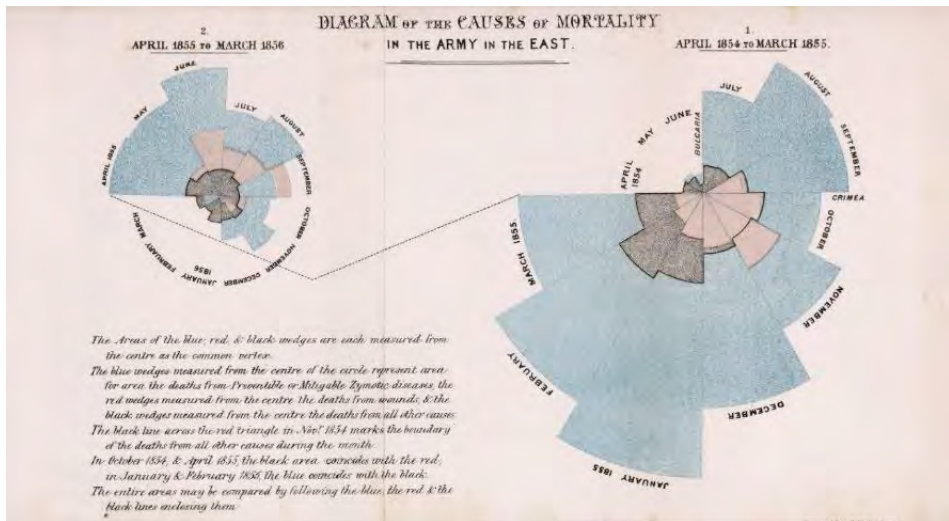
Note. Light red areas indicate the number of people who died from bullets, blue areas indicate the number of people who died from preventable diseases such as cholera, dysentery, frostbite, and typhoid. Black areas indicate other causes. Source: Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg). CC-PD and PD-US-expired.

In descriptive statistics, data are often summarized in numbers, such as the arithmetic mean and the standard deviation or a confidence interval, rather than in graphs. Many studies, including studies on educational improvement, report on descriptive statistics. Several examples, however, show that such summary statistics provide limited information, as different data distributions —depicted in graphs—can lead to the same descriptive statistics and vice versa

(e.g., Anscombe, 1973; Pastore et al., 2017). One example is the Datasaurus (Matejka & Fitzmaurice, 2017; based on the Datasaurus dataset created by Alberto Cairo), see Figure 1.2. Mean, standard deviation, and correlation are the same in both graphs, but as the reader can see, the pattern in the data is completely different. Yet it was not just one other graph they created. In total, besides the Datasaurus itself, they constructed twelve completely different graphs with the same mean, standard deviation, and correlation. This example highlights that graphs show "quantitative and qualitative information, so that a viewer can see patterns, trends or anomalies, constancy or variation, in ways that other forms—text and tables—do not allow" (Friendly, 2008, p. 502). According to Friendly, Galton made several important scientific discoveries through graphing data (e.g., the idea that barometric pressure and wind direction are related).

Figure 1.2 Datasaurus and another scatter plot with the same mean, standard deviation, and correlation to two decimal places


```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```
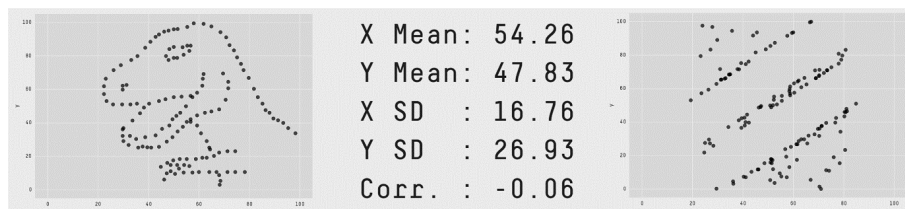
Note. Permission for reprinting granted by Justin Matejka, December 27, 2022. *(*Source*: https://www.autodesk.com/research/publications/same-stats-different-graphs)

The Datasaurus example is humorous. Unfortunately, the misuse of statistics—in the following example combined with probability—can also destroy lives. In 2003, the Dutch nurse Lucia de Berk was sentenced to life in prison for several alleged murders of patients. In 2010, De Berk was acquitted because the conviction was a judicial error, based on "flawed data collection" and "using an over-simplified discrete [hypergeometric] probability model" that did not include "the variation among nurses in incidents they experience during their shifts" (Gill et al., 2018, p. 9). Instead, Gill et al. used another model—the Poisson process. "Since we believe the incidents to be rare, a Poisson process is an obvious choice for modeling the incidents that a nurse experiences." (p. 11). Recently, in a similar case, Daniela Poggiali—an Italian nurse—was accused of murder based on flawed statistics (Gill, 2022) and acquitted in 2021. Gill believes many nurses around the world that are accused of murder are most likely innocent (e.g., Ben Geen and Lucy Letby, UK). For both Lucia and Daniela, Gill and colleagues did a lot of demanding work to clean and depict the data. They found interesting and explainable patterns by simply

graphing the data, patterns such as patients tending to die mostly at half and whole hours in Italy (Gill, 2022). This search for general patterns in hospital data by graphing them had been omitted in the nurses' first trial.

Every day, vast amounts of data are collected. A graphical representation suitable for representing large amounts of data of one statistical variable (also called univariate data) is a histogram. As an example, consider the two histograms depicting the age of Dutch first-time mothers in 1950 and 2021 respectively (Figure 1.3). One advantage of histograms is that they reveal a distribution of the data. They can show patterns in data better than bar charts depicting the mean and standard deviation—the latter being nothing more than a pretty ornament for these two numbers, to paraphrase Lee (1999) or a substitute for a table (Tukey, 1972).

Figure 1.3 Two histograms depicting the age of mothers giving birth to their first child
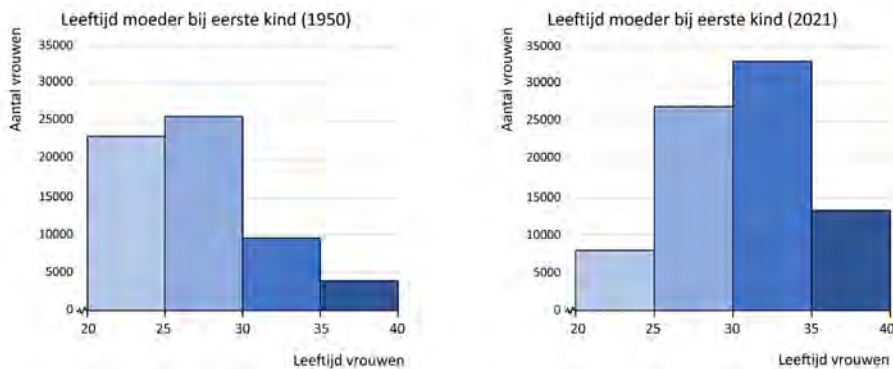
Note. Source: CBS (2022).

Although boxplots and dotplots can similarly disclose a distribution, in some cases histograms do a better job. Pastore et al. (2017) conclude "that appropriate graphical representations can increase reliability in research findings and promote transparency in the way scientific information is shared and disseminated" (p. 2). A second advantage of histograms is that they seem easier than, for example, boxplots (e.g., Bakker et al., 2004; Lem et al., 2013b, 2013c, 2015), although there are some examples of introducing boxplots via hatplots (Konold, 2002) with some degree of success (e.g., Allmond & Makar, 2014; Makar & Confrey, 2003; Saldanha & Hatfield, 2021). In addition, histograms can support both proportion-based and quantile-based reasoning, whereas boxplots only support the latter (cf. Frischemeier et al., 2023). While non-stacked ('messy') dotplots support quantile-based reasoning, proportion-based reasoning with them is more difficult than with histograms. Proportioned-based reasoning with stacked dotplots is similar to histograms. However, the difficulties students encounter with stacked dotplots are similar

to histograms (e.g., Lem et al., 2013a; Lyford, 2017). A disadvantage of histograms (and stacked dotplots) is that the binning process can influence how the shape of the graph appears (e.g., Sahann et al., 2021; Setlur et al., 2022).

1.1.2   What is a histogram?

A histogram is an important graph in research, education, and in media—for example when reporting about the COVID-19 pandemic. A histogram displays a graph with bars that can depict large amounts of numerical data. Although histograms do a better job than descriptive statistics in giving a first impression of data and the patterns in them, they also lead to several misinterpretations. For example, students confuse histograms with look-alikes (Box 1), including case-value plots (Cooper, 2018; Cooper & Shore, 2008, 2010). Before we elaborate on these difficulties in the educational section, we first consider the following question:

Which of the following two statements about the graphs in Figure 1.4 is true? Are the *arithmetic mean* and *variability* in weight higher in the graph on the left, the right, or are they approximately the same for both graphs?

Figure 1.4 A histogram (left) with one statistical variable (weight) and a case-value plot (right) with two statistical variables (given name and weight)
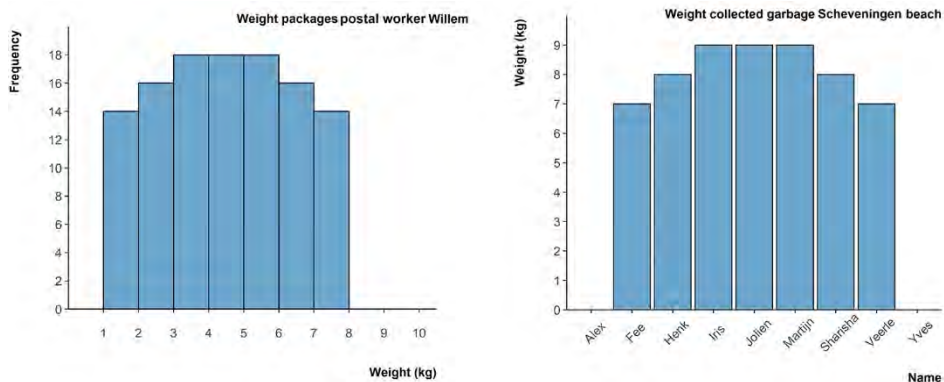
Before answering this question, let us discuss the data depicted in both graphs. Each bar in the histogram (Figure 1.4, left) indicates how many packages there are in that interval (e.g., there are 14 packages with a weight between 1 to 2 kg). Hence, the mean weight of all those packages in the left-hand graph can be read on the horizontal axis and is approximately 4.5 kilograms. For the case-value plot (Figure 1.4, right), nine students were supposed to collect garbage on the beach. Two students handed in zero weight. The rest of the students collected between 7 and 9 kilograms of beach waste. The arithmetic mean weight of beach garbage picked up per student was 57 : 9, which is about 6.3

kilograms. Hence, the arithmetic mean is higher in the case-value plot—the right-hand graph. In the histogram on the left, one might first think that postal worker Willem delivers 7 packages. But if all packages are depicted in a dotplot (Figure 1.5, left), it becomes clear that he delivers 114 packages in total.

Figure 1.5 Dotplots depicting the same weights as in Figure 1.4 for all packages that postal worker Willem delivers (left) and the collected garbage on Scheveningen beach (right)
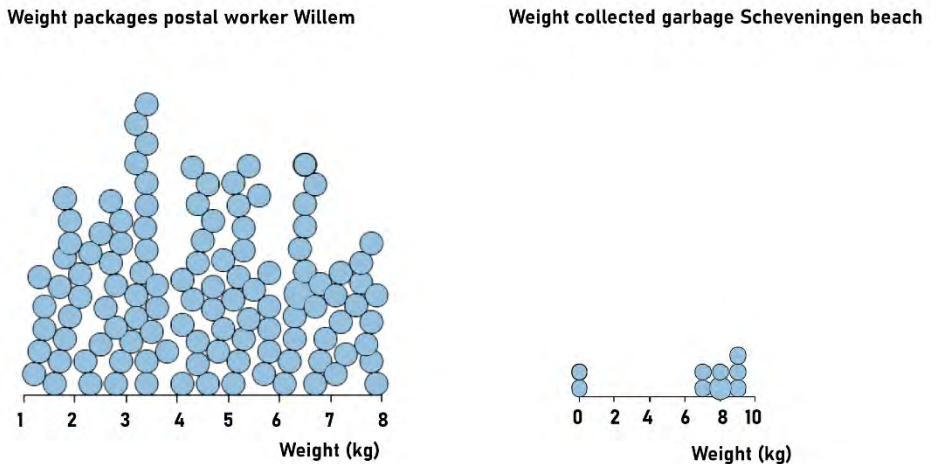
Note. Both dotplots (graph areas) were constructed with VUstat (https://www.vustat.eu).

Some readers may have thought that the mean weight in the left-hand graph was about 16.3 (sum of frequencies divided by seven) or 12.7 kilograms (divided by nine). Like many students, they were possibly confused by the visual similarity to the type of graph on the right (e.g., Bakker, 2004a; Chance et al., 2004; delMas & Liu, 2005) and estimated the mean frequency, instead of the mean weight.

To assess the variability in both graphs, the standard deviation from the mean can be used, which is approximately 1.9 for the histogram on the left and 3.7 for the case-value plot on the right. Hence, the variation is higher in the right-hand graph. This might seem counter-intuitive because the collected weights seem to vary between 7 and 9. However, two students collected zero kilograms of garbage. The weight, therefore, varies between 0 and 9 kilograms in the right-hand graph, compared to 1 to 8 kilograms in the left-hand graph. Outliers can have a huge influence on both the mean and the standard deviation, especially when the dataset is very small (see also Figure 1.5, right). Note that a case-value plot is a graph where each bar represents a measurement of one case. Typically, the horizontal axis depicts a variable

measured at a nominal or ordinal measurement level, and the vertical one at an interval or ratio measurement level[3].

Before it is defined what exactly a histogram is, les us first consider another, more realistic example of data that can be presented in a histogram. With the previous example of a histogram, an attempt was made to give the reader a sense of how difficult it can be for students to interpret graphs. However, we also know from research that graphs of data with which people are familiar, such as American SAT scores from college entrance exams in the USA (Kaplan et al., 2014) are easier to understand. Therefore, in Table 1.1, a few of the 826,192 *reported* infections of COVID-19 in 2020 in the Netherlands are presented (RIVM, 2022). From this table, we take *one* column: *Age group* (see also Box 2)*.* Each row is one person, so in these five rows, we see that there is one person aged 40–49 who got COVID-19, one person aged 50–59, and so on. A dotplot from the original data (using a fictive age instead of an age group) *could* look like Figure 1.6. As an illustration, we present a small subset of cases in the dotplot. From these data, a histogram can also be created by binning age groups in bins of, for example, 10 years (Figure 1.7).

Table 1.1 Example of a part of a data table for COVID-19 infections in the Netherlands in 2020

Date statistics	Date statistics type	Age group	Fictive age
1/1/2020	DOO	40-49	45
1/1/2020	DOO	50-59	53
1/1/2020	DOO	20-29	21
1/1/2020	DOO	60-69	62
1/4/2020	DOO	10-19	16

Note. Source: RIVM, 2022. DOO = Date of (disease) onset. It is not always known whether this first day of illness already involved COVID-19. Fictive age is a variable not present in the original dataset. It was created using the *=RANDBETWEEN(a;b)* function in Microsoft Excel where a and b are the borders of the bin. For example, *=RANDBETWEEN(40;49)* returns a random whole number from 40 to 49. This number was added to make a dotplot for these data.

---

[3] Measurement level refers to the scale used for the measurements of the statistical variable. An example is the variable temperature that can be measured at an ordinal level (e.g., cold, warm, hot), interval level (e.g., Celsius or Fahrenheit with an arbitrary zero point), or ratio level (Kelvin; absolute zero point. When temperature in Kelvin is doubled, thermal energy is also doubled). Besides the four measurement levels mentioned here, other scales do exist (e.g., cyclic for angles).

Figure 1.6 Possible dotplot for a subset of COVID-19 infections in the Netherlands in 2020 (fictive age guessed from the age group through a random function in Excel)
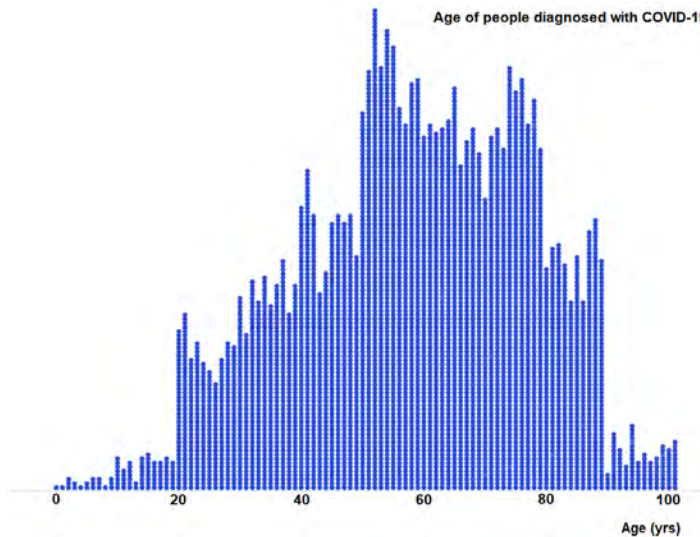
Note. Made with Codap https://codap.concord.org/. For our aim, we can ignore the high spikes in the graphs that are partly due to the random function in Excel, and rounding to whole numbers for age.

Figure 1.7 Histogram for all 826,192 COVID-19 infections in NL in 2020
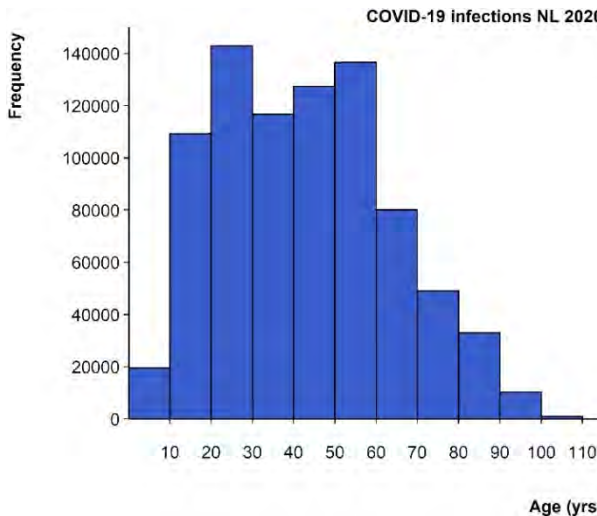
Note. We removed all approximately hundred cases of people aged 0–49 who died due to COVID, as the RIVM removed their specific age group to prevent their identification based on the data. In addition, we removed about fifty people with unknown ages. Given the vertical scale in this histogram, this did not influence the graph's appearance. Note that this graph is *not corrected* for the total number of people in each age group (in which case it would no longer be a histogram).

For large datasets like COVID-19 infections in the Netherlands with over 800,000 cases reported in 2020, the advantage of creating histograms comes to the forefront. Instead of depicting 826,093 dots, now only 10 bars are needed to represent this data set.

Now that the histogram is informally introduced, let us define a histogram as this is rarely done (Humphrey et al., 2014). In many textbooks and online sources, incomplete definitions are given such as "A histogram represents numbers by area, not height." (Freedman et al., 1978, p. 26), "A histogram is a bar graph [bar chart] of a frequency distribution with measurable data on the horizontal axis." (Getal en Ruimte, 2014, p. 14) or only by most noticeable but irrelevant features such as a bar chart "with bars […] that touch" (e.g., Nijdam, 2003, pp. 49–50), often followed by a description of how to construct a histogram (e.g., on Wikipedia, December 11, 2022).

Definition[4]

A histogram is a graph with bars that meets the following criteria:
- It consists of bars that represent groups of numerical data
- It represents data of one statistical variable only (typically continuous)
- The statistical variable is presented conventionally on the horizontal axis[5]
- The statistical variable is measured at an interval or ratio measurement level
- The vertical axis displays the class density, or—when bin widths are equal—relative or absolute frequency (counts) of the depicted statistical variable
- The total density adds up to 1 or the total relative frequency adds up to 100%

The histograms above were all with equal bin widths (e.g., Figure 1.7). In Figure 1.8, an example of a histogram with density along the vertical axis can be found for COVID-19 infections per age group in the Netherlands in 2020. Density here means the proportion of the population per "unit on the horizontal axis" (Freedman et al., 1978, p. 33) with this unit being 10 years of age in our example. For example, the age group 80–110 has a proportion of about 0.02 per 10-year group. Calculating the actual number of people can be done through a multiplication 0.02 × 826,192 × 3 (this 3 is because there are three 10-year age groups in the age group 80–110) which returns 49,572

---

[4] This definition is a refined version of the one given in the next chapter.
[5] For example, a population pyramid (age-sex pyramid) is an exception.

people[6], which is almost the actual total of age groups 80–90, 90–100, and 100–110 together in Figure 1.7.

---

1. Other types of graphs with bars: bar charts, distribution bar charts, and case-value plots

Histograms are often confused with other graphs with bars. One type is a *distribution bar chart* or *distribution bar graph* (univariate, categorical data along the horizontal axis, counts of data in these categories along the vertical axis). Similarly to a histogram, this distribution bar chart contains aggregated data as the height of the bar represents multiple data points (e.g., the blood type figure in the next chapter). These data could also be represented in pie charts whereas data depicted in a histogram cannot.

　　　Another type of graph with bars is a graph in which each bar's height represents one measured value. We call this a *case-value plot* (cf. Garfield & Ben-Zvi, 2008a). There is but different terminology for this type of graph: "value bar chart (aka "case value graph" [delMas et al., 2005], […] and "ordered value bar graph" [Lappan et al., 2014])" (Cooper, 2018, p. 111). A time-plot can "be considered a special case" of case-value plots (Cooper & Shore, 2010, p. 4).

　　　Variants also exist, for example, *stacked* forms of *case-value plots*. The words *bar charts* usually refer to all these graph types together but not to histograms. In the section on education, some examples can be found. Instead of 'bar charts', 'bar graphs' is sometimes used (e.g., Humphrey et al., 2014).

In histograms, bars are often connected. Nevertheless, this is neither a defining nor a distinctive feature to distinguish histograms from other graphs with bars (e.g., Ioannidis, 2003; Rufilanchas, 2017). A density histogram is hard to make using common software (e.g., neither Excel nor SPSS can make density histograms and Excel often makes mistakes with regular histograms)[7].

---

[6] The actual number in our cleaned dataset of 2020 was 47,934. Differences are due to rounding.

[7] There are workarounds, but then the graph is constructed by the user.

Figure 1.8 Density histogram for COVID-19 infections per 10 years in the Netherlands in 2020
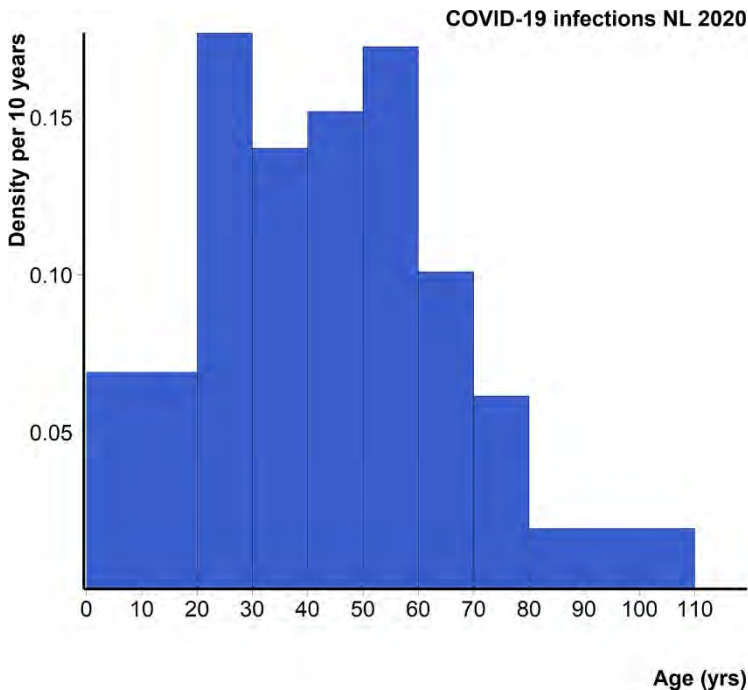
Note. This density histogram was created by Alex Lyford with *ggplot2* using a workaround.

1.1.3   The role of histograms in statistics and statistical literacy

Most citizens read the results of investigations in a newspaper and magazine or see these on television, news websites, and social media. Especially in these times when fake news spreads at lightning speed, it is important that people can critically evaluate results. Critical evaluation is part of statistical literacy:

> …people's ability to *interpret and critically evaluate* statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant (b) their ability to discuss or communicate their reactions to such statistical information, such as their understandings of the meaning of the information, their opinions about the implications of this information, or their concerns regarding the acceptability of given conclusions. (Gal, 2002, pp. 2–3, emphasis in original)

Statistical literacy requires graph comprehension and being able to interpret and produce graphs of data. This ability is also called graphicacy (Balchin & Coleman, 1966) or graph(ical) literacy (e.g., Gillespie, 1993). Graphicacy is "the

ability to read and write (or draw) graphs" (Fry, 1981, p. 383). Fry also includes pictograms and sketches in this definition of graphs, but we prefer to stick to what mathematicians usually refer to when talking about graphs. For this type of graph, three aims can be distinguished: propaganda, analytical, and substitute for tables (Tukey, 1972). In line with Tukey, we are specifically interested in analytical graphs. Some studies focus on people's difficulties with misleading graphs (e.g., Wijnker et al., 2022). In this dissertation, we focus on students' difficulties with correctly constructed graphical representations of data.

---

2.   Histograms have two axes but depict *one* statistical *variable*

A returning topic of debate when discussing histograms with scientist is whether histograms depict one or two variables. An argument for stating it is one statistical variable is that the same data can be presented in dotplots, stem-and-leaf plots, boxplots, and density graphs; all being used for depicting one, statistical variable, and all without a vertical axis. A (density) "histogram does not need a vertical scale" and when income is along the horizontal axis "the area of each block [bar] is proportional to the number of families with incomes in the corresponding class interval" and the total area of a histogram is 100%, or one if proportions are used (Freedman et al., 1978, pp. 25–26). Another argument is that the algorithm for computing the arithmetic mean from histograms is different (i.e., sum of the measured values is divided by the sum of frequencies along the vertical axis, instead of number of bars) compared to, for example, case-value plots (where sum of the measured values is divided by the number of measured values along the horizontal axis, often being the same as the number of bars). "Univariate graphs provide information about the distribution of observations on a single variable. […] The histogram is by far the most commonly used procedure for displaying univariate data." (Jacoby, 1997, p. 13).

        An argument for two variables is that there are two axes. To program software to plot this graph, two variables need to be defined somewhere in the software. Therefore, I often reply that there is only one *statistical* variable. If I had to label the other variable, I would call it a 'plotting' variable.

A histogram can be regarded as a spider in a web of knowledge. Histograms prepare for key concepts such as probability distribution and density in probability theory (Batanero et al., 2004). Histograms may play a central role in learning statistical key concepts such as data, distribution, variability or variation, and central tendency (Garfield & Ben-Zvi, 2004). Each key concept—such as distribution—relies on other concepts (e.g., center, density, skewness, relative frequency) (Bakker & Gravemeijer, 2004). Shape is often also included in this list, but one might wonder whether the focus should be so much on shape. As an example, we invite readers to think about the normal distribution.

What kind of shape do you have in mind? When people think about a normal distribution, they may imagine a bell shape, think about a straight line on normal probability paper, or an S-shape (the cumulative frequency polygon). Others may think of

> …the probability density function, the Galton board, or we think of phenomena that can be modeled with the normal distribution (for example, height). In line with Dörfler's observation that he could not find the concept of the number 5 or the triangle in his mind, we cannot find the concept of the normal distribution in our mind, only representations. (Bakker, 2004a, pp. 31–32)

The bell curve or other appearances of the normal distribution are signs (Bakker & Hoffmann, 2005)—graphs if one likes—of the concepts. Moreover, the shape of a histogram depends not only on the data but also on the binning choices (e.g., Sahann et al., 2021; Setlur et al., 2022). Therefore, the appearances—shapes—are not the concepts themselves. Unfortunately, we cannot learn concepts without signs, without graphs. Therefore, we see a histogram as a means to teach students about those concepts.

In the Netherlands, histograms are taught mostly in Grades 9–12. They have been underrepresented in research literature while they are widely used in practice (e.g., Lem et al., 2014b). When we began this research, it was unclear how histograms could play a role in developing students' statistical literacy as part of critical citizenship. Research on students' difficulties with histograms was hard to find and fragmented. A gap existed between research and teaching practice, at least in my own country, the Netherlands (cf. Bakker et al., 2021). Knowledge of how to effectively teach histograms was lacking. These problems existed for many years, (e.g., Ismail & Chan, 2015; Meletiou, 2000; Pettibone & Diamond, 1972) despite some carefully designed interventions (e.g., Kaplan et al., 2014).

Therefore, this doctoral dissertation concentrates on histograms to develop students' critical citizenship and statistical literacy. In our research proposal for this dissertation, the research question was: *How can pre-university track students in Grades 10–12 learn to draw correct conclusions from histograms*? After the first study, it became clear that the focus of our research should not be on histograms only, but on understanding key concepts that become visible through histograms. To elicit this focus on students' understanding, we changed the overall research question into:

> *RQ: How can pre-university track students in Grades 10–12 be supported in understanding histograms?*

In the Netherlands, mathematics is mandatory for pre-university track (vwo) students and they can choose from three types of mathematics: A, B, and C. Mathematics C is preparation for cultural and art studies and includes statistics, Mathematics B prepares for technical and other scientific studies and does not contain statistics, and Mathematics A concerns applied analysis in economics and health contexts and statistics (Daemen et al., 2020). In addition to Mathematics B only, students can choose Mathematics D, which contains statistics and probability, as well as some other topics that broaden and deepen their mathematical knowledge from Mathematics B. In this dissertation, we mostly concentrate on the group of students with Mathematics A as these students have statistics in their curriculum.

1.2   Educational aims for using histograms

1.2.1   What is the educational problem with histograms?

Students misinterpreting histograms

As explained in the previous section, histograms are often confused with case-value plots (e.g., Bakker, 2004a; Lem et al., 2013a, 2014b). Adding to students' confusion, not every graph depicting frequencies is a histogram. Consider, for example, the number of people who died in a hospital (Figure 1.9, inspired by Gill, 2022). Although, at first glance, this graph might look like a histogram—as it has frequency on the vertical axis—but it is not. First, consider the variable along the *horizontal* axis. This is an ordinal variable (day of the week) and calculating an arithmetic mean of it would make no sense. Compare this to the COVID-19 histogram (Figures 1.7, 1.8), where the mean age (roughly 43 years) is along the horizontal axis and can be depicted by a vertical line that crosses the *horizontal* axis at 43 years. Second, in Figure 1.9, the mean (number of people who died) can be depicted by a horizontal line at 58 people per day, crossing the *vertical* axis at that number. Third, we can assess the variation. Imagine a graph similar to Figure 1.9 in which almost all people died on Saturday or Sunday. This would be a graph with only two high bars and five very low bars. Would that indicate much variation or not? We would assume that this would be considered a lot of variation. In that case, the graph is a time-plot (which can be considered a special case of a case-value bar chart regarding mean and variation; Cooper & Shore, 2010). But if this graph had been similar to a 'histogram' (or, to be precise: a *distribution bar chart*, as the horizontal scale is ordinal), the imaginary graph with only two high bars would *not* be considered much variation at all, as all deaths are concentrated around the same two days.

Figure 1.9 Number of people who died in one hospital per weekday counted over one year
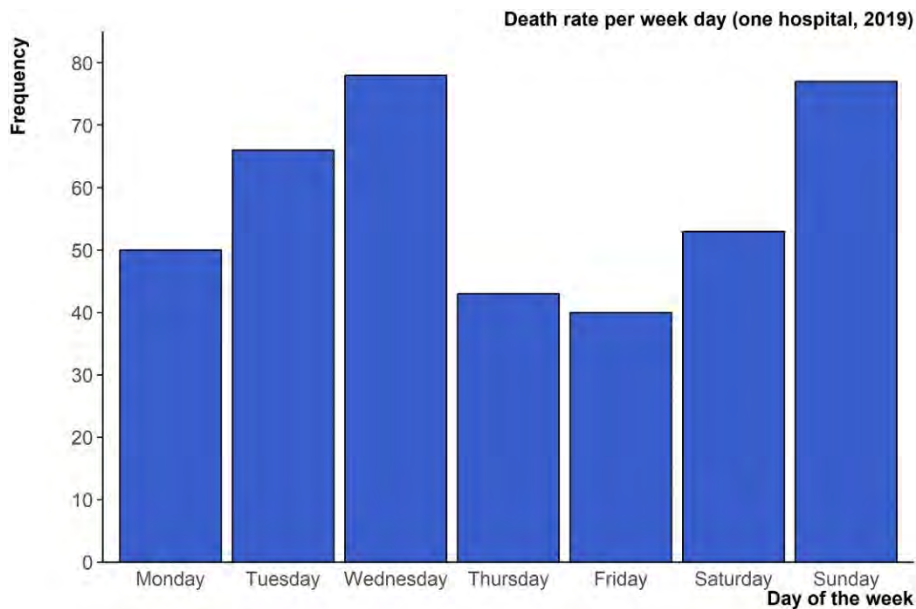
In addition, "The distinction between distribution bar graphs and value bar charts [case-value plots] can blur if frequency is found on the vertical axis and the data itself is not well-defined" (Cooper & Shore, 2010, p. 14). The same holds true for histograms and some time-plots. For example, consider the number of people who died at certain times (Figure 1.10, inspired by Gill, 2022). Although at first glance, this graph might look like a histogram, it most likely is not. However, it really depends on what you consider to be the data. We would expect here that the data are the number of people who died. In that case, the mean number of people who died in a hospital per hour is approximately 17 and can be found along the *vertical* axis. Moreover, a dotplot for number of deaths cannot be made from Figure 1.10, or at least not without discarding the crucial time-of-the-day information, unless the mean time of the day somebody died is the variable of interest, in which case, time would be on the horizontal of this dotplot. Second, how is variation assessed? If the variation in the frequency only is considered (e.g., high and low peaks), the variation in *heights* of bars (vertical variation) is assessed as if this graph is a kind of case-value plot (with a standard deviation for the frequencies of 7.9). If this graph were a histogram, then the mean hour of deaths is roughly in the morning (mean hour: 10.84 or 10:50). For the variation in the data, we would then look at the *horizontal spreadoutness* of the data in combination with the *heights* of bars.

Figure 1.10 Number of people who died due to COVID-19 in one hospital in one year (fictive numbers)
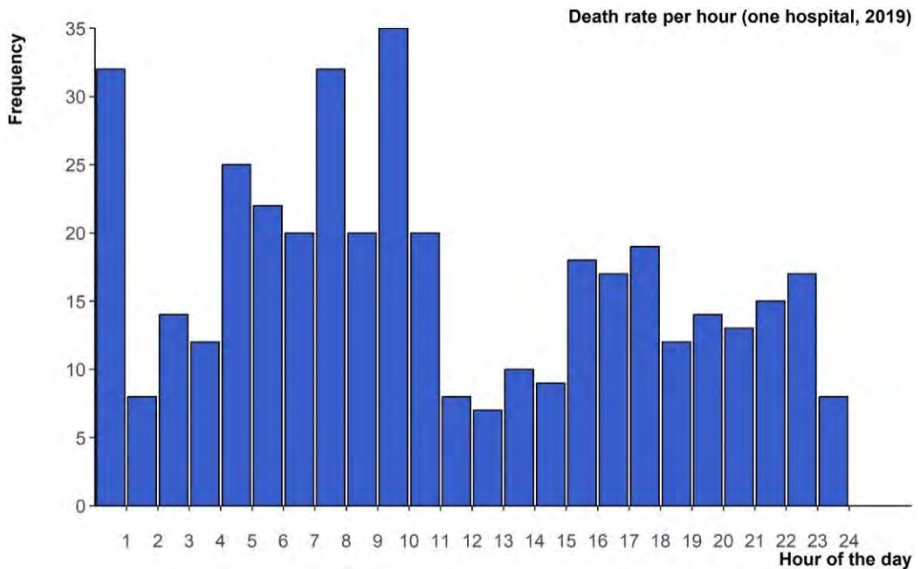
As explained in the previous section, in English, different words are used to distinguish different graphs with bars from each other; see the box on other types of graphs with bars. For example, the graph in Figure 1.10 is called a time-plot (Cooper & Shore, 2010). This naming might seem to be merely extra detail, but as we will see from the eye movements of students in Chapter 3, it is not.

Teachers' difficulties with histograms

When we give workshops to teachers—after presenting them with the graphs from Figure 1.4—we regularly ask them to sort graphs with bars from textbooks and newspapers (Figure 1.11, see also Boels, 2019). Which of these graphs are histograms? Teachers often find this a challenging task. One way to decide on this is to try to find the mean. Another way is to make a graph with dots. Graphs a, b, and e would result in a line graph, and the mean can be found on their vertical axis. Graph d would result in a dotplot, and the mean can be found on the horizontal axis. For graph c, it is impossible to decide what kind of graph it is. We advise avoiding these graphs (e.g., in textbooks), as these emphasize most noticeable features instead of relevant ones[8].

---

[8] See also my videos: https://youtu.be/zpRHhixoYmg and https://youtu.be/5od2uB908PI

Figure 1.11 Some graphs with bars found in textbooks and websites
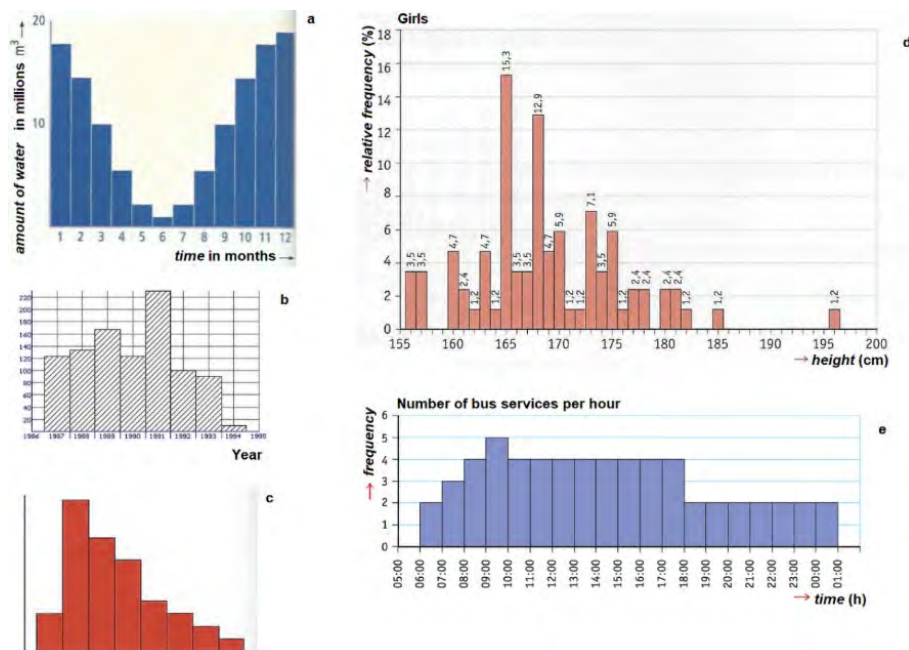
Note. Source: **a)** Moderne Wiskunde, 5 vwo, Mathematics A, 11th edition, **b)** Wisfaq explanation of histograms, 2017, November 1, **c)** Getal en Ruimte, 4 havo, Mathematics A, part 2, 11th edition, **d)** Mathplus, 4 havo, Mathematics A, part 3, 1st edition, **e)** Mathplus, 4 vwo, Mathematics A and C, part 2, 1st edition.

Research in which histograms of research outcomes caused difficulties

The previous section hinted at possible misinterpretations that can occur when using histograms. However, it is not only students that have difficulties with interpreting histograms, but also researchers. Here, we provide two examples. In both examples, histograms were avoided when they should have been used. The conclusions of the research could still be correct and the quality of the rest of the research could still be high. This is up to others to judge. My focus is strictly on the correct use—or avoidance—of histograms.

In the first example, the diameter of savanna trees in Australia—the Banksia Marginata—was measured at breast height (Heyes et al., 2020). Bins were created with diameter on the horizontal scale and number of trees (frequency) on the vertical scale. The issue is with the graphs constructed in the study and the calculations that were done. First, the researchers drew a line graph, which is suitable for *two* statistical variables but not for *one*. Second, they used logarithmic scales for both axes. This means that for larger breast height widths, bin sizes are bigger, requiring a kind of density graph (density histogram, violin plot). Third, they calculated the correlation between

the middle of the bins and the frequency. The latter can be problematic as the choice for binning in histograms can severely influence the shape (e.g., Sahann et al., 2021; Setlur et al., 2022), which in turn influences what correlation will be found.

The second example stems from a course for prospective mathematics teachers (Norabiatul Adawiah et al., 2021). To compare the distribution of scores in two groups—*one* statistical variable measured at ratio measurement level—a kind of double distribution bar chart was used (suitable for one variable measured at *nominal* or *ordinal* measurement level). For teachers, we would expect that data would be presented in two histograms, two boxplots, or two dotplots (e.g., Biehler, 2007; cf. Rodríguez-Muñiz et al., 2022) or that it would be discussed which didactical reasons justify this deviation (see also Chevallard & Bosch, 2014). In addition, for comparison, a group size of $N = 12$ is quite small.

1.2.2 Histograms in Dutch mathematics education

In the Netherlands, there are different curricula for pre-vocational education (vmbo) and general secondary education (havo and vwo: pre-college and pre-university track education). However, in most textbooks, histograms are introduced in Grade 9 (e.g., Getal en Ruimte, 2015) or sometimes in Grade 10, although one textbook that is no longer available (Mathplus, 2014) very briefly touched upon histograms in Grade 7. In addition, as explained earlier, in the pre-university track students choose one type of mathematics, and it is that choice that determines whether histograms are further elaborated on.

A typical introduction in the Grade 9 pre-university track is that students are first asked to aggregate given data into a frequency table. Next, they perform some calculations with frequency tables, such as calculating the arithmetic mean or the median. Finally, students are asked to draw a histogram for a given frequency table. There can be quite a few months or even a whole school year in between those steps. In Moderne Wiskunde (2019, p. 102), dotplots are used in a statistics chapter in Grade 10 Mathematics A and C. The word histogram is not used in this textbook and the authors do not seem to have clarity about what kind of graphs are used for what kind of data. In one task, for example, they ask students to make a dotplot, a bar chart, and a pie chart for the same data (p. 103). As dotplots are suitable for data measured at ratio or interval measurement level (sometimes called numeric or quantitative data) and pie charts are used for nominal or ordinal data (categorical data or qualitative data) this task does not seem to make sense, although they asked students to reflect on the most suitable chart for these data. A quick scan of the three most used Dutch mathematics textbooks in 2021 indicates that they all suffer from inconsistencies,

misinterpretations, and a focus on procedural knowledge instead of developing statistical literacy.

Regarding the latter, for example, students rarely collect their own data. Although histograms are important tools for data analysis, data communication, and interpretation including inferential reasoning (see investigative cycle, Wild & Pfannkuch, 1999), the focus in textbooks is often on *how* to draw a histogram and how to *calculate* something with the data it represents. Interpretation of histograms is rare. Histograms can be used for data exploration, hypothesis generation, communication, interpretation, developing new ideas, drawing conclusions about the collected data or the population, and even sometimes to support data cleaning (e.g., to find typos that may be depicted as outliers).

1.3   Personal motives for choosing statistics education as a research area

Since my training as a mathematics teacher, like many other teachers, I have been concerned with the question of how topics in mathematics can best be taught. Through my experiences in teaching and in previous research, I have noticed that the effectiveness of statistics education in Dutch secondary schools (Grades 10–12) is far from optimal. Many teachers do not feel well-equipped to teach statistics (e.g., Van Dijke-Droogers, 2021). As a result, students are poorly prepared for further study and society. Girls in particular become very frustrated by this as they often choose studies for which statistics courses are an important part of the university curriculum (e.g., psychology). As I was teaching Mathematics A most of the time, I felt personally responsible for their failure and frustration and I wanted to do something about that. This was my first motive.

My second motive stems from students' difficulties. The examples in the previous sections show that graphing data can be a crucial step in data analysis and interpretation. As discussed earlier, some important discoveries were made purely by graphing data. Interpreting graphs seems a simple first step in the statistics curriculum, but in practice, students have little understanding of statistical graphs such as histograms. For boxplots, students' misinterpretations are well known (e.g., Bakker et al., 2004), but for histograms, most teachers seemed unaware of students' difficulties (e.g., Cooper, 2002). I wanted to find out exactly why students have difficulties understanding histograms and what can be done about it.

Given my technical background—I was trained as an electrical engineer—and my training to become a mathematics teacher approximately ten years later, I felt well-equipped for supporting students with topics within

calculus. Statistics, however, was never really part of my training, apart from some procedural knowledge such as how to calculate the mean or standard deviation from a frequency table. In past Dutch teacher training, statistics got little attention. I do not remember that I ever read or talked during my training to become a teacher—approximately twenty years ago—about the investigative cycle (Wild & Pfannkuch, 1999), talking students through the data collection (Ben-Zvi et al., 2018; Cobb & McClain, 2004) or key concepts (e.g., data, distribution), and how these relate to each other and to teaching. Further professionalization of teachers was, therefore, considered necessary when the new curriculum for Mathematics A was introduced in 2015, especially for statistics (cTWO, 2007). By conducting research on a topic I didactically knew little about, I wanted not only to advance teaching and research in statistics education but also to become a better-equipped teacher myself. This was my third motive to engage in this research.

1.4   Overview of this dissertation

Chapters 2 to 6 form the core of this dissertation. Below, a brief description of each chapter and its research question is given. In Figure 1.12, an overview of the studies and chapters can be found. When we started the trajectory for this dissertation, we expected that we would use the literature on students' difficulties with histograms and a small-scale eye-tracking study both as inputs for a larger design study (Bakker, 2018). In that case, design research would have been at the heart of this dissertation. However, during the first study—a review of the literature, see Chapter 2—it became clear that several attempts had already been made to carefully develop interventions to tackle students' misinterpretations. The success of these varied, often even within a single intervention. In a study by Kaplan et al. (2014), for example, after taking an introductory statistics course at a university, upon completion students were better able to distinguish a histogram from a case-value plot. In addition, confusing horizontal and vertical axis when determining the median decreased slightly. "Unfortunately, this may be due to the item construction, rather than actual students' knowledge" (p. 16). Moreover, confusing the horizontal and vertical axis when comparing the mode of two histograms increased. The overall impression we got is that most interventions had not been very successful. Therefore, instead of trying out another intervention, we decided to dig deeper and do what McKenney called "a lot of 'front-end work', [which includes …] understanding the problem better", identifying students' difficulties, and "formulating design criteria" (Bakker, 2018, p. 142). Hence, in the second study (Chapter 3), we decided to figure out on a more fundamental level what students' difficulties with histograms were through a larger *eye-*

tracking study, as we thought that students' gaze patterns could provide insight into their approaches.

Chapter 2. As an overview of students' difficulties with interpreting histograms was lacking in the research literature, the first step was to create such an overview through an extensive review of the literature. The research question for this study was:

> *RQ1: What are the conceptual difficulties that become manifest in the common misinterpretations people have when constructing or interpreting histograms?*

A narrative systematic review of the literature with a configurative synthesis was conducted (Gough et al., 2017). Data were collected through a systematic search in several databases. From each publication in this review, we collected the misinterpretations that were reported or discovered.

Chapter 3. The review results made it possible to address students' conceptual difficulties that become manifest in most common misinterpretations more broadly rather than focusing on a specific misinterpretation. Misinterpretations related to the statistical key concepts data and distribution can be observed when students confuse histograms with look-alikes, including case-value plots. In addition, many of the studies in the literature review draw conclusions from students' final answers (e.g., Whitaker & Jacobbe, 2017). Little was known about students' strategies for reaching these answers. Therefore, it was unclear how to intervene effectively. By observing students' actions, it becomes clear how students use their conceptual knowledge of the data in histograms, hence what strategies they employ. Eye movements can reveal students' strategies (Van Meeuwen et al., 2014). We answer the following research question:

> *RQ2: How and how well do Grades 10–12 pre-university track students estimate and compare arithmetic means of histograms and case-value plots?*

We used eye-tracking as a data collection method, as gaze patterns can provide detailed insight into students' thinking processes, including those processes that students are not aware of or are not able to articulate (Green et al., 2007). We tracked the gazes of students (50) and teachers (18), although teachers are not included in this dissertation for reasons of time (Boels et al., 2019b). Students were asked to estimate or compare arithmetic means. Students' gaze data were qualitatively coded and combined with interview data from cued recall to connect specific gaze patterns—the perceptual forms of gazes—to interpretation strategies.

Figure 1.12 Overview of the studies and chapters in this dissertation
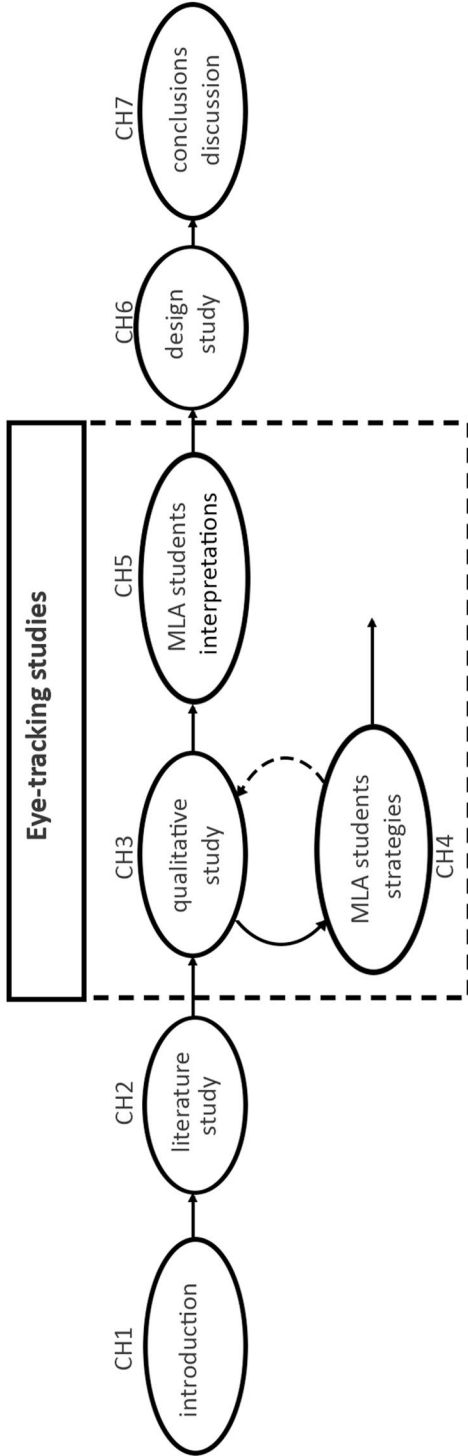
Note. CH = chapter, MLA = Machine Learning Algorithm.

Chapter 4. The three patterns we found in students' gaze data for single histograms sparked us to explore whether automatic recognition of students' strategies might be possible through a machine learning analysis. A potential benefit of automatic recognition would be that targeted intelligent feedback could be given, based on students' strategies inferred from gaze data, during online learning. This, in turn, could help reduce the pervasiveness of misinterpretations among students as described in Chapter 2. Chapter 4 describes a first step in this automation process. The research question for this study was:

> *RQ3: How can gaze data be used to automatically identify students' task-specific strategies on single histograms?*

We made an Interpretable Mathematical Model (IMM) of the gaze data based on heuristics stemming from the previous study. To provide a baseline for the IMM, we used a supervised machine learning algorithm (MLA). The chosen software tool (Mathematica Classify Function) automatically prepared the gaze data and fed these into an automatically chosen MLA. As we also used some single histograms that were not yet analyzed in the previous study, this required another round of qualitative coding. In the chapter on this study, it is explained why that was necessary. The quantitative approach through an IMM and machine learning analysis contributed to the reliability of the results. A similar study for the double histogram tasks is planned for the future.

Chapter 5. The previous studies revealed students' solution strategies when solving histogram tasks in more detail. A local instruction theory in statistics education suggests that having students solving dotplot tasks can support students' learning to interpret histograms (e.g., Bakker & Gravemeijer, 2004; Garfield, 2002; Garfield & Ben-Zvi, 2008a), as dotplots can draw students' attention to the variable being presented along the horizontal axis in both graphs. In this study, previously collected gaze data were re-used to explore whether students' histogram interpretations change after solving dotplot items. We used students' gaze data on four histogram items as inputs for an MLA (random forest) to answer the research question:

> *RQ4: In what way do Grades 10–12 pre-university track students' histogram interpretations change after solving dotplot items?*

In addition, we used students' verbal reports and answers to investigate whether changes in gaze patterns reflect changes in students' approaches.

Chapter 6. The literature research (Chapter 2) also made clear that existing interventions were not sufficiently successful in teaching students to correctly interpret histograms. The students' solution strategies (Chapter 3) showed that

many of these Dutch students lacked understanding of how and where data are represented in histograms. Interpreting dotplots may assist students' understanding of histogram (Chapter 5) but it was still unclear how an intervention could be designed that would support students' learning of statistical key concepts through interpreting dotplots and histograms.

Therefore, in this last study, we made a start with a *design study*. From the previous studies, we got the impression that students lacked experience with dotplots and sufficient attention to how these artifacts—histograms, dotplots—become tools in statistical reasoning. We suspected that students' education might have lacked an embodied grounding of how histograms are constructed. Therefore, using embodied instrumentation approach as a theoretical lens, we designed a learning trajectory that drew upon findings and insights from previous studies. The research question for this study was:

> *RQ5: What sequence of tasks designed from an embodied instrumentation perspective can support students' understanding of histograms and the underlying key concepts?*

Chapter 7. This chapter presents general conclusions and discussion. We answer the main research question. The theoretical and methodological insights, implications, and recommendations for research and educational practice are elaborated.

Chapter 1

2

Conceptual difficulties when interpreting histograms: A review

"Amazing the things you find when you bother to search for them." [9]
Sacagawea

"A mistake is just another way of doing things." [10]
Katharina Graham

This chapter is based on
Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019). Conceptual difficulties when interpreting histograms: A review. *Educational Research Review, 28,* Article 100291, 26 p.
https://doi.org/10.1016/j.edurev.2019.100291

---

[9] Divya Raghav (2021), 10+ Best Sacagawea quotes from the influential explorer. Kidadl.
https://kidadl.com/quotes/best-sacagawea-quotes-from-the-influential-explorer
[10] Garson O'Toole (2023, April 24). Quote origin: A mistake is just another way of doing things.
https://quoteinvestigator.medium.com/quote-origin-a-mistake-is-just-another-way-of-doing-things-754f6ae01548

Abstract Histograms are widely used and appear easy to understand. Nevertheless, research indicates that students, teachers, and researchers often misinterpret these graphical representations. Hence, the research question addressed in this chapter is: What are the conceptual difficulties that become manifest in the common misinterpretations people have when constructing or interpreting histograms? To identify these conceptual difficulties, we conducted a narrative systematic literature review and identified 86 publications reporting or containing misinterpretations. The misinterpretations were clustered and—through abduction—connected to difficulties with statistical concepts. The analysis revealed that most of these conceptual difficulties relate to two key concepts in statistics: data (e.g., number of variables and measurement level) and distribution (shape, center, and variability or spread). These key concepts are depicted differently in histograms compared to, for example, case-value plots. Our overview can help teachers and researchers to address common misinterpretations more generally instead of remediating them individually.

2.1   Introduction

Statistical literacy is a core competence for citizenship, and, therefore, an important goal of statistics education for students of all ages (Ben-Zvi et al., 2017). It includes the ability to interpret graphical representations of statistical data (Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2007). Graphical representations of statistical data can be found in newspapers, schoolbooks, research articles, government policy reports, television, news bulletins, and other common sources of information. "Graphical representations serve as useful tools to communicate aspects of a distribution as they facilitate a focus on aspects of the data that may be missed with the use of descriptive statistics alone." (Leavy, 2006, p. 90); see also Pastore et al. (2017). Different representations reveal different aspects of the data. Many real-life examples show that lives can literally be saved if people master the ability to switch between different representations of data to reveal different aspects. One example is from Nightingale, who saved many lives with her famous polar graph (Martineau, 1859) which showed that more soldiers died from preventable diseases—caused by bad hygienic circumstances in the hospitals— than from the war wounds caused by the Crimean War.

A graphical representation widely used to represent the distribution of univariate scale data is the histogram. What researchers consider a histogram is rarely defined. In addition, some researchers (e.g., Stevens & Palocsay, 2012; Wong, 2009), teachers, and citizens use—often implicitly—a definition of a histogram that deviates from what statisticians refer to as a histogram (e.g., Cooper & Shore, 2010; Friel et al., 2001). In the statistics literature (e.g., Bruno & Espinel, 2009; Cooper & Shore, 2010; Pearson, 1895; Shaughnessy, 2007), a regular histogram is defined as a graph with bars that meets the following criteria (see Figure 2.1 for an example and a non-example):

- The data of only one statistical variable are presented on the horizontal axis;
- The data are measured at interval or ratio measurement level;
- The variable is preferably continuous;
- The vertical axis typically displays the class density, or—when bin[11] widths or class intervals are equal—relative frequency or frequency[12].

---

[11] Ioannidis (2003) uses the word 'bucket' instead of 'bin.'

[12] In some languages the word frequency refers to relative frequency only and the word count is used to address absolute numbers. An example is found in French textbooks where the word *effectifs* is used for absolute frequency and the word *fréquence* is used for relative frequency (e.g., Derouet & Parzysz, 2016). In English and in our manuscript, the word frequency means absolute frequency.

Connected bars are neither a defining nor a distinctive feature to distinguish histograms from other graphs with bars (e.g., Ioannidis, 2003; Rufilanchas, 2017).

At first sight, histograms may appear easy to understand, but research indicates otherwise (e.g., Lem et al., 2014b). In fact, many errors, misconceptions, and mistakes in interpreting histograms have been documented in the literature (e.g., Bruno & Espinel, 2009; Derouet & Parzysz, 2016; Friel et al., 2001; Kaplan et al., 2014; Lem et al., 2013c). However, a systematic overview of these misinterpretations—a term we use as an umbrella for the ways in which people interpret histograms incorrectly—has not yet been compiled. Research repeatedly showed the persistence of the misinterpretations despite various attempts to improve statistics education (e.g., Ben-Zvi & Garfield, 2004b; Ben-Zvi et al., 2017; Chance et al., 2004; Cohen, 1996; Garfield & Ben-Zvi, 2007; Kaplan et al., 2014; Shaughnessy, 2007). Hence, there is a need to reflect on what conceptual difficulties may lie at the basis of these persistent misinterpretations. The aim of this review is, therefore, to make an inventory of the misinterpretations that occur when people use histograms, as well as to categorize these misinterpretations along the conceptual difficulties that become manifest in these misinterpretations. In this review, the word 'people' refers to students of all levels, as well as to teachers, researchers, teacher educators, and others. The question guiding this review is:

> *What are the conceptual difficulties that become manifest in the common misinterpretations people have when constructing or interpreting histograms?*

Organizing misinterpretations by conceptual difficulties that may lead to them seems to have several advantages. First, it provides a better understanding of the misinterpretations (in terms of types or common difficulties). Second, once the conceptual difficulties that become manifest in the most common misinterpretations are made plausible, researchers and educators can address these more generally instead of treating or remediating misinterpretations one by one. Such a pedagogical route would be in line with the current view in statistics education, which aims to ensure that students develop an understanding of the key concepts of statistics in relation to each other. In the statistics education literature, the term 'big ideas' was once used more often than the now more common term 'key concept'. We use 'key concept' and 'big ideas' as interchangeable terms. Up to now, research usually focuses on a specific misinterpretation (e.g., of the standard deviation) instead of multiple misinterpretations that together are a manifestation of a conceptual difficulty with a key concept (e.g., of the distribution). Third, this overview is useful for

all researchers in education and not only for mathematics education research because many education researchers either use statistics to analyze the results of their research or teach statistics in—for example—psychology or education. For researchers using statistics, graphing is a logical first step in analyzing quantitative data—for instance, when looking at the distribution of univariate data—and is often advised to do before calculating a measure (Ben-Zvi & Garfield, 2004b; Pastore et al., 2017). Fourth, research indicates that we need subject-topic-specific information for certain aspects of teaching and learning this topic (Leinhardt et al., 1990; Pareja Roblin et al., 2018).

Figure 2.1 Example of a histogram (left; ratio measurement level) and a distribution bar graph (right; nominal measurement level)
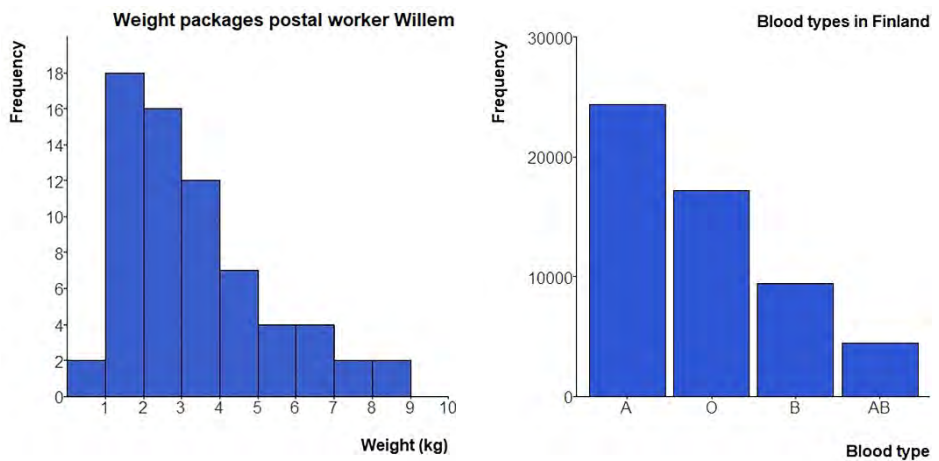
2.2   Theoretical background

2.2.1   Graphical representations

Statistical graphs often serve as the analysis of data or inquiry—as Gal (2002) phrases it—and communication of results. This requires graph comprehension (Curcio, 1981, 1987; Friel et al., 2001). Difficulties with graphical representations have been extensively studied (e.g., Arcavi, 2003; Carpenter & Shah, 1998; Larkin & Simon, 1987; Leinhardt et al., 1990; Tufte, 1983/2001; Tversky, 1997). Statistical graphs represent not only data but also statistical concepts—especially graphs that represent data in an aggregated form (e.g., boxplots and histograms). In turn, statistical concepts are inextricably represented in some form—sometimes numerically, sometimes graphically, or both. For example, for most people, the concept of the normal distribution is

inextricably connected to the bell shape as a graphical representation (Bakker & Hoffmann, 2005).

Therefore, in this review, we focus on the relation between the graphical representation and key concepts in statistics. Although some misinterpretations might be unique to the graphical representation itself, we anticipated that most misinterpretations would be a manifestation of a conceptual difficulty. Some conceptual difficulties may appear with other graphical representations too, for example, with a boxplot, which is also a graphical representation of univariate data measured at interval or ratio measurement level (e.g., Bakker et al., 2004; Lem et al., 2013c).

The literature on statistical graphs revealed that experts and novices analyze graphs in different ways. Experts tend to view a graph globally, while novices seem to focus on the local features of the graph (Khalil, 2005). Konold et al. (2015) showed that some elementary school students regard the data as a pointer to the context or situation, which is in line with the findings from other researchers that students see the graph as a picture (e.g., Friel et al., 2001; Leinhardt et al., 1990). According to Konold et al., other students focus on individual cases in the graph, for example, the shortest person, or where a specific person can be found in the graph. Yet other students see the data in a graph as classifiers—for example, for the mode or "the winning outcome" (p. 314). Elementary school students rarely see the data as aggregates, meaning that their focus is mostly not on the entire distribution. Which perspective is useful depends on the question posed to the data.

2.2.2 Misinterpretations and conceptual difficulties

In this review, we distinguish between conceptual difficulties and misinterpretations. In line with other research (Lem et al., 2013c), we use the term 'misinterpretation' to denote a repeatable and explicit mistake or error that occurs in different people (Leinhardt et al., 1990) and that relates to the conclusion being drawn from a given graph. The term 'conceptual difficulty' is widely used in the literature on physics and chemistry education when people have an incorrect, naïve or incomplete idea of a concept (e.g., Battaglia et al., 2017; Garnett & Treagust, 1992; Hammer, 1996). As a clear definition was not found in this literature, we define a conceptual difficulty as having not fully grasped or understood the key concept at hand. People who have fully grasped the key concept are not expected to show misinterpretations when drawing conclusions from graphs. When we identify a misinterpretation, we can, therefore, conclude that it is a manifestation of a conceptual difficulty.

An example may further clarify the distinction between a misinterpretation and a conceptual difficulty. When statistics teachers state that a graph has more variability because the graph is bumpier (meaning: more

difference in heights of the bars; e.g., Dabos, 2014) they assess the variability of the frequency bars in the histogram instead of the variability of the variable at hand. We infer from this behavior (i.e., showing a misinterpretation) that these teachers have difficulties with the statistical concept of variability, which is part of the key concept of distribution. This is further clarified through the examples in the next section.

2.2.3   The key concepts of data and distribution

The research on the teaching and learning of statistics identified several key concepts that underlie statistical investigations (Garfield & Gal, 1999)—as the core goals of statistics education. These statistical key concepts encompass several other concepts such as trend, model, sample, and graphical representation (e.g., Bakker, 2004a; Ben-Zvi et al., 2017; Gal & Garfield, 1997; Pfannkuch & Ben-Zvi, 2011). The statistical concepts are intricately connected (Bakker & Derry, 2011). Which statistical concepts are at stake depends on the particular context and research question posed to the data. Figure 2.2 summarizes how the various statistical concepts fit together when it comes to solving a statistical problem involving univariate data that can be represented in a histogram.

During the analysis of the groups of misinterpretations (the axial codes, see section 2.3.2), it became clear that the usual theoretical framework of key concepts in statistics—a collection of the statistical concepts they are related to—lacked a specification of the relationships between these statistical concepts. We, therefore, propose a network of statistical concepts based on the theoretical framework of key concepts found in the literature (see Figure 2.2). As it is unlikely that there is a generic relationship between these statistical concepts, we focused on those relevant to solving statistical problems that may involve the representation of univariate data in a histogram. Our contribution consists of three parts. First, we added connections between the statistical concepts, which led to a coherent network that, from our analysis[13], turned out to be relevant. In this network, we outlined how these connections can be understood. Second, we linked this network to the statistical investigation by assigning a specific statistical concept to a specific part of the statistical investigation—such as posing a question or collecting data (Wild & Pfannkuch, 1999). Assigning the concepts to the statistical investigation clarifies the consequence of misinterpretations for statistical investigations and inferential reasoning in education and research. Third, we added measurement level and number of variables as

---

[13] For example, concepts related to hypothesis testing are not included in this network as these did not emerge from our analysis.

separate statistical concepts, as from the grouping of our data it became clear that these concepts were lacking in the existing theoretical framework of key concepts. In addition, we added some statistical concepts that, beforehand, we did not expect to find in this review (e.g., correlation and covariance). These statistical concepts do not make sense for histograms. For example, correlation is only possible with at least two variables, whereas a histogram depicts only one variable. From the coding (axial codes), it nevertheless became clear that misinterpretations related to this statistical concept sometimes played a role, so we added this to the network.

We now discuss the two key concepts that turned out to be most relevant during the analysis phase. The descriptions of these key concepts are taken from Garfield and Ben-Zvi (2004, p. 400).

- Data: […] data represent characteristics or values in the real world […]
- Distribution: a representation of quantitative data that can be examined and described in terms of shape, center, and spread [variability], as well as unique features such as gaps, clusters, outliers, and so on.

Because we know from the literature in this review that the key concepts data and distribution are hard to grasp for most people, we synthesize the main characteristics in two examples.

The key concept of data

The key concept of *data* includes how many variables are depicted in the graph (see the letter F in Figure 2.2) as well as the measurement level (nominal, ordinal, interval, or ratio) of its attributes (see K in Figure 2.2). In Figure 2.3, the key concept of data is explained through the example of babies born in a hospital in Queensland, Australia (Dunn, 1999). For our explanation, only two variables of this data set are used: a number referring to each baby girl that was born (instead of her name) and her weight in grams. To visualize these data, a so-called case-value plot or value bar chart is used, which is a special type of bar graph that shows a value (birth weight) for every case (baby girl; see Figure 2.3a).

Figure 2.2 Network of statistical concepts relevant for interpreting histograms and their place in the investigative cycle
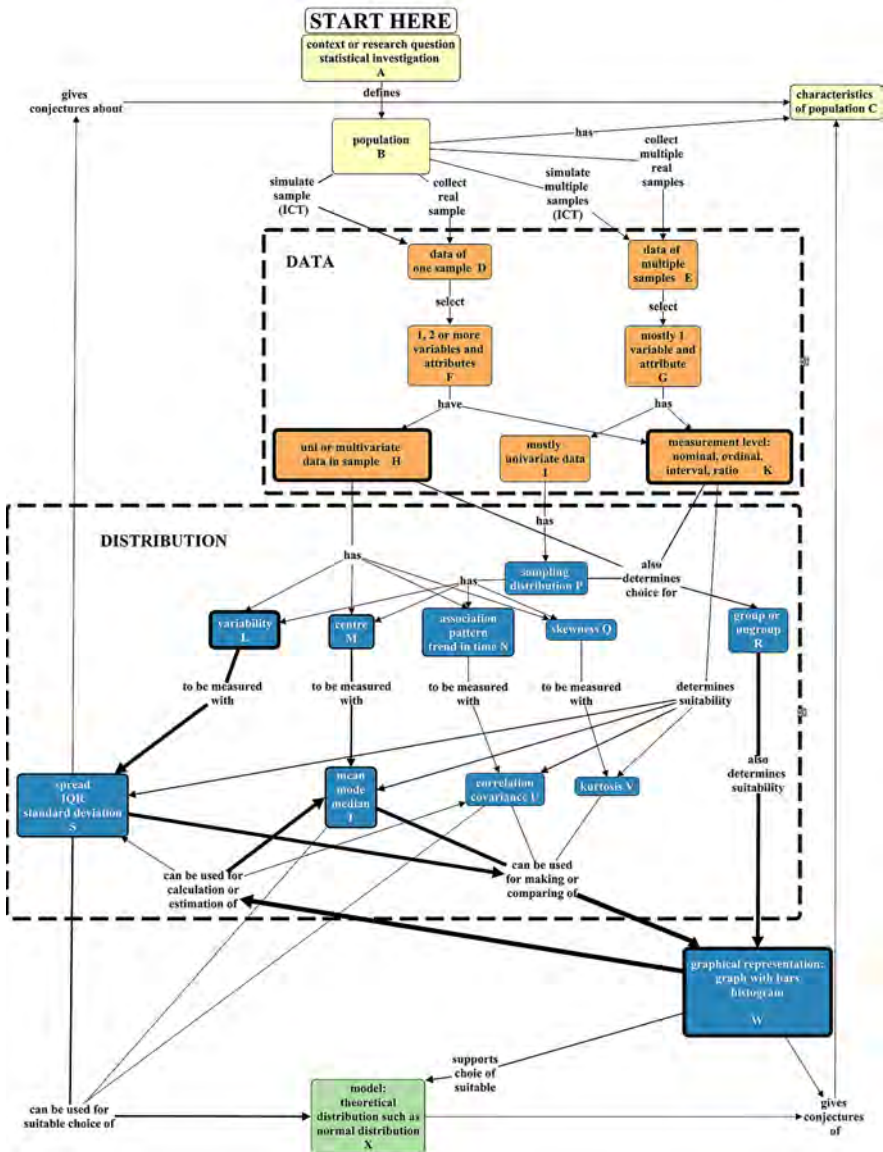
Note. Statistical concepts are located in the colored rectangles; the sizes of rectangles have no meaning. The thick lines of both arrows and boxes indicate frequently reported misinterpretations, see section 2.4 Results. The color or grayscale of the rectangles refers to different aspects of statistical investigations:

☐ posing a question   ☐ collecting data   ☐ analyzing data   ☐ making inferences

The key concepts DATA and DISTRIBUTION encompass several statistical concepts as indicated by the large, dotted rectangles. Arrows indicate relationships.

From the case-value plot with two statistical variables (see Figure 2.3a)—thus, a bivariate distribution—a histogram can be constructed through six intermediate steps. These intermediate steps are needed to tackle one of the most common misinterpretations related to the key concept of data. This misinterpretation—existing not only among many students, but also among some researchers, and some mathematics teachers—is that the number of axes determines the number of statistical variables measured, thus defining whether the distribution is univariate or bivariate (see Figure 2.3; e.g., Cohen, 1996). In the first step, one variable is removed from the graph. The resulting series of graphs is, therefore, univariate, including the histogram (see Figure 2.3b–g).

During three of the six steps described here, and during a seventh step outside figure 2.3, information reduction[14] occurs (Gal & Garfield, 1997). The first information reduction is the removal of the names of the baby girls (here anonymized; see Figure 2.3b), possibly inducing, for example, the misinterpretation that bars in a histogram can be reordered (e.g., Bruno & Espinel, 2009). The second information reduction occurs when the dots are stacked (see Figure 2.3f), possibly inducing, for example, the misinterpretation that only the middle value of the bar is observed (e.g., Biehler, 1997). The third information reduction occurs when the dots are removed from the bars, making it necessary to use a second axis for the height of the bars (density or frequency), possibly inducing the misinterpretation that two statistical variables are depicted instead of one (see Figure 2.3g, e.g., Baker et al., 2002; Dabos, 2014). When bin widths are unequal, another step is needed. A fourth step in information reduction is, therefore, using frequency density instead of frequency (not shown in Figure 2.3; Boels & Shvarts, 2023) possibly inducing, for example, wrong labeling of the vertical axis (e.g., Derouet & Parzysz, 2016).

The key concept of distribution
The key concept of *distribution* encompasses shape, center, and variability (see L–V and part of W in Figure 2.2). The distribution depends on the type of data (see H–K in Figure 2.2). In line with Cooper and Shore (2010), we argue in this section that shape (part of W in Figure 2.2), center (see M and T in Figure 2.2), and variability (see L and S in Figure 2.2) are assessed differently depending on the type of graph at stake (see W in Figure 2.2). Identifying the mean and variation in a histogram—a univariate distribution—can be done by drawing a vertical line for the mean and examining the horizontal spread of the bars, meanwhile taking the heights of the bars into account, see Figure 2.4, left.

---
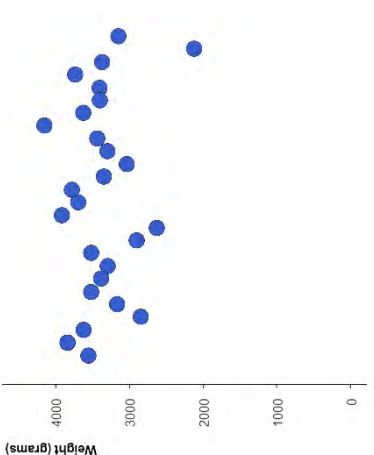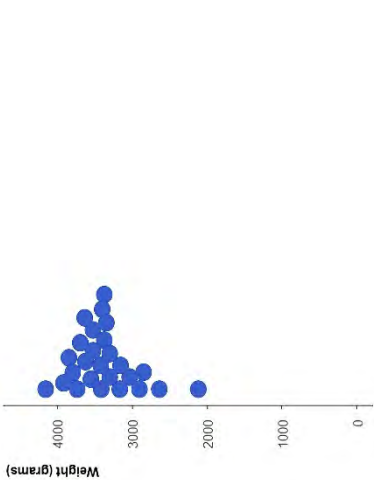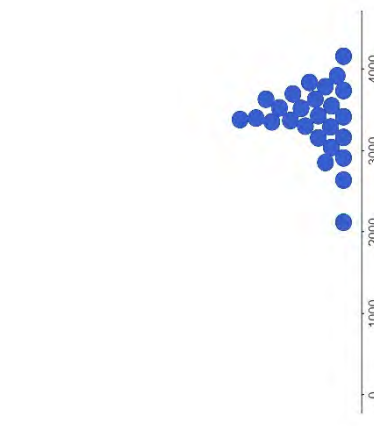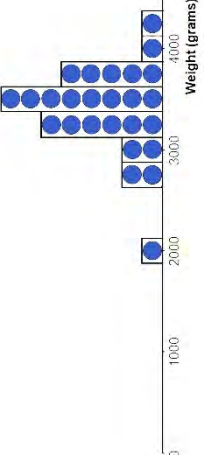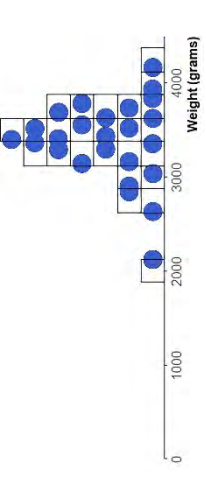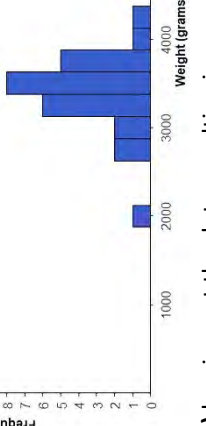
[14] We prefer information reduction over the term data reduction, as the original data themselves are not reduced—only aggregated—making other aspects, such as patterns in the data, more visible.

Figure 2.3 Example of the steps[15] needed to convert a bivariate distribution—using the variables birth weight and name—into a histogram, thus, a univariate distribution of the variable birth weight

A six-step visual on how to convert a case-value plot into a histogram; all graphs are plots of the birth weight of baby girls that were born in a hospital in Australia (Dunn, 2018).

Birth weight baby girls, hospital, Queensland, Australia, December 18, 1997

a) a case-value plot with two variables: name and birth weight
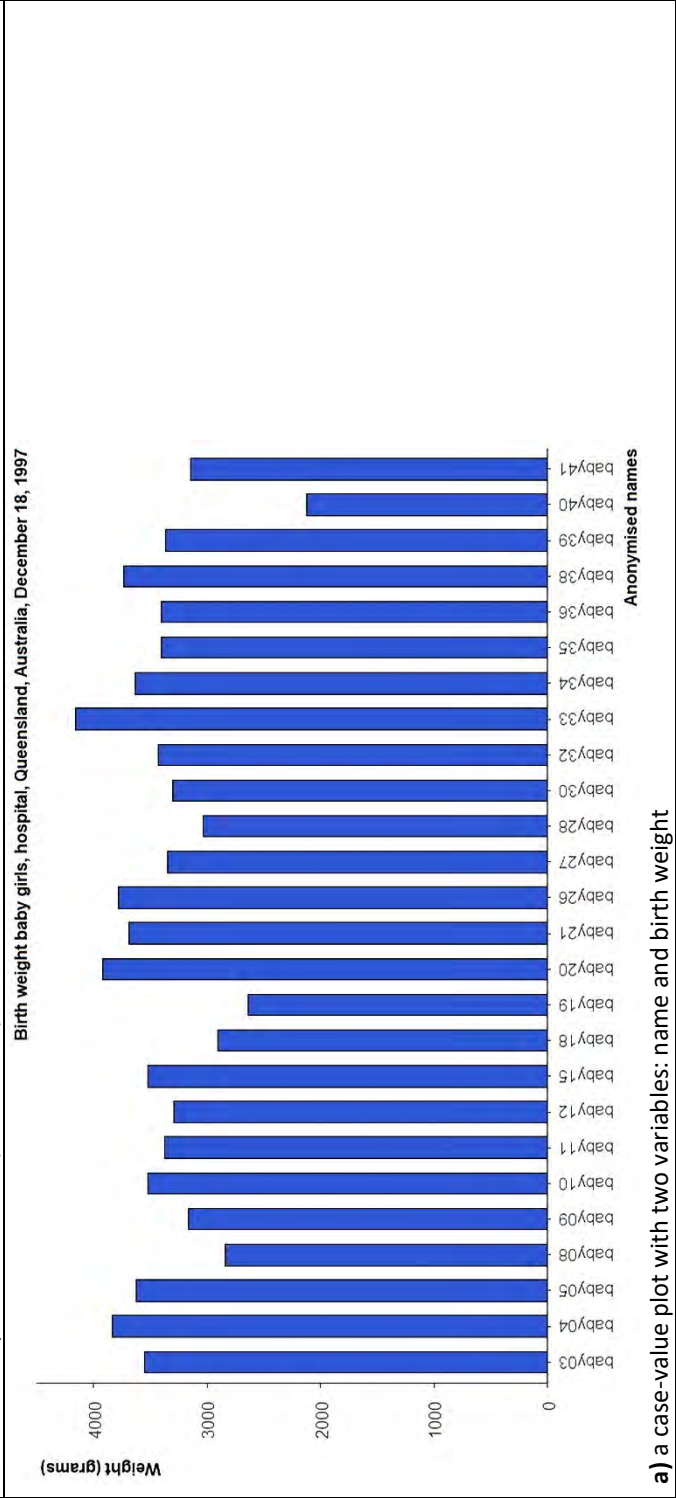
[15] Most of these steps are available in educational software such as CODAP, VUStat, TinkerPlots, Tableau and InZight. The R-code for making these graphs is available upon request from the first author.

43

b) univariate graph after removing the variable 'name'

c) all dots are slid to the left into a horizontal dotplot. Note the disappearance of the horizontal axis

d) dotplot after rotating and mirroring of the vertical axis. Note the absence of the vertical axis

e) binning the dots in a bar; roughly at each measured value

f) grouping the dots in each bar in a stacked dotplot with histogram overlay

g) leaving out the dots, resulting in a histogram. Note the reappearance of a vertical axis in this univariate graph

Identifying the mean and variation or spread in a case-value plot—a bivariate distribution—can be done by drawing a horizontal line for the mean and examining the variation of the heights of the bars around this line, see Figure 2.4, right. Note that in a histogram, less variation in the heights of the bars often indicates more variability of the variable represented on the *horizontal* axis (here: weight), whereas more variation in the heights of the bars in a case-value plot always indicates more variability of the variable represented on the *vertical* axis (here: weight). Although the graphical representations in Figure 2.4 look quite different, the underlying distribution of the variable at hand (weight) is the same. This key concept of distribution is often misunderstood as people tend to think of a distribution as the shape of the graph and not as an abstract statistical concept leading to, for example, not recognizing different graphical representations of the same data (e.g., delMas et al., 2007).

Figure 2.4 Different orientation of the mean value—the dotted line—in a histogram (left) and a case-value plot (right). Both graphs are based on the same weight data, and, therefore, depict the same *distribution* of weight

Birth weight baby girls Queensland, Australia, December 18, 1997
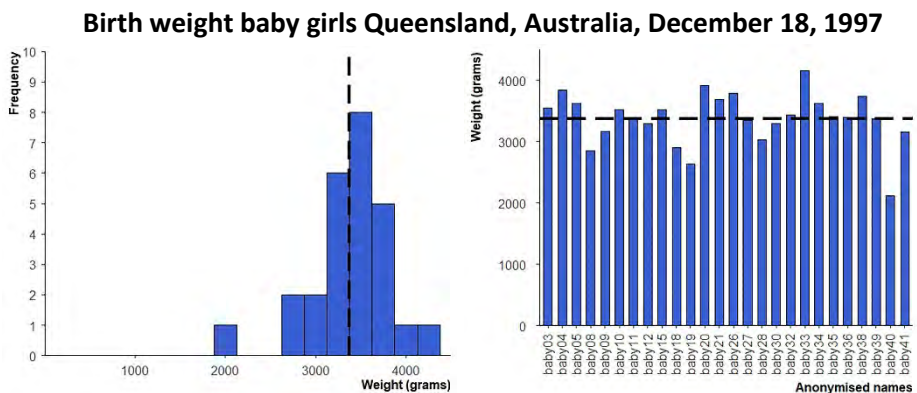
2.3   Method

A narrative systematic review of the literature with a configurative synthesis was conducted (Gough et al., 2017) with a query-based search strategy in the following databases: PsycINFO, Web of Science, Scopus, ERIC, and Google Scholar, see Figure 2.5 for the flowchart. These five databases are commonly used for scientific literature in mathematics and statistics education.

2.3.1   Search strategy

This chapter includes publications that describe or contain misinterpretations when constructing or interpreting histograms by people (students, teachers, researchers, and others). A publication was excluded when histograms were

Table 2.1

M		M					
=			) V	V	k		
U	"		)	V	o		
-	"		7	\	o		
O	#		8	h	u		
	)		U k@	h	†		

Note.

**

Education Literacy -Automated -Bank -Bayesian -Compression -Disability -DNA -Driver - Forensic -Genetic -MRI -Neural -Nuclear -Optics -Patient -Photon -Radiology - Segmentation -Spelling -Texture -Violence}. Exact reproduction of results in Google Scholar is not possible. While Google Scholar can lead to more included publications than other databases, some research might not be found when using Google Scholar alone (e.g., Haddaway et al., 2015). Therefore, Haddaway et al. advise to also use other databases.

Figure 2.5. Flow chart of publication selection process. When exclusion could not be decided on the basis of an abstract, the full text was studied
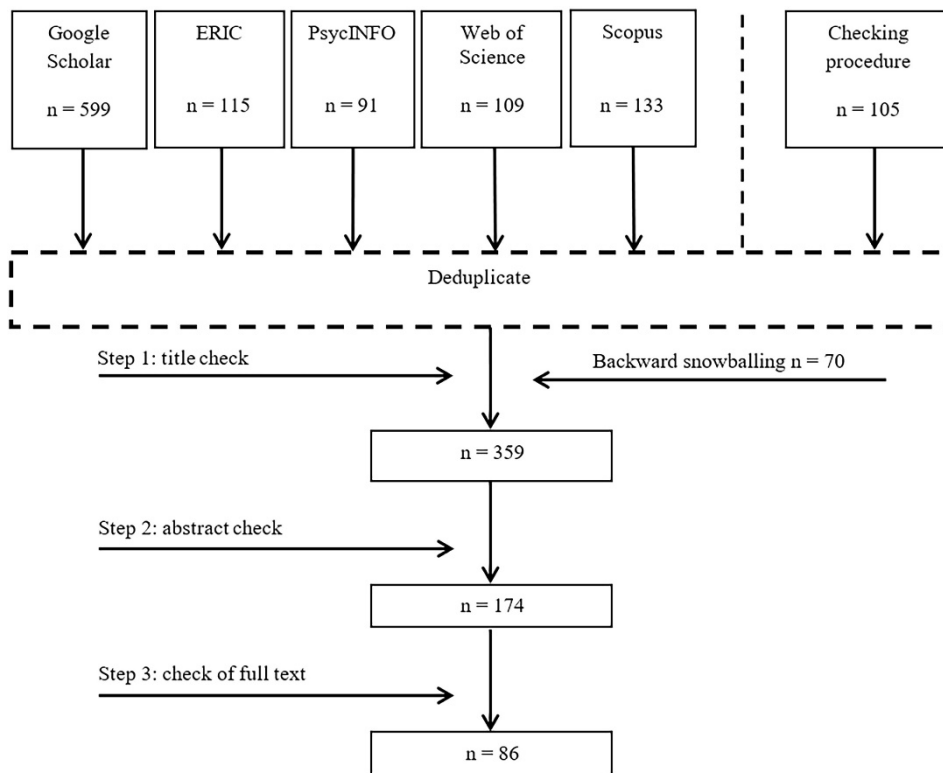
2.3.2 Data analysis

For every publication included in this review, we collected the misinterpretations that were either reported or detected in the publication. To identify the conceptual difficulties that become manifest in the most common misinterpretations, we grouped these misinterpretations into axial codes. Using the key statistical concepts as a lens, we inferred through abduction (Peirce, 1994) that misinterpretations stem from a lack or misunderstanding of these concepts. Abduction is the process of generating explanatory hypotheses. Hoffmann (2011) states that we can stop this process "when an

abductive insight has been achieved" which he defines as "the experience that what someone created in abductive reasoning" is plausible and gives an acceptable argument for the phenomenon (p. 572). As explained in section 2.2., the following holds. People who have fully grasped a key concept are not expected to show misinterpretations when drawing conclusions from graphs. When we identify a misinterpretation, we can, therefore, conclude that it is a manifestation of a conceptual difficulty.

Table 2.2 Example of some searching strategies, search terms, and number of identified and relevant publications in database PsycINFO in 2016. Since then, some changes in OVID databases have taken place including replacing or changing subject headings in the APA Thesaurus

Search in	Keywords for inclusion (search terms)	Number of publications identified	New relevant publications
Abstract	Histogram and mistake and education and literacy	0	0
	Histogram*	Almost 600, so more keywords were used	
All fields	Histogram and education**	Over 200, so more keywords were used	
	Histogram and education and literacy	8	1
Abstracts of predefined category 2240 statistics & mathematics	Histogram	40	0
Title	Histogram	43	0
Total		91	1

Note. Using four keywords for inclusion and none for exclusion led to zero publications identified, so the search strategy had to be slightly adapted by using fewer keywords. *Including a second keyword led to almost zero publications identified and no new relevant publications. **Other combinations were tried resulting in no new relevant publications.

How the network of statistical concepts was used is now explained with two examples. The first example is the misinterpretation of students who used two statistical variables when asked to draw a histogram (Baker et al., 2002). This misinterpretation is categorized as indicating a problem with understanding

the key concept of data (see box F in Figure 2.2: 1, 2 or more variables and attributes), as it indicates that these students do not differentiate between a histogram—which represents a univariate distribution of one variable—and a bivariate distribution of two statistical variables (the latter often being depicted in a scatterplot). A second example is students who do not understand that a distribution that looks unimodal in a histogram can turn out to be bimodal if the bin width is made smaller (Karagiannakis, 2013). This misinterpretation is categorized as indicating a problem with understanding the key concept of distribution, as it indicates that these students do not understand the influence of grouping on the graphical representation, which is displayed by the arrow from grouping (see box R in Figure 2.2: group or ungroup) to graphical representation (see box W in Figure 2.2: graphical representation: graph with bars, histogram). As further explained in the codebook (see Appendix A of this chapter for the full version), the selective code grouping was assigned here.

We used open, axial, and selective coding (Corbin & Strauss, 1990) to cluster the identified misinterpretations exhaustively and mutually exclusively into three categories: (1) data-related conceptual difficulties, (2) distribution-related conceptual difficulties, and (3) miscellaneous. Three examples of axial codes (a group of *misinterpretations*) are: 'larger frequency thus larger mean', 'bell-shaped = histogram', and 'bumpier = higher variability'. From these axial codes, the selective codes were created through abduction from the network of statistical concepts (see Figure 2.2). Provided with the codebook and the open codes (description of what was reported or found in the publication) and axial codes (the first grouping of the misinterpretations), an external coder was asked to assign one of eleven selective codes to the description of the misinterpretations. Of the more than 300 descriptions of misinterpretations (open codings), 73 were coded by the first author and an external coder. The interrater reliability—Cohen's kappa—was .84, suggesting a reliable coding procedure with "almost perfect" agreement (Landis & Koch, 1977, p. 165). A summary of the codebook is given in Table 2.3; a full version can be found in Table A.1 in the Appendix of this article.

The selective codes in the codebook categorize the misinterpretations at the level of a specific concept that were then merged into three categories of conceptual difficulties. At this final level, categories summarize whether the conceptual difficulties that become manifest in the misinterpretations are related to the data represented, or related to the distribution represented, or neither of these two (miscellaneous). The level of selective codes identifies subcategories of specific concepts that are misinterpreted. These subcategories are characterized briefly in the last column of the codebook and are illustrated with the types of misinterpretations listed. The characterization

ends with a note detailing when not to assign this code so as to make the second coder aware of the boundaries of a particular code (subcategory) (Boyatzis, 1998).

Some misinterpretations are possibly caused by the translation into English. In English, different words are created to distinguish histograms (one variable; numerical measurement level, see Figure 2.1, left) and distribution bar graphs (one variable; categorical measurement level, see Figure 2.1, right) from case-value plots (see Figure 2.3a; two variables) on the one hand and time-plots (also two variables) on the other. Other languages may lack such different words. Several researchers refer to a graph with bars as a histogram while it is not. If this misinterpretation was held by researchers from non-English-speaking countries, it might be due to translation only. Therefore, these specific misinterpretations were excluded from the results (Kramarski, 1999; Mevarech & Kramarsky, 1997).

Table 2.3 Summary of the codebook for classifying the misinterpretations; letters (e.g., K) refer to the network of statistical concepts

Phase	Conceptual difficulty	Selective codes
Orientation on histogram	Data-related	Number of variables (F) or measurement level (K) or both (K, F).
Interpreting histogram	Distribution-related	Variability (L, S), center (M, T), shape (W) and grouping (C)
	Miscellaneous	Context (A), Population (B), ICT[18] or unknown
In review but not included in results	Translation	

---

[18] ICT is found along the arrows from population to a sample. ICT is indicated only where relevant for this review.

2.4   Results

Table 2.4 Overview of publications in which misinterpretations were identified

Misinterpretations related to difficulties with the concept of data	Misinterpretations related to difficulties with the concept of distribution
Abrahamson & Wilensky, 2007; Agro, 1977; Baker et al., 2001, 2002; Bakker, 2004a; Bruno & Espinel, 2009; Capraro et al., 2005; Chance et al., 2004; Clayden & Croft, 1990; Cohen, 1996; Cooper & Shore, 2008; Corredor, 2008; Dabos, 2014; delMas et al., 2005; delMas et al., 2007; Derouet & Parzysz, 2016; Enders, 2013; Eshach & Schwartz, 2002; Friel & Bright, 1996; Gilmartin & Rex, 2000; Hawkins, 1997; Humphrey et al., 2014; Ismail & Chan, 2015; Kaplan et al., 2014; Kramarski, 2004; Kulm et al., 2005; C. Lee & Meletiou-Mavrotheris, 2003; Lem et al., 2013c; McKinney, 2015; Meletiou & Lee, 2002; Meletiou, 2000; Redfern, 2011; Ruiz-Primo et al., 1999; Sorto, 2004; Stevens & Palocsay, 2012; Stone, 2006; Strasser, 2007; Tiefenbruck, 2007; Watts et al., 2016; Whitaker & Jacobbe, 2017; Wong, 2009; Yun, Ko, & Yoo, 2016; Zaidan et al., 2012.	Baker et al., 2001; Batanero et al., 2004; Biehler, 1997; Bruno & Espinel, 2009; Capraro et al., 2005; Chan & Ismail, 2013; Chance et al., 2004; Cohen, 1996; Cooper & Shore, 2008; Cooper & Shore, 2010; Corredor, 2008; Dabos, 2014; delMas & Liu, 2005; delMas et al., 2005; delMas et al., 2007; Derouet & Parzysz, 2016; Friel & Bright, 1995, 1996; Gilmartin & Rex, 2000; González, 2014; Huck, 2016; Ismail & Chan, 2015; Kaplan et al., 2014; Kaplan et al., 2009; Karagiannakis, 2013; Kelly et al., 1997; Konold et al., 1997; Kukliansky, 2016; Kulm et al., 2005; Lee & Meletiou-Mavrotheris, 2003; J. T. Lee, 1999; Lem et al., 2011, 2013a, 2013c, 2014b, Madden, 2008; Martin, 2003; McGatha et al., 2002; McKinney, 2015; Meletiou & Lee, 2002; Meletiou, 2000; Meletiou-Mavrotheris & Lee, 2005; Mevarech & Kramarski, 1997; Olande, 2014; Roth, 2005; Rumsey, 2002; Sorto, 2004; Stevens & Palocsay, 2012; Stone, 2006; Tiefenbruck, 2007; Turegun & Reeder, 2011; Vermette & Gattuso, 2014; Whitaker & Jacobbe, 2017; Whitaker et al., 2015; Wong, 2009.
Misinterpretations related to miscellaneous concepts	Language or translation
Abrahamson, 2006, 2008, 2009; Abrahamson & Cendak, 2006; Abrahamson & Wilensky, 2007; Baker et al., 2001; Behrens, 1997; Biehler, 1997; Carrión & Espinel, 2006; Chance et al., 2004; Cohen, 1996; delMas et al., 2005; 2007; Friel et al., 2001; Hawkins, 1997; Kaplan et al., 2014; Konold et al., 1997; Madden, 2008; McKinney, 2015; Nuhfer et al., 2016; Prodromou & Pratt, 2006; Shaughnessy, 2007; Slauson, 2008; Stone, 2006; Whitaker & Jacobbe, 2017; Whitaker et al., 2015; Yun & Yoo, 2011; Yun et al., 2016.	Kramarski, 1999; Mevarech & Kramarsky, 1997

The results show that the conceptual difficulties that become manifest in the most frequently reported misinterpretations fall into three different categories: data-related, distribution-related and other. The misinterpretations that are a manifestation of difficulties with the concept of data include: not understanding how many statistical variables are depicted in a histogram (only one) and not understanding that a histogram is suitable for numeric variables only (see Figure 2.2 F and L). The misinterpretations that are a manifestation of difficulties with the concept of distribution include: (a) not knowing how shape (part of W, Figure 2.2), center (see Figure 2.2 M and T), and variability (see Figure 2.2 L and S) are depicted in a histogram and (b) not understanding the effect of grouping into bins in a histogram (see Figure 2.2 R). In line with Bakker and Hoffmann (2005), our research shows that these two conceptual difficulties cannot be isolated from their sign—the histogram. The third category of miscellaneous conceptual difficulties is more loosely related to the sign—the histogram—and entails difficulties that occur due to the software used, and/or confusion about whether the sample or the population is depicted in the histogram, and the context. The most common misinterpretations resulting from these conceptual difficulties are elaborated further in the next sections. Table 2.4 gives an overview of the publications included in this review. The full details of all misinterpretations can be found in the data paper (Boels et al., 2023) and more summaries of the findings are given in the online extra materials. The misinterpretations described or detected in the publications—including almost 16,000 students, teachers, and researchers—are incorporated in this review. This includes slightly over 400 elementary school students, almost 7,000 secondary school students, and approximately 8,000 college and university students. The remainder includes college statistics teachers, mathematics teachers, and researchers. Most participants are from the USA (see Appendix A of this chapter).

2.4.1 Misinterpretations related to difficulties with the concept of data

Identifying the measured variable only
As explained in the theoretical background, by definition a histogram displays the distribution of one statistical variable[19]. Twenty-five publications reported or showed misinterpretations regarding the measured variable. A widespread misinterpretation is that a histogram could display the data of two variables, which was reported or found in nine sources (e.g., Cohen, 1996; Gilmartin &

---

[19] Some statistics educators prefer the more general term of 'attribute' (W. Finzer, personal communication, July 12th, 2018). As other people may think that attribute only refers to a nominal measurement level, we avoided this term.

Rex, 2000; Meletiou & Lee, 2002; Meletiou, 2000; Stevens & Palocsay, 2012; Zaidan et al., 2012) and which is related to the misinterpretation that the number of bars is seen as the number of cases (Dabos, 2014; Ismail & Chan, 2015; Sorto, 2004). Another often-found misinterpretation is that the frequency is seen as the measured value (Bakker, 2004a; Chance et al., 2004; delMas & Liu, 2005; Friel & Bright, 1996; Kaplan et al., 2014; Lem et al., 2013c) and that the horizontal axis is seen as a timescale when it is not (Dabos, 2014; Kaplan et al., 2014; Meletiou & Lee, 2002; Meletiou, 2000; Zaidan et al., 2012). This confusion is aggravated as frequency and number (count) are commonly interchangeable terms[20]. The definition of a histogram nevertheless implies that the vertical axis depicts the frequencies or number counts of the measured values that are depicted on the horizontal axis. Consequently, a time-plot—with, for example, years on the horizontal axis—is not a histogram, as it is nonsensical to count how often a year occurs in a year. Furthermore, it is often stated that the bars of a time-plot must be connected when intervals are consecutive, but this is only true for histograms[21].

Figure 2.6 Case-value plot or time-plot with two statistical variables (year and number of unemployed). Data source: Statistics Netherlands (CBS, 2018)
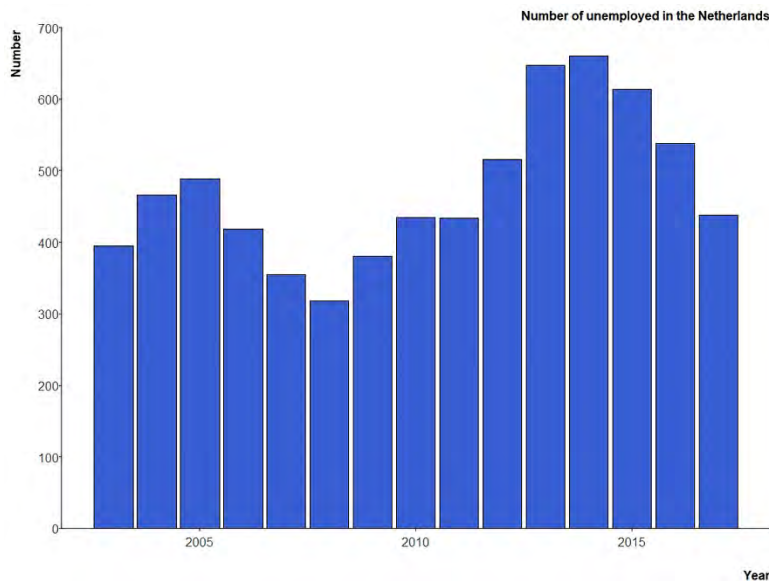
Note. Many people incorrectly think this graph is a histogram because the variable on the horizontal axis is numerical. In such cases, connected bars are often—mistakenly—used.

---

[20] See also our footnote in the introduction on the influence of language on the interpretation of the term frequency.

[21] Some researchers also use separate bars in histograms, e.g., Ioannidis (2003).

Identifying the measurement level only

Eighteen publications reported or contained misinterpretations regarding the measurement level. Five of these publications reported people referring to a normal distribution—which is only possible for numerical data—while the measurement level of the data was nominal or ordinal (delMas et al., 2007; Humphrey et al., 2014; Kaplan et al., 2014; Redfern, 2011; Whitaker & Jacobbe, 2017). Nine publications reported or contained 'histograms' with nominal or ordinal measurement level (e.g., Stone, 2006; Tiefenbruck, 2007; Watts et al., 2016; Wong, 2009). People showing this misinterpretation may consider the blood type graph (see Figure 2.1) as 'right skewed' or 'not normally distributed'. These people overlook that the measurement level is nominal, and, therefore, the bars are not in scale order and the theoretical model of a normal distribution is, therefore, not applicable.

Three publications identified the misinterpretation that the interval is a 'label' with, for example, students and authors of schoolbooks treating this label as a nominal measurement level, neglecting the numerical scale (Bruno & Espinel, 2009; Derouet & Parzysz, 2016; Humphrey et al., 2014).

Figure 2.7 Example of incorrect 'histogram' with labeled bars (data from Fisher, 1947)
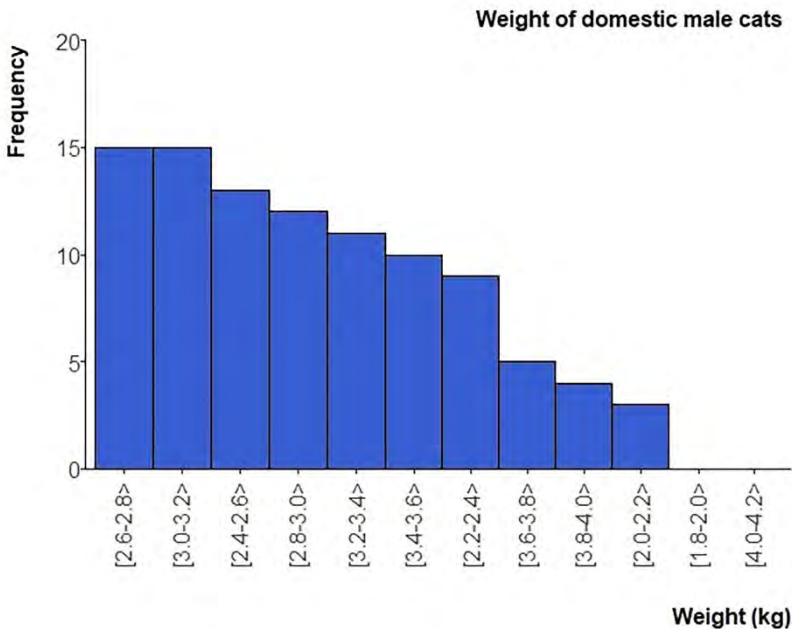
Another misinterpretation is the use of histograms for Likert scales when words combined with numbers are used. An example of how seriously this can go wrong when used by non-statisticians can be found in McKinney (2015)

where the following strange[22] attribution for a 5-point Likert scale is used: none at all (1), very little (2), strong degree (3), quite a bit (4) and a great deal (5) for a Self-Efficacy Scale for Teaching Mathematics Instrument (SETMI). This SETMI was developed by McGee (2012) and is used in several other studies (e.g., McCampbell, 2014).

Identifying the measured variable and the measurement level

Seventeen publications reported or contained misinterpretations regarding both the number of variables and the measurement level. The most often reported or found misinterpretation (10 publications) is that people think that there is no difference between a histogram and a bar graph, or that the only difference is that bars are connected in a histogram, neglecting the required measurement level (Capraro et al., 2005; Clayden & Croft, 1990; Eshach & Schwartz, 2002; Gilmartin & Rex, 2000; Humphrey et al., 2014; Kramarski, 2004; Kulm et al., 2005; Sorto, 2004; Stevens & Palocsay, 2012; Tiefenbruck, 2007). Six publications contained or reported the misinterpretation that a histogram could be used for nominal or ordinal data and two variables (Baker et al., 2001, 2002; Dabos, 2014; delMas & Liu, 2005; Eshach & Schwartz, 2002; Ruiz-Primo et al., 1999). Four publications reported the misinterpretation that bars could be rearranged in a histogram, for example, from highest to lowest bar (Dabos, 2014; Humphrey et al., 2014; Kaplan et al., 2014; Whitaker & Jacobbe, 2017).

2.4.2 Misinterpretations related to difficulties with the concept of distribution

As explained in the Theoretical background section, the number of measured variables as well as the measurement level define the type of graphical representation, which in turn influences the interpretation of the distribution: shape, center, and variability. For example, variability can be seen as weighted deviation from the arithmetical mean (Cooper & Shore, 2010). In a case-value plot with nominal data on the horizontal and numerical data on the vertical (two measured variables), the relevant measured value is on the vertical axis and variability can be seen as variation in the heights of the bars. In a histogram, the only measured value is on the horizontal axis, and, therefore, the horizontal spread of these measurements must be considered—in combination with the heights of the bars. Several studies report that students

---

[22] For instance, strong degree (3) is in the middle of the scale so it should be a more neutral word, such as undecided. Furthermore, strong degree (3) and a great deal (5) seem synonyms; quite a bit (4) seems a bit less strong than strong degree. This scale does not even seem to be ordinal, but rather nominal and therefore, a histogram is inappropriate (and calculating a mean is nonsensical).

and teachers confuse variation in the frequencies in a histogram—the heights of the bars—with variation in the measured value—hence, the variability in a histogram (e.g., Lem et al., 2013a). In this section, four groups of misinterpretations are reported regarding variability, center, shape, and information reduction through grouping.

Variability

Twenty-six publications reported on misinterpretations regarding the statistical concept variability or regarding the variability combined with the statistical concept's center and/or shape. Eleven publications reported the misinterpretation that a higher difference in the heights of the bars only implies more variation in the data (Cooper & Shore, 2008; delMas et al., 2007; Kaplan et al., 2014; Karagiannakis, 2013; Madden, 2008; Meletiou & Lee, 2002; Meletiou, 2000; Meletiou-Mavrotheris & Lee, 2005; Olande, 2014; Stone, 2006; Vermette & Gattuso, 2014). Eight publications reported an overgeneralization of the idea that a certain shape (normal, uniform, or symmetrical distribution) has the highest or lowest variability (Cooper & Shore, 2010; Dabos, 2014; González, 2014; Kaplan et al., 2014; Meletiou-Mavrotheris & Lee, 2005; Turegun & Reeder, 2011; Vermette & Gattuso, 2014; Whitaker & Jacobbe, 2017). Range can be regarded as a simple or preliminary measure of variability, especially for secondary school students. Seven publications reported misinterpretations of the variability in the data when range was used (Cooper & Shore, 2008; Dabos, 2014; Kaplan et al., 2014; Lem et al., 2013c; Madden, 2008; Meletiou-Mavrotheris & Lee, 2005; Olande, 2014) and two reported misinterpretations about variability and center when range was used (Kukliansky, 2016; Lem et al., 2013a). Various misinterpretations regarding the standard deviation in a histogram are reported, including that a certain shape or ordering of the bars (e.g., ascending or descending heights) leads to the largest or smallest standard deviation, that a larger mean implies a larger standard deviation and that gaps between bars (frequency zero) do not influence the standard deviation (delMas & Liu, 2005). Others found the misinterpretation that standard deviation and mean in a histogram are the same (Chan & Ismail, 2013) or that once the means in both histograms are the same, the standard deviation is the same as well (Kukliansky, 2016). Misinterpretations regarding variability are also found among teachers (e.g., González, 2014). Variability is the variation of the data, for example, around the mean—see Figure 2.8. As the mean is depicted differently in a histogram than in a case-value plot, the variability also has to be assessed differently. In a case-value plot, the variability is the variation in the heights of the bars. In a histogram, the variability is the weighted horizontal spread of the bars.

Figure 2.8 Center and thus variability is assessed differently in a histogram (left) compared to a case-value plot (right)
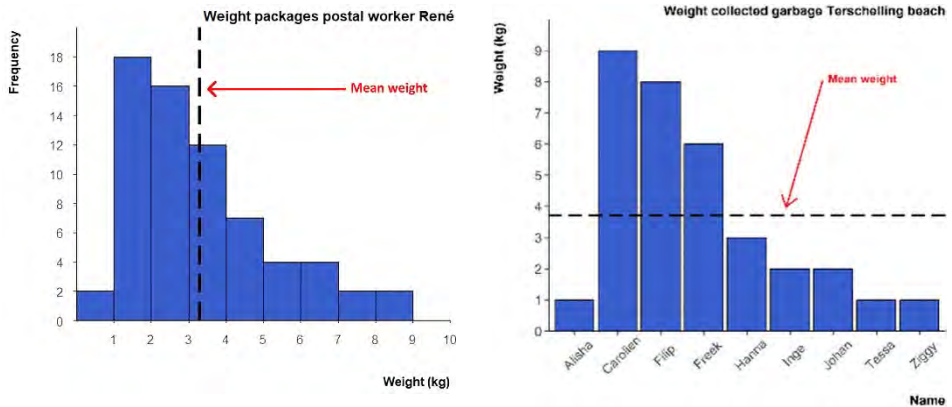
Center

Thirteen publications reported on misinterpretations regarding the statistical concept of center. Four publications reported a misinterpretation where the mean of the frequencies (vertical axis) was used instead of the mean of the measured values of the variable (horizontal axis, see Figure 2.8) (Cooper & Shore, 2008; Lem et al., 2013a, 2013c, 2014b). Five publications reported a similar misinterpretation regarding the median (Cooper & Shore, 2008; Ismail & Chan, 2015; Kaplan et al., 2014; Lem et al., 2013a), the mode (Huck, 2016; Ismail & Chan, 2015; Kaplan et al., 2014) or both (Kaplan et al., 2014). All these misinterpretations are related to the type of graphical representation, as whether the frequency is a statistical variable or not depends on the type of graph. For example, in a time-plot, the frequency is the measured value. Other misinterpretations include that the median is seen as the middle class (Stevens & Palocsay, 2012), that it is seen as the midpoint of the scale on the horizontal axis, or as the midrange (Cooper & Shore, 2008).

In many Introductory Statistics courses, rules of thumb are taught for the position of mean and median in relation to the skewness of the distribution (thus the shape in the histogram). One such rule of thumb is that the mean is typically lower than the median in left or negatively skewed distributions. Although this holds true in many situations, Huck (2016) states that this was helpful when people lacked strong computers, but nowadays these kinds of rules are no longer needed as they can also mislead us when analyzing results. Huck claims that "Unfortunately, the application of those rules can make one think data are skewed left when they are skewed right (or vice versa)." (p. 26). Therefore, we carefully need to reconsider questions that test, for example, if students know the rule of thumb that the mean is bigger

than the median in right-skewed distributions (Cooper & Shore, 2008; delMas et al., 2007; Karagiannakis, 2013; Lee & Meletiou-Mavrotheris, 2003; Whitaker & Jacobbe, 2017).

Information reduction through grouping

People have difficulties with the information reduction (Gal & Garfield, 1997) present in histograms. As explained in the theoretical background, one step in information reduction is that several values are grouped into one bin. Bakker (2004a) already pointed out that this grouping is difficult for students in Grades 7 and 8. Fifteen publications reported or contained misinterpretations regarding the grouping in bins. Misinterpretations include not using or mentioning density for unequal bin width (Derouet & Parzysz, 2016; Gilmartin & Rex, 2000; Kelly et al., 1997; McGatha et al., 2002) and choosing a wrong bin width or wrong boundaries for the bins (Bruno & Espinel, 2009; delMas et al., 2005; Martin, 2003; McKinney, 2015; Whitaker & Jacobbe, 2017). Three publications reported misinterpretations regarding the measured values, either that all possible values in a bin are measured (Lem et al., 2013c; Meletiou, 2000) or that only the middle value of a bar is measured (Biehler, 1997).

Shape

Twenty-eight publications reported or contained misinterpretations concerning the graphical representation of a histogram itself. Six reported that students cannot link a histogram to a corresponding boxplot (Corredor, 2008; delMas et al., 2005; delMas et al., 2007; Karagiannakis, 2013; Lem et al., 2011, 2015). Ten reported or contained misinterpretations regarding graph conventions (Baker et al., 2001; Batanero et al., 2004; Bruno & Espinel, 2009; Lem et al., 2013c; Martin, 2003; McGatha et al., 2002; Mevarech & Kramarsky, 1997; Roth, 2005), for example, that connected bars are for easier comparison (Capraro et al., 2005; Kulm et al., 2005). Some authors state that histograms are not suitable for discrete variables (Batanero et al., 2004; Cohen, 1996; Friel & Bright, 1995, 1996; Tiefenbruck, 2007). However, data are always discrete due to the accuracy of the measurement instrument. Therefore, we decided not to exclude discrete variables. Students using graphs with poles instead of bars can be found in McGatha et al. (2002).

2.4.3 Misinterpretation related to miscellaneous concepts

In addition to the two aforementioned categories, there are less frequent miscellaneous difficulties that can be summarized as: not understanding the histogram in relation to the given context, not understanding the difference between a histogram of a sample and a histogram of a population, and the influence of ICT (ICT often does not differentiate between histograms and

other types of graphs with bars[23]; e.g., see Abrahamson, 2006 for an example). Some descriptions in publications do not provide enough details for specifying the type of misinterpretation and are classified as unknown (Baker et al., 2001; Behrens, 1997; Biehler, 1997; Carrión & Espinel, 2006; Chance et al., 2004; Konold et al., 1997; Shaughnessy, 2007; Yun & Yoo, 2011).

Context

Nine publications reported misinterpretations due to the context. One example of a misleading context is height (Whitaker & Jacobbe, 2017), as students in this specific context more easily interpret the height of the bars in a histogram as the measured height, leading to the confusion of a case-value plot with a histogram. The misinterpretation of a time scale on the horizontal axis can sometimes also stem from the context and is described in section Identifying the measured variable only. Furthermore, students and teachers occasionally use context knowledge or personal experience instead of the data (Friel et al., 2001; Madden, 2008; Shaughnessy, 2007). The opposite equally occurs where students have trouble linking the histogram to the original data collection or context (delMas et al., 2005; Yun & Yoo, 2011). This is in line with research from Kaplan et al., (2018) who showed that students' descriptions of histograms systematically differ depending on the specific wording of the question (including the word distribution or variable or both in the question) as well as the context (income or hours of sleep).

Sample or population?

Seven publications reported misinterpretations regarding the population. Five of these report the misinterpretation that the histogram of a sample and the histogram of a population have the same properties—for example, the same shape or distribution (Chance et al., 2004; Hawkins, 1997; Slauson, 2008; Stone, 2006; Whitaker & Jacobbe, 2017). Not distinguishing between sample and population might also lead to ignoring the effect of random noise (Biehler, 1997; Nuhfer et al., 2016).

Influence of ICT

Although ICT can be a helpful tool to understand statistics, it can also introduce new misinterpretations. The most common misinterpretation is embedded in the software where no distinction is made between a histogram and a bar graph (Hawkins, 1997), often leading to histograms with strange or even wrong

---

[23] Excel, for example, creates a kind of bar chart with intervals below, instead of a histogram (see https://trumpexcel.com/histogram-in-excel/#Creating-a-Histogram-using-Data-Analysis-Toolpak for an incorrect example of a 'histogram' with unequal bin widths). Although this was more prominent in older versions, the way Excel handles unequal bin widths or values that are higher or lower than the specified categories, is not correct, and more in line with how bar charts would be created.

boundaries of the bins (Abrahamson, 2006, 2008, 2009; Abrahamson & Cendak, 2006; Abrahamson & Wilensky, 2007; McKinney, 2015; Prodromou & Pratt, 2006). Two publications reported the misinterpretation that the number of classes is fixed, possibly due to a fixed number of classes in the software (Cohen, 1996; Yun et al., 2016).

2.5   Conclusions and discussion

In this review, the aim was to make a systematic inventory of the misinterpretations that occur when people use histograms, as well as to categorize these misinterpretations along the conceptual difficulties that become manifest in them. It turned out that the most common conceptual difficulties could be related to two key concepts in statistics: data and distribution. The category misinterpretations that are related to the difficulties with the key concept of data includes misinterpretations about the number of variables depicted in a histogram and the measurement level of the data, including the wrong application of theoretical models. The category of misinterpretations that are related to difficulties with the key concept of distribution includes misinterpretations about variability, center, shape, and information reduction through grouping. The third and more diverse category of misinterpretations is related to other conceptual difficulties and includes having trouble linking the context to the histogram, not understanding the difference between a histogram of a sample and of a population, and the influence of ICT. The analysis of the publications in our review also led to the identification of a network of statistical concepts specific to interpreting histograms, see the theoretical background section. From our analysis, it furthermore became clear that two statistical concepts needed to be added to the key concept of *data*: number of variables and measurement level. These two concepts were not yet explicitly part of the collection of key concepts in statistics.

Furthermore, our review study reveals that most publications investigate students' or teachers' notions of shape and variability, which is an important topic for college and university students. Hence, these publications focus on misinterpretations that are related to difficulties with the key concept of distribution. Although misinterpretations regarding identifying the number of variables and the measurement level of their attributes are more often observed, research specifically addressing these misinterpretations is scarce. The latter two sub-categories of misinterpretations are related to difficulties with the key concept of data. The data-related conceptual difficulties may be underlying the distribution-related conceptual difficulties, as the data (number of statistical variables and measurement level) define the type of graph, and in

turn how variability and center are depicted in the graphical representation (e.g., Cooper & Shore, 2010). We speculate that the persistence of people's misinterpretations of histograms is partly due to overlooking the impact of data-related conceptual difficulties. This might also result in the underreporting of misinterpretations regarding data-related conceptual difficulties, as well as misinterpretations regarding shape and center.

Our findings are in line with findings about mathematical graphs from Leinhardt et al. (1990), such as the tendency to overgeneralize. An example of overgeneralization is the idea that the number of axes is the number of measured variables (true for a case-value plot, but false for a histogram; see the section Identifying the measured variable only). Another example is the overgeneralization of the effect of shape (e.g., uniform distribution) on variability (see the section Variability) and of theoretical models (normal distribution, see the section Identifying the measurement level only). Leinhardt et al. also found interference with the context or daily life observations (see the section Context).

According to Friel et al. (2001), the basic level of reading the data is often not very difficult for students for most graphs. This may be true for reading off a particular value, but our review shows that many misinterpretations are related to the data depicted in a histogram, hence, to reading the data (thus the key concept of data). In addition, during the application of the theoretical framework of statistical key concepts, it became clear that not only are the statistical concepts important, but also the connections between them, such as, for example, that grouping in bins influences the shape of the distribution, thus the graphical representation of the data. We, therefore, proposed a coherent network of statistical concepts relevant to research questions that may involve the interpretation of histograms (see the section Data analysis).

Systematic reviews of the literature have limitations. A geographical selection bias seems to exist. A large proportion of the studies in this literature review was carried out in the United States, followed by European countries (see the Appendix). The English-speaking countries generally pay more attention to statistics in their curriculum than other countries (e.g., Franklin, 2019). This suggests that the problem may be bigger than what was found here. We do not want to suggest representativeness, as we were mainly interested in the types of conceptual difficulties that become manifest when people (students, teachers, researchers, and others) interpret histograms.

Furthermore, we speculate that the misinterpretations identified in this literature review also hold for Asian, African, and South American countries, as well as for Australia. The reasons for this speculation are that in some countries statistics is not yet or only recently part of the curriculum, for

example in Thailand (e.g., Burrill & Ben-Zvi, 2019; Franklin, 2019; González & Chitmun, 2019), and there are some, although not yet many, studies of Asian countries indicating misinterpretations when interpreting histograms (Ismail & Chan, 2015; Yun & Yoo, 2011; Yun et al., 2016).

Several implications for future research and education arise from this review. The first is that data-related conceptual difficulties seem to be understudied, and, therefore, would require more explicit attention from researchers. Ignoring the difficulties with the concept of data may possibly explain the persistence of misinterpreting histograms. Researchers, teachers, and teacher trainers are encouraged to be more aware of the differences regarding distribution and data (number of variables, measurement level), the differences between a case-value plot, a distribution bar graph, and a histogram, and the consequences of these differences for shape as well as for assessing variability and center. Furthermore, in languages that lack distinct words for case-value plots, distribution bar graphs, and histograms, our suggestion is to create and introduce those words and implement them in the statistics education curriculum from elementary school level up to the university level, as this will support the awareness of the differences. In addition, this literature review adds to the framework of key concepts in statistics education that there is a hierarchy in those key concepts. The key concept data (number of variables and measurement level) is fundamental for a deep understanding of the key concept distribution as shape, center, and variability are depicted differently in different types of graphical representations.

The second implication is that the role of information reduction seems to be understudied (see the section Theoretical background). The literature on information reduction is very scarce. Bakker (2004a) is one of the few examples indicating the difficulty of the idea of grouping. Nevertheless, indications for this difficulty are also found in other research (Ismail & Chan, 2015; Lem et al., 2013c; Meletiou, 2000; Sorto, 2004). Researchers, teachers, and teacher trainers are advised to be aware that information reduction plays an important role in the following four stages when turning a case-value plot into a histogram. The first stage is when one of the measured variables is removed (resulting in, for example, a dotplot). Students who see the data in a graph as a pointer to the situation (Konold et al., 2015) and students who consider a histogram as a case-value plot might not have understood the case information removal phase. The second phase is when the dots in a dotplot are stacked into classes with a certain bin width. People who think that only the middle value of a bar is observed might not have understood this grouping phase. The third phase is when the dots are omitted from the bar, making it compulsory to use a second axis when absolute frequencies are used. People

who regard a histogram as a bivariate distribution might have problems with this third phase. The fourth phase, which is hardly studied, is when the frequency is turned into frequency density. This phase is of key importance for the transition to continuous probability distributions. The other two understudied areas are the difference between the histogram of a sample and of a population and the influence of the context. These two areas are only loosely related to histograms.

The third implication is that future research is needed in those countries that are not yet included in our review to substantiate our claim that the identified conceptual difficulties can be found all over the world and are not due to a specific way of teaching or an educational system, as a geographical gap seems to exist in the research literature. Also, the active promotion of publishing in English journals of work published earlier in other languages is needed to make this literature available for many more researchers, as well as the translation of English literature into other languages.

As an implication for task design in research and education, this review makes it clear that items containing graphs with bars without context or labels cannot be identified with regard to the type of graph and must be avoided in schoolbooks as well as assessments and research items. Furthermore, for languages that lack different words for different types of graphs with bars, the advice is to create such words—and use these in education as well as research—to distinguish histograms, distribution bar graphs, case-value plots, and time-plots.

An implication for education—now that the conceptual difficulties that become manifest in the most common misinterpretations are made plausible—is that researchers and educators can address these more broadly rather than treating or remediating misinterpretations one at a time. Such a didactical itinerary would be consistent with the current view in statistics education which aims for students to develop an understanding of the key concepts of statistics and their interrelationship. Our overview opens up the possibility of systematically dealing with these misinterpretations first in research and eventually in elementary and secondary schools and statistics introductory courses, as well as developing and testing materials specially designed to tackle these misinterpretations. Teachers and teacher trainers now have access to an overview of all the common misinterpretations identified in the publications. This adds to their Statistical Knowledge for Teaching (SKT, see Groth, 2007). According to Pareja Roblin et al. (2018), an overview is very important as "positive student outcomes were associated with curriculum materials […] that provide teachers with information about students' ideas" (p. 260).

One might conclude that histograms are too difficult to use and teach. Can we do without them in education and research? Our answer is no. First, histograms reveal some aspects of the distribution that other graphs do not (e.g., Pastore et al., 2017). Second, histograms are omnipresent in research and education and should, therefore, be learned. Third, the alternatives entail some of the same disadvantages such as the height misinterpretation in dotplots (Lyford, 2017), as well as other disadvantages such as an irregular shape (dotplots) or an even more advanced step in information reduction (boxplots; Lem et al., 2014a). Fourth, it is the key concepts underlying a histogram that are hard to grasp (the key concepts of data and distribution). Unfortunately, we cannot learn those key concepts without signs (e.g., histogram), as the representation of the data as well as how the distribution manifests itself (through its shape) strongly depends on the specific type of graph with bars, as we explained in the theoretical background section. It is when interpreting histograms that these underlying conceptual difficulties become manifest, making histograms a good diagnostic instrument for teachers and researchers as well.

Appendix A Codebook, samples, and misinterpretations

In this Appendix, the provenance of participants included in the review, an overview of the identified misinterpretations (axial codes), and the full codebook can be found. This is supplementary to Chapter 2 of this dissertation and is published on the website of Educational Research Review.

A.1   Provenance of participants of studies included in this review

The participants of studies included in the publications in this review come from the United States of America (12,959) followed by Europe (1572). Asia (1298) and South America (84) are relatively underrepresented (see Figure A.1); no African studies were found during the search. When the misinterpretation was in the publication itself or when the number of participants was not given, the count was set to zero.

Figure A.1 Spread of the participants included in the studies in this review
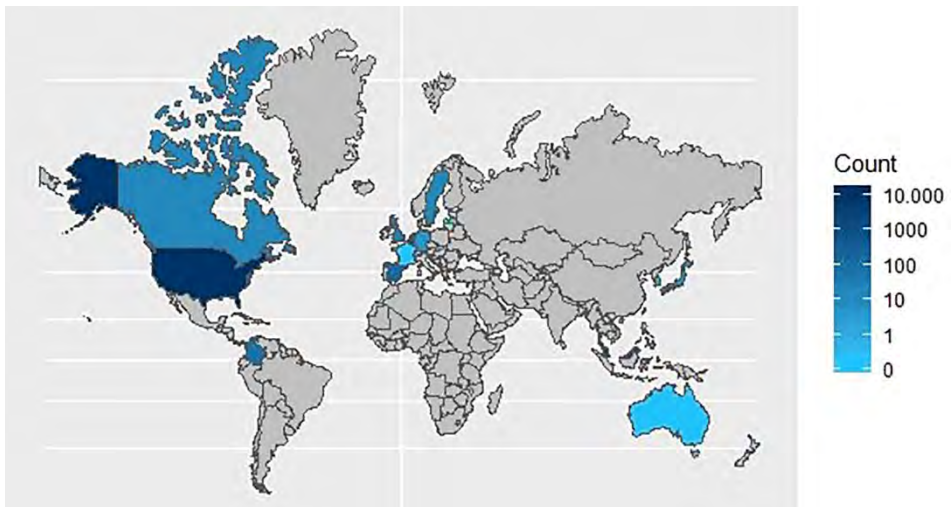
A.2   Samples included in the publications

Table A.1 Publications included in this review as well as number of participants

Publication (shorted reference)	Number of participants[24]
Abrahamson_2006_01	0
Abrahamson_2006_02	0
Abrahamson_2007	0
Abrahamson_2008	0
Abrahamson_2009	2
Agro_1977	0
Baker_2001	52
Baker_2002	12
Bakker_2004a	580
Batanero_2004	117
Behrens_1997	0
Biehler_1997	4
Bruno_2009	29
Capraro_2005	134
Carrion_Perez_2006	0
Chan_2013	412
Chance_2004	0
Clayden_1990	18
Cohen_1996	0
Cooper_2008	186
Cooper_2010	0
Corredor_2008	84
Dabos_2014	52
delMas_2005_01	12
delMas_2005_02	542
delMas_2007	763
Derouet_2016	0
Enders_2013	80
Eshach_2002	10
Friel_1995	76
Friel_1996	76
Friel_2001	0
Gilmartin_2000	0
González_2014_01	4
Hawkins_1997	0
Huck_2016	0
Humphrey_2014	0

---

[24] Zero students either indicates that no numbers were given (often) or that this was not relevant (sometimes) as—for example—the misinterpretation was in the ICT used in this publication.

Publication (shorted reference)	Number of participants[24]
Ismail_2015	412
Kaplan_2009	67
Kaplan_2014	341
Karagiannakis_2013	9
Kelly_1997	25
Konold_1997	4
Konold_2015	15
Kramarski_1999	0
Kramarski_2004	0
Kukliansky_2016	256
Kulm_2005	134
Lee_1999	0
Lee_2003	162
Lem_2011	167
Lem_2013a	167
Lem_2013c	125
Lem_2014	114
Lem_2015	188
Madden_2008	56
Martin_2003	0
McGatha_2002	24
McKinney_2015	0
Meletiou_2000	33
Meletiou_2002	33
Meletiou_2005	35
Mevarech_1997	92
Nuhfer_2016	0
Olande_2014	13
Prodromou_2006	6
Redfern_2011	0
Roth_2005	1
Ruiz_1999	0
Rumsey_2002	0
Shaughnessy_2007	0
Slauson_2008	53
Sorto_2004	42
Stevens_2012	4727
Stone_2006	0
Strasser_2007	0
Tiefenbruck_2007	0
Turegun_2011	41
Vermette_2014	12
Watts_2016	0
Whitaker_2015	3324

Publication (shorted reference)	Number of participants[24]
Whitaker_2017	1881
Wong_2009	0
Yun_2011	0
Yun_2016	0
Zaidan_2012	122
Total number of participants	15926

Table A.2 Number of publications and participants per type of school

Type of school	Number of publications	Number of participants
college	5	418
college and university	2	763
middle school	7	46
elementary school	8	420
elementary and secondary school	1	15
secondary school	12	6763
secondary school and college	1	542
university	25	6757
unknown or unclear	3	29
work (teachers or researchers)	5	169
n.a.	17	4
Total	86	15926

A.3   Overview of axial codes

The table below gives an overview of the axial codes that were informed by the open codes (description of the misinterpretations). Some of the axial codes might only make sense to readers if they are combined with the description of the misinterpretations. Readers interested in specific axial codes are, therefore, referred to the first author.

Table A.3 Axial codes used in the review

Axial codes	Axial codes
all values = middle bar	mean = st.dev.
area	mean > median left skewed
ascending/descending order = smallest st.dev[25].	measure of variability
	median = frequency of mode
bar = observed value	median = median of frequency or scale

---

[25] st.dev. is standard deviation

Axial codes	Axial codes
bar --> all values observed	median = middle class
bars can be reordered	median = midpoint hor. axis
bars in middle = smallest st.dev.	median = midrange
bell-shaped = histogram	miscellaneous
bell-shaped histogram = higher variability	modality cannot change
bell-shaped histogram = lower variability	mode = highest frequency
bumpier = higher variability	more bars = higher variability
bumpier = lower/higher variability	no influence gaps on st.dev.
choice of graph	nominal measurement level
choosing wrong variable	not include effect random noise
combining 2 histograms	not noting difference in range
comparing sample(s) and population(s)	number bars = number cases
connected bars	ordinal Likert scale
context	ordinal scale
density	outliers
depicted variables	random = bell-shape
discrete = bar graph	range
effect bin width	range = variability
evenly spread	relating graphs
fixed number of classes	same frequency thus same st.dev.
frequency = measured value	same mean thus same st.dev.
frequency on vertical then always histogram	same range and shape thus same median
graph convention	sample = population
higher bars, not higher average and st.dev.	shape doesn't influence variability
	skewness
histogram = bar graph	smallest st.dev. Is not smallest spread from center
histogram = dot plot	
histogram = scatterplot	standardization = normality
histogram = stem & leaf	statistical language
histogram --> boxplot	swap of axis
histogram 2 variables	symmetrical histogram = lower variability
hor. axis = data	
horizontal = time	table preference
label bars	total n is not sum frequency
larger frequency thus larger mean	uniform shape = highest variability
larger frequency thus larger median	uniform shape = lowest variability
larger frequency thus larger variability	use context
larger mean thus larger st.dev.	variation = variation frequency
larger n[26] = more variation	wrong bin width
larger n = more variation and higher mean	wrong boundaries bins

---

[26] n is the number of observations shown in the graph

Axial codes	Axial codes
larger range or variation frequency thus larger mean	wrong chance
largest max = more values above mean	wrong data collection
largest range = more values above mean	wrong description
largest st.dev. is not largest spread from center	wrong labels
largest st.dev.: uniform shape	wrong mean
largest st.dev.: U-shape	wrong median
mean < median right skewed	wrong skewness
mean = mean of frequency	zero frequency problems
mean = mean of hor. axis	

A.4   Full text of the codebook

The codebook categorizes the misinterpretations found in the literature, see Table A.4 below. These misinterpretations are not only found in students' work or reasoning, but also sometimes in the work or reasoning of researchers, teachers, software makers, journalists, etcetera. The misinterpretations in the codebook are categorized at two different levels. At the first level, categories summarize whether misinterpretations are related to the data represented, or related to the distribution represented, or none of these two (miscellaneous).

        The second level of codes identifies subcategories of possible origins of misinterpretations. These subcategories are characterized briefly in the last column of the codebook and are illustrated with types of misinterpretations listed. The last column ends with a note when NOT to assign this code, alerting the second coder to the boundaries of a particular code (subcategory).

Table A.4 Abstract of the codebook for data-related misinterpretations

Selective code	Assign this code if…
Identifying the measured variable only (VO)	The misinterpretation is related to identifying what was measured (which variable), but not to the measurement level of this variable. This includes the following misinterpretations: - frequency is regarded as a measured value (focus on the wrong – vertical – axis) with correct measurement level (interval/ratio – horizontal axis) - it is stated that frequency or number count on the vertical axis implies a histogram - a histogram is chosen to depict two variables, or a histogram is confused with a scatterplot - the graph is called a histogram, but has time-scale on the horizontal axis and the frequency is not a count of a time interval

Selective code	Assign this code if…
	represented on the horizontal axis (e.g., the percentage of unemployed in 2018, 2019, …, or the number of mortalities during certain time periods in a day). An example of a correct histogram with time-scale would be if it was counted how many times the postman worked 5 hours (per day), how many times 6 hours per day, and so on. When in doubt, calculate the arithmetic mean. If that is the mean of the time, it is a histogram. If it is the mean of something else (number of mortalities, number of unemployed), it is NOT a histogram and this code should be assigned. - two histograms are combined without extra information available - a wrong choice is made of variable(s) that are or will be depicted in the histogram (e.g., a histogram of the distribution of age is asked, and salary was chosen on the horizontal axis) - the variable depicted in the histogram is not the one depicted in the table - people use the number of bars as the number of measured values - a statement is made such as: bars on the right-hand side of the graph occur later in time - the time-scale issue is also related to the context - correlation, association or trend in time is mentioned Do not assign this code if: - there is also an issue with the measurement level (see the code identifying the measured variable and measurement level)
Measurement level only (MO)	The misinterpretation is related to only the measurement level of this variable, including: - what is called a 'histogram' is actually a distribution bar graph (nominal or ordinal data on horizontal, (relative) frequency on the vertical) - a statement is made about continuous distributions (e.g., normal distribution) in a graph with nominal or ordinal data - a bar graph (nominal or ordinal data) with a 'bell-shape' is chosen instead of a histogram - the intervals (e.g., the interval "[6, 7>") are seen as labels, thus nominal or ordinal measurement level (often inferred by the researchers when instead of a number scale on the horizontal

Selective code	Assign this code if…
	axis, the interval notation is used as a label underneath each bar)[27]
	Do not assign this code if:
	- the measurement level of the data represented is ratio or interval (exception: intervals with labels; see above)
	- a statement about the normal distribution is related to variability in a histogram (e.g., a normal distribution has the lowest standard deviation; see the code variability) or to randomness (e.g., if the sample is random, the population has a normal distribution; see the code population)
Identifying the measured variable and measurement level (VM)	Both previous misinterpretations are at stake, for example if:
	- frequency is regarded as a measured value (focus on the wrong—vertical—axis) in combination with a wrong measurement level (nominal or ordinal measured values; horizontal axis)
	- a statement is made that bars can be rearranged
	- bar graph and histogram are used as synonyms
	- it is unclear whether two variables are involved or one
	Do not assign this code if:
	- the authors who made the mistake are not native English speakers nor statistics teachers (see the language code)
	- a statement is made that a histogram depicts a relationship between two variables (without indication of measurement level or with interval or ratio measurement level; see the code identifying the measured variable only)
	- the software does not distinguish between bar graphs and histograms (see the code ICT)

The distribution-related codes category consists of four interrelated codes: variability, center, shape and grouping. Variability, for example, can be measured in terms of standard deviation and this is a measure relative to the center of the distribution. Similarly, the shape of the distribution is often assessed in relation to variability (e.g., an incorrect statement such as: a uniform distribution has the highest possible variability). We, therefore, made choices. If misinterpretations can be assigned to two subcategories (codes), the following hierarchy applies: grouping takes priority over all other subcategories, then variability, then measures of center, and finally shape. We

---

[27] Some researchers report this as a mistake or misconception, and we, therefore, included this in the review. One can discuss if labeling itself really is a problem, although there is indeed evidence that labeling can lead to or does stem from misinterpretations.

made this hierarchy keeping in mind that, for example, variability is often regarded as a more difficult concept to grasp than center.

Table A.5 Abstract of the codebook for distribution-related misinterpretations

Selective code	Assign this code if…
Variability (VY)	(A measure of) variability is wrongly used in relation to the shape of the distribution in a histogram, including the following misinterpretations: - variability is assessed as variation in the heights of the bars (thus variation in the frequency), instead of variation around the mean (e.g., range of measured value, IQR). This includes using words like 'bumpy'. - measures of variability such as standard deviation and IQR are wrongly used in a histogram, (e.g., if two symmetrical histograms have the same mean; their standard deviation is the same) -a statement about symmetry in relation to variability (e.g., a more symmetrical histogram has a variability). Do not assign this code if: - a statement is made about continuous distributions (e.g., normal distribution) in a graph with nominal or ordinal data (see the code measurement level only) - if the histogram is compared with another type of graph (see the code shape) - people use the number of bars as the number of measured values (see the code identifying the measured variable only)
Center (CE)	(A measure of) center is wrongly used in relation to the shape of the distribution in a histogram, including the following misinterpretations: - mean, mode or median are assessed of the frequency - heights of the bars are used for mean, mode, median - shape (e.g., skewness) in the histogram is assessed with measures of center - people cannot estimate measure(s) of center from a histogram[28] Do not assign this code if: - center and variability are both assessed (see the code variability[29])

---

[28] Some people might argue that a histogram is normally not used or made for estimating the mean. Nevertheless, as variability in a histogram is assessed, for example, relative to the mean of the data, one has to have at least a rough estimation of what the mean is to correctly interpret the variability.

[29] The reason for this choice is that variability is regarded as variation around a measure of center.

Selective code	Assign this code if…
	- people use the number of bars as the number of measured values (see the code identifying the measured variable only)
Shape (SH)	The misinterpretation is related to how the distribution of the data is depicted in a histogram, including graph conventions. The following are examples of misinterpretations: - dotplots and stem-and-leaf plots are called histograms - a histogram is wrongly matched to or compared with another type of graph (e.g., boxplot of the same data) - graph conventions for histograms are not met; including statements about connected bars or discreteness of data or when space is left between bars in a histogram[30] - any graph can depict the shape of a data distribution (e.g., the authors state that people do not understand that a graph like a histogram is needed to describe shape, center and variation[31]) - the intervals of bars with a frequency of zero are left out or a bar is used when the frequency is zero - statements are made that a table is more precise than a graph (thus not taking variability[32] and random noise into account) - not knowing the purpose of different graphical representations, including histograms - area or density is not correctly used - outliers are missed or not taken into account Do not assign this code if: - a statement is made about continuous distributions (e.g., normal distribution) in a graph with nominal or ordinal data (see the code measurement level only) - shape or gaps are used in relation to center (see the code center) - shape or gaps are used in relation to variability (see the code variability) - density or area are wrongly used in relation to the bin width (see the code grouping) - the variables in the graphs are not the same (see the code identifying the measured variable only) - people use the number of bars as the number of measured values (see the code identifying the measured variable only)

---

[30] We are aware that some researchers will not regard this as a misinterpretation.

[31] We are aware that other graphs exist that describe shape, center and spread.

[32] Variability caused by sampling.

Selective code	Assign this code if…
Grouping (GR)	The misinterpretation is a misunderstanding of the process of information reduction[33] encompassed in a histogram leading to a possibly different shape or modality (e.g., bimodal, depending on the bin width), including the following misinterpretations: - a statement is made that all values in a bar or only the midpoint are/is measured or when the data in the histogram are used as raw data (not taking into account the information reduction caused by grouping into bins) - a wrong bin width is chosen (e.g., different bin widths without using density on the vertical axis), or not enough (e.g., two) or too many bins (e.g., a 'bin' for every value in a continuous distribution, often resulting in a graph with only frequencies of one) in relation to the given context - area or density is wrongly used in relation to the bin width Do not assign this code if: - the wrong bin width is generated by the software (see the code ICT)

Table A.6 Abstract of the codebook for miscellaneous misinterpretations

Selective code	Assign this code if…
Context (CO)	The misinterpretation is due to the context or the research question, including the following: - shape, center, and variation are incorrectly, or not at all, related to the context - a wrong description of the distribution in a histogram is given in relation to context - the context or personal knowledge is used instead of the given data in the histogram Do not assign this code if: - correlation, association, or trend over time is mentioned (see the code identifying the measured variable only).
Population (PO)	The misinterpretation of the histogram is related to the distinction between sample and population, including the following: - a statement is made indicating that sample and population in a histogram are the same (e.g., distribution, shape)

---

[33] One could argue that grouping is related to data reduction and, therefore, should be categorized in the data-related category. However, the data reduction (information reduction) directly influences the shape as well as the modality. Therefore, it was classified as distribution-related.

Selective code	Assign this code if…
	- the histogram (of a sample) is regarded as a precise representation of the population (random noise in sample or population is not taken into account) - a statement is made that the sampling distribution with z- or T-scores has a normal distribution thus the population has a normal distribution
ICT (IT)	The misinterpretation is embedded in the software by the software designers, including the following: - the software does not distinguish between bar graphs and histograms - histograms produced by ICT have wrong or strange boundaries - (use of ICT leads to) the idea that a fixed number of classes is needed (e.g., 10) with or without ICT, regardless of situation
Unknown (U)	A misinterpretation is mentioned by the authors of the publication but is not specific enough to be coded.
Translation (T)	The misinterpretation may be caused by translation, including the following: - if authors use the word bar graph for a histogram—or as a synonym—and are not from a native English-speaking country. Do not assign the code if - the authors are teachers/researchers of statistics (education)[34].

---

[34] In statistics, a distinction is made between histograms and other types of graphs with bars. Researchers and teachers investigating statistics education are expected to be aware of this, even in non-English speaking countries.
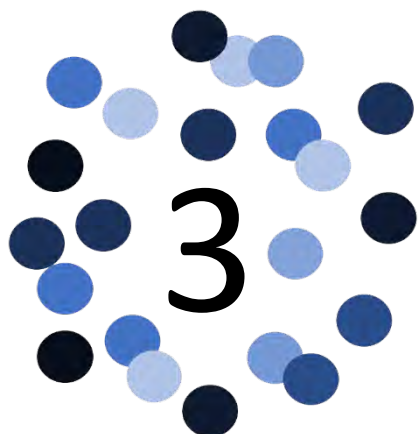
3

Secondary school students' strategies when interpreting histograms and case-value plots: An eye-tracking study

"I never am really satisfied that I understand anything; because, understand it well as I may, my comprehension can only be an infinitesimal fraction of all I want to understand about the many connections and relations which occur to me, how the matter in question was first thought of or arrived at…" [35]
Ada Lovelace

This chapter is based on
Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2022). Secondary school students' strategies when interpreting histograms and case-value plots: An eye-tracking study [Manuscript submitted for publication]. Freudenthal Institute, Utrecht University.

---

[35] https://www.goodreads.com/author/show/3950749.Ada_Lovelace

Abstract Many students persistently misinterpret histograms. A literature review made plausible that students' difficulties with statistical key concepts become manifest in most common misinterpretations. In line with suggestions from that review, in the present study, we address students' conceptual difficulties more broadly, rather than focusing on a specific misinterpretation. Students' difficulties related to the statistical key concepts data and distribution can be observed when students confuse histograms with look-alikes, including case-value plots. However, little was known about students' strategies when solving histogram tasks. Using students' gaze data, we address the question: how and how well do Grades 10–12 pre-university track students estimate and compare arithmetic means of histograms and case-value plots? We designed four item types: two requiring estimation of the mean and two requiring comparison of means. Gaze data of 50 [15–19 years old] Grades 10–12 students solving these items were combined with data from cued recall. We found five strategies. Two hypothesized most common strategies for estimating means were confirmed: a typical case-value plot strategy and a histogram (interpretation) strategy. A third, new, count-and-compute strategy was found. Two more strategies were found for comparing case-value plots and histograms: a distribution-informed histogram strategy and a case-value plot strategy, both taking specific features of the distribution into account. In 43% of the trials, students used a correct strategy for estimating the mean from one histogram. In 50% of the trials, students used a correct strategy for comparing two histograms. Surprisingly, some students used a distribution-informed histogram strategy for comparing two case-value plots. As several of the students' strategies related to how and where the data and the distribution of these data are depicted in histograms, future interventions should aim at supporting students in understanding of these concepts in histograms. A methodological advantage of eye-tracking data collection is that it reveals more details about students' thinking processes than thinking aloud protocols. Teachers can use gaze patterns (scanpaths) to draw students' attention to correct and incorrect interpretations of graphs. We speculate that gaze data can be re-used to underpin ideas about the sensorimotor origin of learning mathematics.

Keywords Eye-tracking (ET); Histograms; Problem-solving strategy; Graphs; Statistics education; Misinterpretation.

3.1   Interpreting histograms

The ability to understand and communicate through graphs—called graphicacy (Balchin & Coleman, 1966)—is an important skill for citizens (Ben-Zvi & Garfield, 2004a). "Looking at the data" is essential (Watson & Moritz, 1999), as graphs can reveal data patterns that might stay hidden when computational measures or hypothesis tests alone are used and even lead to opposite conclusions (Pastore et al., 2017). In addition to this general importance of statistical graphs, histograms are considered of key importance for introducing continuous probability distributions. Nevertheless, difficulties with understanding and interpreting histograms have been reported for several decades now (e.g., Cooper, 2018; Pettibone & Diamond, 1972).

A review of the literature revealed many misinterpretations regarding histograms (Chapter 2). Based on that review, we conjectured that overlooking the importance of data-related conceptual difficulties in previous research—and interventions addressing these difficulties—might partly have contributed to the persistence of people's misinterpretations. In addition, we proposed to address these conceptual difficulties more broadly rather than focusing separately on each specific misinterpretation. In the present study we, therefore, examine several data- and distribution-related conceptual difficulties concerning histograms.

Students' difficulties related to the statistical key concepts data and distribution can be observed when students confuse histograms with look-alikes, including case-value plots (Cooper, 2018). Figure 3.1 illustrates one common confusion when the questions are posed whether the arithmetic mean and variation are higher in the graph on the left, or the right, or approximately the same for both? To answer these questions, one first needs to understand what data are depicted in these graphs. In the histogram (left), weight of packages is on the horizontal scale and ranges between 0–9 kg. The first bar contains twelve packages with a weight between 0–1 kg. In total, there are 138 packages. The mean weight is around 4.5 kg and can be estimated on the horizontal axis. In the case-value plot (right), nine students collected garbage. Garbage weights range between 6–9 kg with a mean of around 7.7 kilograms, which can be estimated on the vertical axis. At first glance, the variation in both graphs might seem the same. However, in this example, the variation is the highest in the histogram, regardless of whether an informal measure of variation (range) is used or a more formal one (standard deviation).

Figure 3.1 A histogram (left) with one statistical variable (weight) and a case-value plot (right) with two statistical variables (name and weight)
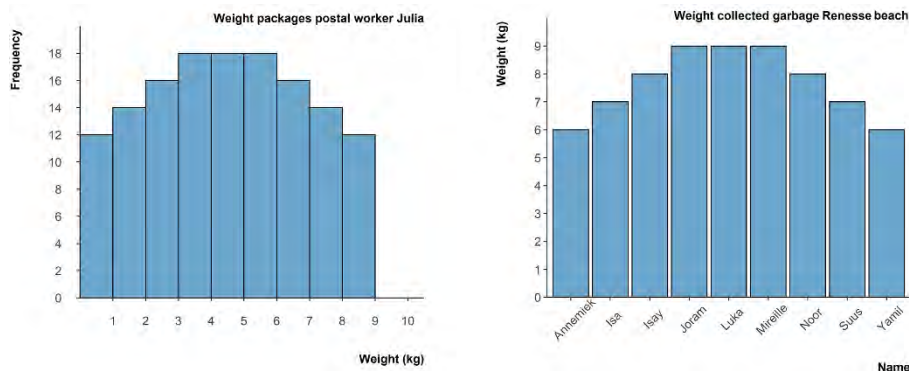
The example above shows that understanding data in histograms includes an understanding of what, how many, and where the variables are depicted in a histogram. For example, a histogram (Figure 3.1, left) has one statistical variable located along the horizontal axis (here: weight). A case-value plot (Figure 3.1, right) has two statistical variables—here: name and weight—represented along two different axes.

Many secondary school students struggle with interpreting histograms. Most of 412 Malaysian Grade 10 students most incorrectly calculated the mean from a histogram (Ismail & Chan, 2015) by dividing the sum of the frequencies by the number of bars—an approach only correct for finding the mean from a case-value plot if frequencies were the measured values. In a study by Whitaker and Jacobbe (2017), around 3,700 Grades 6–12 United States students answered one or more questions about histograms. A common misinterpretation when comparing histograms was that the least variability from the mean was understood as the least variation in the heights of bars. Thus, students compared frequencies instead of measured values. Similarly, in an item about height, some students thought that taller bars in a histogram indicate taller instead of more people. In summary, people find it difficult to identify the statistical variable and its measurement level depicted in a histogram (Chapter 2). This identification is part of the key concept of data (Ben-Zvi & Garfield, 2004a).

Despite several decades of carefully designed interventions students' difficulties persist (e.g., Delport, 2020). As many of them relate to confusing histograms with case-value plots, we decided to examine how this confusion arises. Various studies draw conclusions from students' final answers (e.g., Whitaker & Jacobbe, 2017). Some studies used other approaches, including classroom observations (Bakker, 2004a), concurrent think-aloud protocols and

observations or interviews (Stone, 2006), or students' written explanations to open answer questions (Whitaker & Jacobbe, 2017). However, little was still known about students' detailed thinking processes for reaching their answers—and thus their strategies—when interpreting histograms and case-value plots.

In this study, we aim to better understand how students interpret data in histograms and case-value plots. By observing students' actions—estimating and comparing means from these graphs—it becomes clear how students use their conceptual knowledge of the data in these graphs, hence, what strategies they employ. Therefore, the research question is: *how and how well do Grades 10–12 pre-university track students estimate and compare arithmetic means of histograms and case-value plots?* We use eye-tracking as a tool to observe some of students' actions. In the following section we elaborate on some advantages and disadvantages of using such tool.

3.2 Theoretical background

3.2.1 The mean as a representative value

Modeling and interpreting distributions as a center with variation around it is important for statistical thinking and modeling. Historically, the mean evolved from estimations of a representative value (Bakker & Gravemeijer, 2006). Therefore, students need to learn to estimate the center of a data set (Bakker, 2003)—represented in a histogram—and judge the variation around this center. Bar representations can be used for visually estimating the mean (Cai et al., 1999). Moreover, Gal (1995) argues that: "Asking students to compute or estimate averages from data presented in […a] histogram […] can often reveal certain strengths and flaws in students' knowledge" (p. 99).

Most secondary school students (Grades 7–12; 12–18 years old) know how to calculate the arithmetic mean from raw data. Applied to a histogram, this calculation could be multiplying the height of each bar with its middle value, summing the results, and dividing this answer by the sum of the frequencies. Besides this school-learned algorithm, several approaches exist for finding an average from a histogram including mode, average as reasonable—centered within the data—midpoint or median, and the point where the data are in balance (Mokros & Russell, 1995).

Comparing groups is important for statistical literacy, can motivate students (e.g., Konold & Pollatsek, 2002), and is recommended for introducing hypothesis testing (Watson & Moritz, 1999). Group comparison is also considered a way for teaching statistical reasoning (Makar & Confrey, 2004). "All students […] should be encouraged to draw comparisons between

groups." (GAISE II, Bargagliotti et al., 2020, p. 10). The mean—as a representative value of a group—can be used for such comparisons. We, therefore, designed items for which students needed to compare and estimate means. For comparing groups, Frischemeier and Biehler (2016) found six approaches: comparing center (means, median), spread (IQR, range), shift, skewness, percentages (in histograms: areas) below or above a certain value called p-based and values belonging to (e.g., first) quartiles—called q-based.

3.2.2 The potential of eye-tracking for graph interpretation research

Eye movements are our most frequent motor movements and play an important role in our cognitive processes (Spivey & Dale, 2011). Gaze data can provide a special window into students' thinking processes on a micro level. Although it is generally accepted that such processes can be inferred from eye movements (e.g., Kok & Jarodzka, 2017), there is no simple relation between the two (e.g., Orquin & Holmqvist, 2017). For example, not every eye movement is part of students' solving strategy (e.g., Schindler & Lilienthal, 2019).

Gaze data can provide evidence of actions that are related to specific concepts (e.g., Chumachemko et al., 2014; Schindler et al., 2021; Schindler & Lilienthal, 2019), but a form of data triangulation is often useful. In this study, we combined the gaze data with data from cued recall. According to Van Gog et al. (2005), the advantage of recall is that it often contains more information on the 'why' and 'how', hence on students' strategies, compared to thinking aloud. The disadvantage, however, is that students may have forgotten their strategy after completing all items. This risk can be reduced by cueing students: having them look back at their eye movements (e.g., Kragten et al., 2015; Van Gog & Jarodzka, 2013).

The use of gaze data on graph items for finding students' strategies has several advantages over other data collection methods. First, most people cannot manipulate their gazes as they are not aware of their eye movements. Second, eye movements can reveal thinking processes that students are not aware of or cannot articulate (Green et al., 2007). Third, thinking aloud can slow down or alter the problem-solving process and influence eye movements (e.g., Dickson et al., 2000). Fourth, students who think aloud might only report what is readily available or what they think is expected (Wilson, 1994). Last, gaze data can reveal students' strategies toward the answer. This makes it possible to detect whether students used a correct strategy despite an incorrect answer—or vice versa.

Eye-tracking is quite often used for investigating how students solve mathematical problems (e.g., Lai et al., 2013; Lilienthal & Schindler, 2019). A review of eye-tracking studies in mathematics education showed that gaze

data are particularly suitable for studying processes and subconscious mathematical thinking (Strohmaier et al., 2020). Research in the last two decades demonstrates that gaze data are suitable for studying students' thinking processes for many other graph types (e.g., Andrá et al., 2015). In a study using stacked dotplots—similar to histograms but with all cases visible and vertically stacked—novices and experts both mostly used global comparison methods like displacements of means and modal clumps when comparing two groups, experts more than novices (Khalil, 2005). In addition, several novices used local comparison methods, mostly comparing similar parts of both graphs. Recently, Schreiter & Vogel (2023) found similar patterns when tracking the gazes of Grade 6 students. Taken together this suggested that eye-tracking would fit our aim to study details of students' strategies while solving statistical graphs.

3.2.3 Considerations for choosing gaze metrics

Various gaze metrics can be used in eye-tracking research: temporal, count, or spatial. A review study on the use of gaze data in education (Lai et al., 2013) shows that cognitive researchers often use temporal metrics to analyze gaze data (e.g., total gaze or fixation duration or total dwell time, time to first fixation, total reading time) followed by count (e.g., fixation count). A fixation is where people look at the screen [36]. Saccades are relatively fast transitions between two fixations.

Some of the metrics—for instance, total dwell time—are nowadays considered a threat to the validity of the research (Orquin & Holmqvist, 2017). To calculate such metrics, areas of interest (AOIs) that seem relevant for the given item need to be defined. The choice of number and size of AOIs influences results on those metrics. The least used measures are spatial (e.g., scanpath, fixation position). Spatial measures can both be used independently or combined with AOIs.

Temporal metrics have the advantage of being easy to compute. The disadvantage, however, is that they often provide only global insight into students' thinking processes, as many gaze data details are ignored. Traditional temporal metrics, for example, can hide visual scanning patterns (Goldberg & Helfman, 2010). Such general measures do not seem to provide topic-specific guidance needed for learning or instruction on that topic. Spatial measures, such as scanpaths, seem better suited for providing detailed information about

---

[36] A "period of time during which a specific part of [item on screen] is looked at and thereby projected to a relatively constant location on the retina […] operationalized as a relatively still gaze position in the eye-tracker signal implemented using the [Tobii] algorithm." (Hessels et al., 2018, p. 22)

students' thinking processes (Hyönä, 2010) and are advised for problem-solving research (Tai et al., 2006). The use of students' scanpaths for identifying strategies is rare (Epelboim & Suppes, 2001). The disadvantage of using scanpaths is that it often requires time-consuming qualitative analyses of eye-movement data (Alemdag & Cagiltay, 2018), although machine learning algorithms may be helpful in analyzing scanpaths from heatmaps, (Schindler et al., 2021), order of AOIs (Garcia Moreno-Esteva et al., 2018) and geometrical vectors (Chapter 4; Jarodzka et al., 2010; Schreiter & Vogel, 2023).

From the pilot study—described in the next section—with six university students (Boels et al., 2018), we learned that students' scanpaths—thus their strategies—typically differ within the graph area (Figure 3.2). In contrast to most existing eye-tracking studies that define scanpaths as a sequence of AOI transitions (e.g., Garcia Moreno-Esteva et al., 2018), here a scanpath is defined as a sequence of fixations and saccades within an AOI (graph area). Therefore, we characterized the perceptual form of a sequence of fixations and saccades on the graph area.

3.2.4 Pilot study

In an exploratory pilot study with six university students, we found two main strategies for estimating the mean from single histograms and case-value plots: a histogram strategy and a case-value plot strategy (Boels et al., 2018). First, in a 'typical' histogram strategy, students interpreted the graph at hand as a histogram. Although this strategy can initially be applied to both graph types (Figure 3.2, top), it is only appropriate for histograms. In this strategy, students visually search for the mean on the horizontal axis, often using a point on this axis where the distribution in the graph seems in balance. Students' gazes go up and down between a point on the horizontal axis and the top of the bars, resulting in a specific perceptual form of their scanpaths on the graph area: a vertical line. Although there are other correct strategies for finding the mean from a histogram, students did not use these in the pilot study, possibly because they were asked to estimate from the graph on the screen.

Second, in a 'typical' case-value plot strategy, students visually search for the mean on the vertical axis, meanwhile often 'flattening' the bars, sometimes referring to an imaginary horizontal line (Figure 3.2, bottom). In this strategy, students interpreted the graph as a case-value plot and their gazes went back and forth between a point on the vertical axis and the middle of the bars, resulting in a horizontal scanpath on the graph area.

Figure 3.2. Examples of gaze patterns on histograms and case-value plots
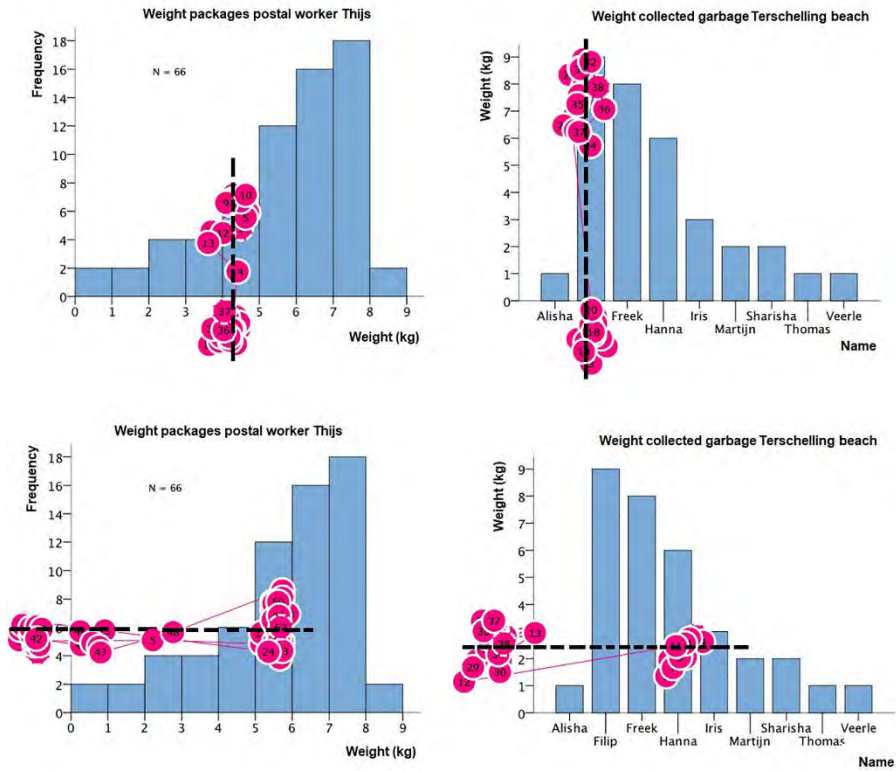
Note. Top: histogram (interpretation) strategy on histogram (left; similar to Item06) and initial application to a case-value plot (right, Item04). Bottom: case-value plot (interpretation) strategy applied to a histogram (left,) and to a case-value plot (right). Circles indicate fixations; thin lines between the circles indicate saccades. In this, and following figures with scanpaths, axis, and graph titles are translated into English whenever possible. Horizontal or vertical line segments—indicating scanpaths—are superimposed for the reader's convenience. The barely visible numbers on the circles indicate the order of fixations.

Our pilot study made us hypothesize the existence of the above-described two main student strategies for estimating the mean in histograms and case-value plots. In the present article, we check the commonality of these strategies in a larger sample than Boels et al. (2019a). Furthermore, we now (1) also explore strategies for comparing two graphs, (2) report on how gaze data reveal the imaginary object students talked about (e.g., horizontal line), (3) address the uniqueness and potential of considering the perceptual form of scanpaths on one AOI compared to other metrics and (4) discuss how gaze data could be re-used to underpin ideas about the sensorimotor origin of learning mathematics. Furthermore, we added a section on (5) the potential of eye-tracking for graph

interpretation research, (6) considerations for choosing gaze data metrics, and (7) lessons learned on doing eye-tracking research for novices in this field (Appendix A of this chapter).

3.3   Method

3.3.1   Participants and school curriculum

Gaze data were collected from 50 Grades 10–12 pre-university track students (Table 3.1) from a Dutch public secondary school [15–19 years old; mean = 16.31 years]; 23 males, 27 females. Each student was individually given the items in a separate room in their own school. Participation was voluntary; informed consent was signed and permission from the Utrecht University ethical committee was obtained. Participants received a small gift for their participation. Dutch students can choose four different mathematics levels starting from Grade 10 (Table 3.2).

Table 3.1 Grade level and age of participants. One participant did not provide details on grade, another one not on age (see also Chapter 3)

Grade	Number of participants		Age	Number of participants
10	20		15	12
11	17		16	19
12	12		17	10
Unknown	1		18	7
Total	50		19	1
			Unknown	1
			Total	50

Note. Due to legislation, data on ethnicity cannot be collected. In the Netherlands, there is hardly any difference between public and private schools, nor between city, suburban, and rural schools. Private schools are rare.

Table 3.2 Choice of mathematics course per grade level

Mathematics	Grade 10	Grade 11	Grade 12	Unknown	Total
A	13	10	4	1	28
A and B			1		1
B	5	3	3		11
B and D	2	3	4		9
C		1			1
Total	20	17	12	1	50

*Note.* The content of the different types of mathematics is, according to Daemen et al. (2020), as follows: A—applied analysis and statistics in economics/health; B—analytic geometry, and analysis, formal or applied in engineering/science; C—liberal arts topics, e.g., statistics, logic, symmetry, perspective and D—more analysis, Euclidean geometry, statistics, other applications in engineering/science.

The textbooks used in this secondary school had one histogram—referred to as bar graphs—in Grades 7 and 9, and around five in Grade 8. Students with Mathematics A, D, or C re-encounter histograms in Grades 10–12; students with Mathematics B re-encounter histograms in Grade 11 as part of a school-selected topic. Textbooks sometimes confuse histograms and case-value plots and pay no attention to relevant differences.

3.3.2 Materials

Students solved 25 items. Only the first twelve are relevant for the present study and included four different item types: two single graph types—a histogram or a case-value plot—and two types with two histograms or two case-value plots (Figure 3.3). For each item type, we constructed three versions. The question for all single graph items was what the approximate mean weight was, either per package (histogram) or per person (case-value plot). The question posed in the double graph items was which graph had the higher mean weight, with three answer options: left, right, or that both had (approximately) the same mean weight. Most items were right-skewed and double graphs all had pair-wise similar skewness, shapes, and symmetry (Figure 3.3, middle and right). In two double graph items, for example, one graph was shifted to the right, which is relevant for histograms but irrelevant for case-value plots as bars can be reordered in the latter. As students often do not understand the influence of bars with zero 'height' (e.g., delMas & Liu, 2005), we added some items with 'zero' bars. We also chose a non-ambiguous context (weight), as some misinterpretations are related to the context or appear when context is missing (e.g., Lem et al., 2013c; Meletiou, 2000).

**Figure 3.3** Graphs used in Item02 (left) and Item03 (middle and right)
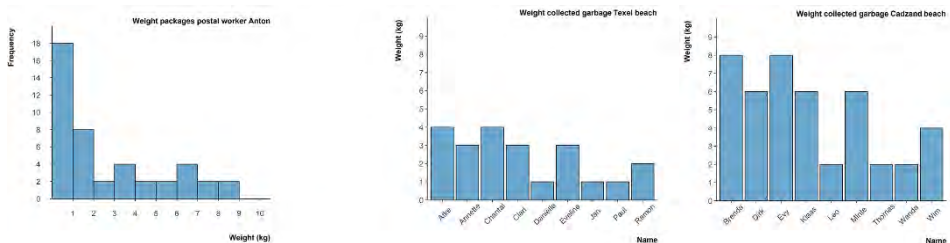
Note. The answer for the histogram (left) is approximately 2.7 [1.6–3.8]. The answer for comparing the case-value plots (middle and right) is that the mean is higher in the right-hand graph.

3.3.3 Eye-tracking apparatus

A Tobii XII-60 (sampling rate: 60 Hz) was placed on an HP ProBook 6360b between the 13-inch screen (refresh rate: 59 Hz) and keyboard (Figure 3.4). Participants used a chin rest. Gaze data were recorded and processed with Tobii Studio software version 3.4.5 (Tobii, n.d.-a). The calibration procedure consisted of a 9-point calibration; this software has no built-in validation procedure. Therefore, we included a validation screen in the set-up at the start, after each item, and at the end. Fixations and saccades are calculated by the Tobii software.

Good data quality can be hard to achieve in an eye-tracking study. Nevertheless, no students were excluded from the data set, as the data loss per trial (averaged over all participants) and per participant (averaged over all items) were below the exclusion point (34% or more). Accuracy and precision are especially important when using temporal or count measures as—for example—a low accuracy can result in fixations being classified to a different AOI than where participants actually looked. In our qualitative study—where scanpaths from videos of eye movements were used—accuracy and precision have less influence on the final results. The mean accuracy is 56.6 pixels (1.16°) with the highest accuracy on the graph area (mean 13.4 pixels or 0.27°). The average precision (0.58°; RMS-S2S; Holmqvist et al., 2023) is considered good.

Figure 3.4 Set-up of the experiment (person in the picture did not participate). Red oval: eye-tracker placed on the laptop
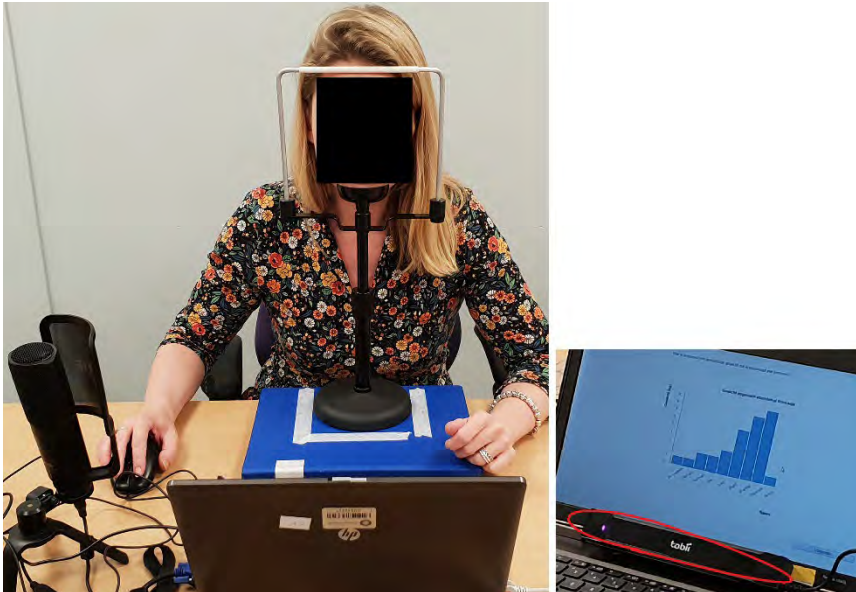
3.3.4 Data collection

Data were collected on the following: characteristics of participants (e.g., age, prior knowledge), students' answers, answer correctness, and solution strategy through cued recall. Students' gaze data on graph tasks behind a laptop were collected. The Tobii software collected *x*- and *y*-coordinates of the eyes on the screen for each time stamp and produced videos of smoothed eye movements overlaid on the screen image for each item. The video also displayed which answer students clicked on in the double graph, multiple-choice items. For the single graph items, students answered verbally. Furthermore, we asked half of the students immediately after they finished the 25 items what strategies they used. To this end, we used students' own gaze data for a cued recall of what they thought when solving the item.

We illuminated the location where students looked—through a kind of spotlight—and made the rest of the graph darker. We preferred this method over having students look back at their fixations (e.g., red dots) for two reasons. First, it may prevent students from making different eye movements when looking back—and describing the corresponding strategy—instead of the strategy they initially used (M. Kragten, personal communication, March 8, 2017). Second, this made visible the exact information that the learner had looked at, instead of the information being covered by—for example—a red dot (Jarodzka et al., 2013).

3.3.5 Data analysis

We first discuss how we analyzed and coded single graph items based on the hypothesis of the existence of two typical strategies as found in the pilot study. Next, we discuss the analysis of double graph items.

Single graph items

As students were asked to estimate the mean, answer ranges for correct answers were set to the mean +/- 1.1 based on spread in experts' answers, see Appendix A of this chapter. A pragmatic iterative approach was used for coding gaze data (Tracy, 2013) that alternated between existing explanations and emergent interpretations. Hence, although we started with two predefined categories (deductive approach, Twining et al., 2017), we used open coding that allowed other strategies to emerge (inductive approach). The unit of analysis was one trial. Note that in mathematics education we usually talk about an item, task, or problem. In eye-tracking research, the gazes of a student solving one such item are called a trial.

To analyze students' scanpaths, the first author qualitatively coded video data [37] of eye movements on single graph items (Table 3.3). In the pilot study (Boels et al., 2018), we found two strategies that led to two predefined coding categories: a 'typical' histogram strategy associated with a vertical scanpath on the graph area and a 'typical' case-value plot strategy associated with a horizontal scanpath on the graph area (Figure 3.2). In the present study, a third strategy emerged from the data during coding: count-and-compute, associated with a zigzag scanpath from the vertical axis to the top of bars, sometimes followed by gazes on almost all names or numbers on the bottom of the bars along the horizontal axis (Figure 3.5). This strategy led to a correct answer for case-value plots but not for histograms. Our interpretation—supported by triangulation with verbal data—is that students added the heights of all bars and divided by the number of bars:

StudentL13:	Looking at the number of people [frequency] and at the weight and again I added up [frequencies] here and divided [this sum] by the number of people.

Table 3.3 Gaze data single graph items (* predefined categories)

Code	Assign this code if most of the gazes on the graph area are
Case-value plot strategy*	- horizontal (e.g., from vertical axis to a bar in the graph) and approximately on the same height - at specific numbers on the vertical axis Do not assign this code if one of the count-and-compute strategy options hold (see below).
Count-and-compute strategy	- horizontal and clearly go from the vertical axis to (almost) all (top of the) bars (zigzag pattern or repeating Z-pattern) - jump underneath the horizontal axis or bottom of bars from one to the next (for almost all bars)
Histogram strategy*	- vertical (e.g., from horizontal axis to the top of a bar in the graph) and approximately on the same position - at specific numbers on the horizontal axis Do not assign this code if one of the count-and-compute strategy options hold (see above).

For coding the verbal reports, a predefined codebook was used with both typical strategies (Table 3.4). In some cases, the verbal data allowed us to distinguish variations of strategies that might be associated with the same scanpath. In the median and mode strategy, students correctly understood where to find the data in the histogram but took an incorrect measure of

---

[37] We use static gaze plots to report results instead of video stills which would require much more figures to show scanpaths.

center. As we found no consistent way of detecting this strategy from the gaze data, we reported this in the verbal data only and did not use it for the final coding in the scanpath analysis. As anticipated, some students described a strategy during the recall that differed from the strategy that was visible from their gazes and answer.

Table 3.4 Verbal data single graph items (* predefined categories)

Code	Assign this code if a student talks about
Case-value plot strategy*	- using the height of one or more bars to determine the mean - a horizontal line
Count-and-compute strategy	- adding the heights of the bars and divides the answer by the number (9) of bars
Histogram strategy*	- using a balance or balancing point to find the mean
Median strategy	- making areas left and right in the graph the same. Can only be assigned for histograms.
Mode strategy	- using the weight that belongs to the highest bar. Can only be assigned for histograms.

Final coding was obtained by triangulating the coding of the gaze data for one trial with the coding of the verbal data of the same trial when available. Whenever there was a discrepancy, the video of the gazes was reconsidered. The video code—most often similar to the final code—was only recoded if obvious signs for a specific type of strategy were missed. When the discrepancy between the coding of video and verbal data remained, the first coder decided on the strategy code. An example of such a discrepancy is when a student's gazes (Figure 3.6) and answer (ten) indicated a case-value plot strategy but verbal data in the transcript (below) indicated a histogram strategy. Therefore, the video and final code remained identical: a case-value plot strategy.

StudentL01:   Then I looked to see which one occurred more often. That was eight. And then it went down a bit so it should be between four and eight because those are the highest values that occur. And then somewhere in the middle [six].

Figure 3.5 Count-and-compute strategy
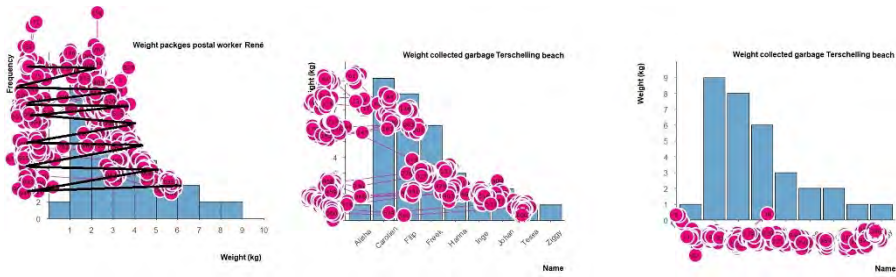
Note. Parts of scanpaths alternating between top of bars and vertical axis for both a single histogram (left, Item01) and single case-value plot (middle, Item04) with fixations on almost all names (right) indicating counting bars. Note the characteristic zigzag pattern (black line segments superimposed for convenience of the reader in the left graph. A similar pattern is noticeable in the middle graph).

Figure 3.6 Static gaze-plot with a case-value plot strategy of studentL01 on Item06: reading off frequency values from the vertical axis
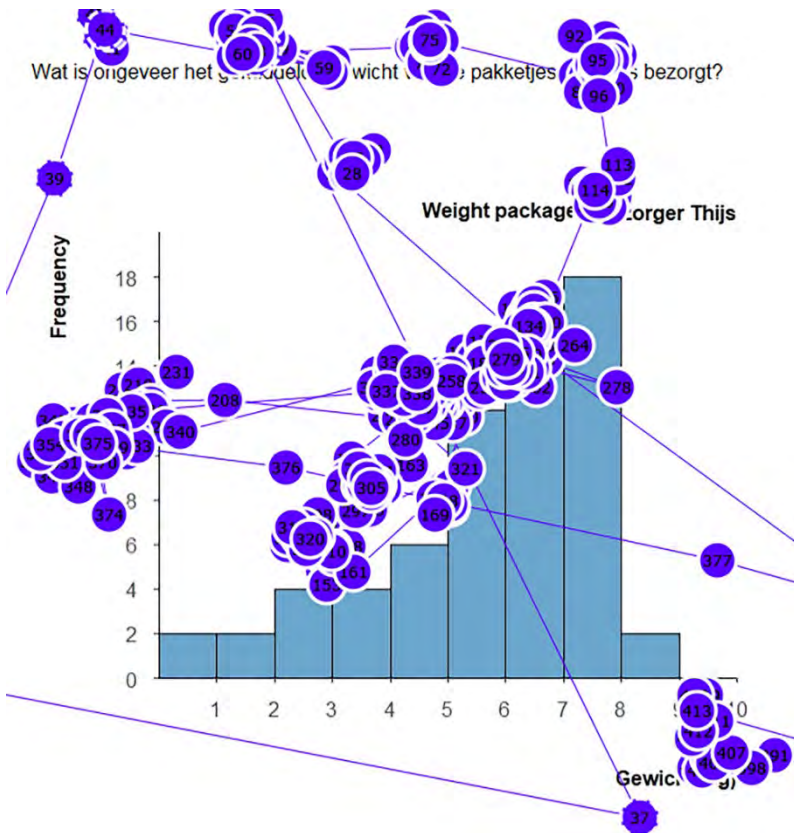
Note. Note that there are no fixations on the horizontal axis even though the title/label weight is read.

If the strategy was still not clear after reconsidering the video coding, or the first coder could not reach a decision, the final code 'unclear' was used. The resulting final code is presented in the Results section. Coding reliability was checked by a second coder who coded ten percent of the trials, and the verbal data associated with these. Interrater agreement for the final codes was .89 (Cohen's Kappa), which is considered almost perfect agreement (Landis & Koch, 1977).

Double graph items

For the double graph items, we explored strategies (inductive approach) per trial. Pilot data, and, therefore, coding categories were not available. Most scanpaths indicated a distribution-informed strategy that takes specific characteristics of the distribution of the data in the graphs into account (Table 3.5). The differences in gazes between distribution-informed histogram and case-value plot strategies are subtle (Figure 3.7), and, therefore, sometimes hard to classify. In both strategies, students looked at the position of the zero frequency bar and weight range (histogram) or position of zero weight bars and number of bars (case-value plot) but the given—correct—answers differ. Small differences in scanpaths—such as the pattern on the horizontal and vertical axis and labels—combined with the given answer—determined the correctness of this distribution-informed strategy. Note that the researchers analyzed the videos, not the gazeplots. Videos of gazes showed how the student's gazes on the graph area progressed through time which made it slightly easier for a trained viewer to interpret the scanpath pattern. Gaze data were combined with verbal data when available. For the verbal data on double graph items, we used open, axial, and selective coding (Table 3.6). For the approach to obtaining a final code, see the single graph items.

Table 3.5 Example of codebook for gaze data on double graph items, histogram strategy code

Assign this code if the gazes are	Assign this code if the answer is
- on the top of bars in both graphs, indicating comparing double heights - going back and forth along the horizontal axis - see the codebook for single graph items	Same or Texel (Item03) Willem (Item05) Kees (Item09, Figure 3.7)

Figure 3.7 A correct distribution-informed strategy for comparing the means of two histograms in Item09 (top row) using similar shape, shifted to the right and for comparing means of two case-value plots in Item07 (bottom row) using shape and number of bars
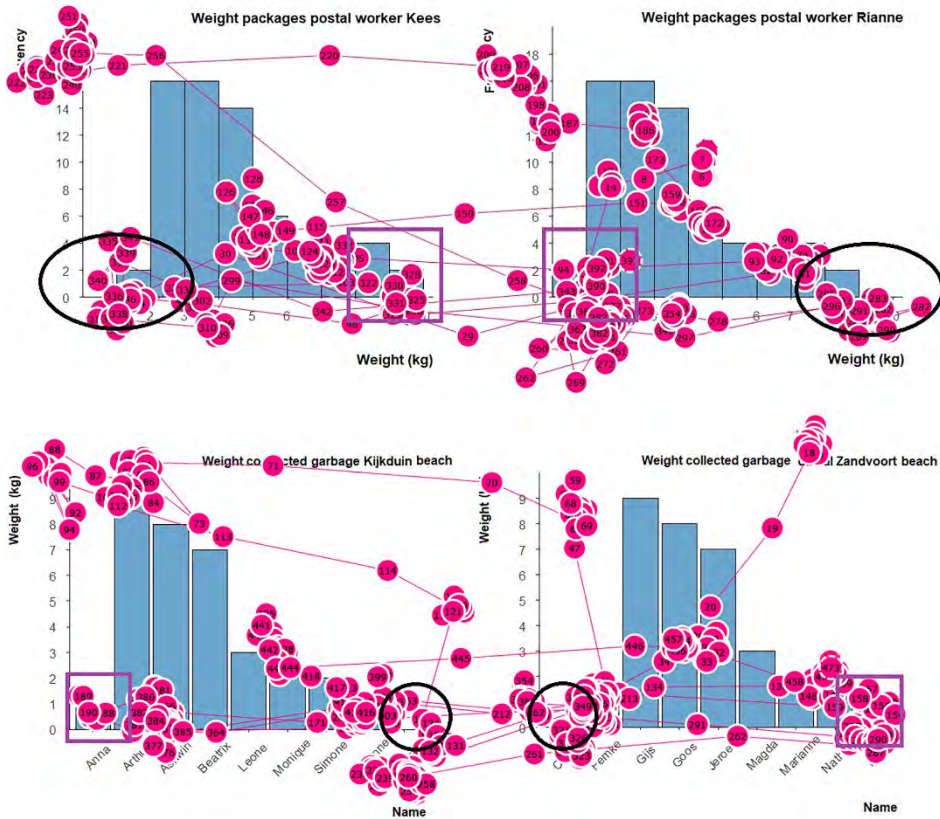
Note. Students specifically compared the position of the 'zero' bars (black ovals) and others on similar positions (e.g., purple squares). Correct answer top row: Kees' packages had a higher mean weight; bottom row: mean weight is the same for both graphs.

Table 3.6 Example of codebook for verbal data double graph items, case-value plot strategy code

Assign this code if a student talks about
-the two zero bars or missing bars lower the mean
-comparing the heights of the bars
-higher/more/fewer bars in one graph thus higher mean
-bigger area thus higher mean
-same number of bars and same heights
-similar bars but reordered

3.4 Results

More details can be found in Appendix A of this chapter.

3.4.1 Strategies on single graph items

The most common strategy for single graph items was the 'typical' case-value plot (interpretation) strategy that we already found in our pilot study—associated with a horizontal scanpath and only correct when applied to case-value plots—with 37% and 71% for histogram and case-value plot trials respectively (Table 3.7). The 'typical' histogram (interpretation) strategy—associated with a vertical scanpath—was used for single histograms in 43% of the trials. The most common strategies hypothesized based on the pilot study were, therefore, confirmed. A third strategy was found—count-and-compute, associated with a zigzag scanpath (Figure 3.5)—in which students add heights of bars and divide by the number of bars.

Table 3.7 Strategies, percentage of trials (*N* = 150) per item type (correct strategy in **bold**)

	Histogram strategy	Case-value plot strategy	Count-and-compute strategy	Unclear
Single histogram	**43%**	37%	18%	2%
Single case-value plot	0%	**71%**	**29%**	0%

Note. Although a correct count-and-compute strategy is also possible for histograms, such a strategy was not found. In case-value plots, count-and-compute only returns the mean if the sum of the heights of bars is divided by the number of cases.

StudentL18 describes a correct 'typical' histogram strategy for single Item02:

StudentL18:	I was mostly […] looking at the small counts.
	[…]
Researcher1:	Can you explain why you answered three here?
StudentL18:	Because the first one [bar] was really long anyway and the rest were all pretty small. So to my feeling that made more sense. Because it was kind of in the middle as such. Not very far from the middle [of the horizontal scale].

Most strategies applied to single histogram items (55%) were incorrect, namely case-value plot and count-and-compute strategies. During recall, some students reported a strategy that returns the median—dividing the area into two equal parts, see the excerpt of studentL11 below—or the mode—the position of the highest bar, in line with the literature (Frischemeier & Biehler,

2016; Watson & Moritz, 1999). Both strategies were associated with the same vertical scanpath, and, therefore, reported as a histogram strategy.

StudentL11:   A bit where the area was somewhat the same already from the right, so [the point on the horizontal axis] where the areas cancel each other out [are equal].

3.4.2 Strategies on double graph items

In 50% of the trials with double histograms, students used a histogram strategy for comparing the means (Table 3.8). Most students used a distribution-informed strategy using specific features of the graph such as: same symmetry and positions of the bar thus same mean, or similar shape but moved to the right thus higher mean (Figure 3.7) in short: using 'shift' and 'shape' (Frischemeier & Biehler, 2016). This has some similarities with what Khalil named 'local slices' (2005). However, in contrast to Khalil, our participants saw the full graph at once. The 'typical' strategy in which students estimate means—similar to what is used for single graphs—is rare for double graphs. A distribution-informed histogram strategy—shift—is described for double histogram Item09 (Figure 3.7, top):

StudentL22:   So his [postal worker Kees] mean is always one step higher anyway and [postal worker] Rianne's one step lower because the [Kees'] graph shifts one step [to the right].

Table 3.8 Strategies, percentage of trials (N = 150) per item type (correct strategy in **bold**)

	Histogram strategy	Case-value plot strategy	Count-and-compute strategy	Unclear
Double histograms	**50%**	49%	0%	1%
Double case-value plots	9%	**87%**	**3%**	1%

In trials with double case-value plots, students frequently (87%) used a case-value plot strategy for comparing means. Most students applied a distribution-informed strategy using specific features of the graph such as same shape and number of bars, double heights, or more area. In contrast to histograms, in case-value plots, more area does indicate a higher mean (see also the discussion on the use of totals in Watson & Moritz, 1999). A count-and-compute strategy—similar to the single graph items—was also found for double graph items but much less frequently (3%).

To our surprise, in 9% of the double case-value plots trials students used a distribution-informed histogram strategy (Table 3.8), resulting in an incorrect answer (Figure 3.8: same instead of Renesse). As a student stated:

StudentL11:   [...] the same, because the graphs are almost the same, only the left graph has two extra bars on the outside, but if you average it, they cancel each other out as well.
Researcher1:   And why [...]?
StudentL11:   Because they are exactly the same [symmetry]. But I now see that is not correct because it is obviously not a frequency [on the vertical axis].

3.4.3   Students' answers compared to strategy correctness

Regarding answer correctness (Table 3.9) compared to strategy correctness, case-value plots were solved considerably better than histograms and usually also with a correct strategy. Case-value plot strategies were often used for histogram items, resulting in both lower correct answers and lower correct strategies. The discrepancy between answers and strategies (Table 3.9) is mostly due to students' preferences for whole and half numbers. For example, single case-value plot Item04 (Figure 3.5, middle/right) scored the lowest of these plots (76%). For Item04, ten students answered 5—just outside the answer range [2.6–4.8]—and two students overestimated the mean by answering 5.5 and 6.

Table 3.9 Overview of the percentage of correct answers (*N* = 150 trials) and strategies per item type

Item type	Correct answers	Correct strategy
Single histogram	43%	43%
Single case-value plot	83%	100%
Double histograms	39%	50%
Double case-value plots	74%	90%
Total	60%	71%

Students performed best on the three single case-value plots items—on average, in 83% of the trials they correctly estimated the mean and scored highest on Item08 (90%). Of the three double-case-value plot items, Item10 scored low, with 40% of the students answering correctly. Gaze data revealed that there was no difference between students who answered incorrectly and those who answered correctly regarding the fixations on the white space (Figure 3.8) above the names Alex and Yves, as both groups had or did not have fixations in these areas. Only two students explicitly mentioned that Alex and Yves did not collect any litter, but they were initially confused by the

graphs—see the excerpt below. Several students did not understand the role of 'bars' with frequency zero, which is in line with the literature (e.g., delMas & Liu, 2005).

StudentL07   Yes, I had my doubts about this one too, because there is a bar missing on the left and then I want to compare it and that is not possible. [...] then I start to doubt easily whether it is good or not that bar is there and then I thought well if the bar is not there then in this case the low numbers are not there so then the average weight of the right one will be higher. Because it is less distributed. It's all the same numbers, so to speak.

Researcher1:   You said Renesse, so the left one [...]. While on the left, the two bars are extra compared to the right.

StudentL07:   Oh no never mind, average weight. No then because on the left those bars are there, it means that something was picked up at all instead of nothing on the right.

Figure 3.8 White space (ovals) above the names Alex and Yves (right hand graph in Item10)
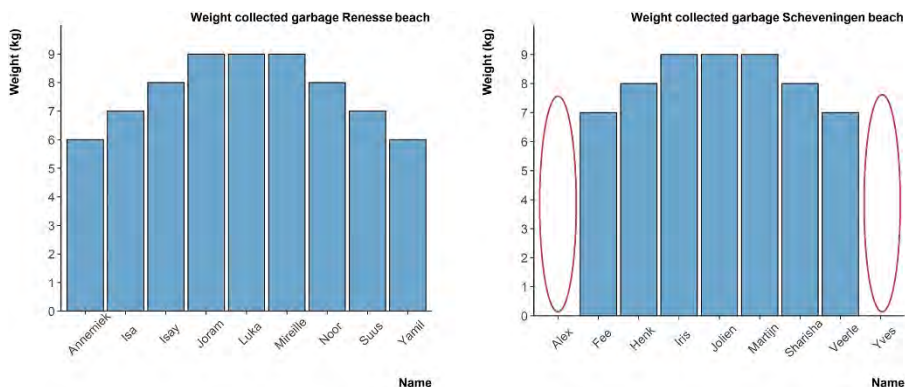
3.5   Conclusions and discussion

In this study, we investigated how and how well Grades 10–12 pre-university track students interpret histograms and case-value plots. We found five strategies. Our study confirmed the two most common strategies for estimating the mean from single graph items found in the pilot study—a typical case-value plot strategy associated with horizontal gazes and a typical histogram strategy associated with vertical gazes. A third, new 'typical' count-and-compute strategy was found—in line with other findings (e.g., Watson & Moritz, 1999)—for both single and double graphs, and associated with a zigzag scanpath. Although this could be seen as a variation of the case-value plot strategy, we decided to report this separately due to its algorithmic character.

These three typical strategies for estimating the mean from graphs were rarely used for double graph items.

Furthermore, two new strategies were found for comparing means by considering specific features of the graphs. In a distribution-informed histogram strategy, students compared 'shifts', ranges and symmetry (spread, shapes, cf. Frischemeier & Biehler, 2016). To our surprise, some students applied a distribution-informed case-value plot strategy on double histogram items. These strategies are item-specific, as scanpaths differ depending on the distribution. Furthermore, some students ignored bars with frequency or measured value zero, even when they looked at them (cf. delMas & Liu, 2005).

Discussing these findings, first, we note that several students talked about a mathematical object not present in the item (e.g., horizontal or vertical line). Although imagined, this object played a role in finding the mean. This imagined object was visible in the gaze data on the graph area of several students on single as well as some double graph items. When students estimated the mean from the graph (for single or double graphs), the same object was used for similar graphs. When comparing means in the double graph items, the object was often item-specific and included for example the 'bars' with zero frequency or weight or range. This is in line with findings from other studies in which objects appeared in the form of triangles, center points, or lines (Alberto et al., 2022). Further research, for example, using machine learning algorithms or latent class or profile analysis (e.g., Hickendorff et al., 2018), is needed to find out whether students can be grouped meaningfully purely based on their scanpaths on the graph area and be related to the framework of Frischemeier and Biehler (2016).

Second, as many students tended to focus on most noticeable features, the different signs—case-value plots and histograms—are perceived as similar. Our results might indicate that perception-action loops (Shvarts et al., 2021) for case-value plots are stronger—or older—than for histograms leading to both graphs being recognized as case-value plots. Future research is needed to find out how specific perception-action loops for histograms can be built.

Third, we speculate that our gaze data can also be re-used to underpin theoretical ideas about embodied aspects of thinking processes in learning mathematics. Piaget (1952) showed that learning starts from reflexes. Sensorimotor experiences (e.g., touch, vision) induced learning through accommodation to the environment, assimilation (incorporating objects that fulfill what is needed and rejecting objects that do not), and individual organization which "exists, inasmuch as organization is the internal aspect of this progressive adaptation" (p. 41). According to Vygotsky, thinking is an especially complex form of behavior (1926/1997). For example, Vygotsky describes an experiment with participants sitting between two objects with

their eyes closed. When asked to think "hard of either of these objects […] The movement of the [participant's] eyes and the straining of his muscles always occurred in the direction his thinking was aimed at." (p. 155). Here, Vygotsky associates specific eye movements—motor actions—with thinking processes. Moreover, stronger and more focused thoughts are associated with clearer and more complex motor actions. This is, according to Vygotsky, also true for mathematical thoughts. For example, young children performing addition or subtraction move their lips, tongue, forehead or cheeks. "Even the most abstract thoughts of relations […] are related ultimately to particular residues of former movements now reproduced anew." (p. 162). Residues of movements can be found in students' gazes—the horizontal line representing the estimated mean—which are sometimes described by students as the line that makes all bars equally high. Students' language described imagined motor actions in line with their gestures as if these occurred. What is particularly new in our study—compared to, for example, the research on trigonometry tasks (Alberto et al., 2022)—is that we found residues of movements even though students could not perform the described action. Eye movements can, therefore, be used as evidence of sensorimotor coordinations that constitute and contribute to mathematical competencies (Abrahamson et al., 2015).

3.5.1 Education

For educational practice, insight into students' graph-based reasoning may contribute to new perspectives on teachers' own thinking processes. Teachers sometimes also misinterpret histograms (Boels et al., 2019b; Dabos, 2014). Therefore, awareness of fundamental differences between various types of graphs that share similar most noticeable features is an important part of teachers' Statistical Knowledge for Teaching (Groth, 2013).

   Second, gaze data provide fine-grained information on students' where students looked. Although the relation between a scanpath and students' thinking processes is not straightforward, the present study support the growing body of literature that gaze data can reveal students' reasoning (e.g., Lai et al., 2013). This allows educators to link more closely to students' thinking processes. For instance, teachers can use students' scanpaths for drawing students' attention to correct interpretations of graphs and pay explicit attention to relevant—instead of most noticeable—differences between graphs. As several students explicitly looked at the axis labels and still used an incorrect strategy, this implies that just telling students to carefully read the axis labels might not be enough. We also noticed that some students checked their answer prior to reporting it. Although such findings are beyond the scope of this study, eye-tracking seems well-suited for disclosing such thinking processes. This is left for future research.

Third, the present study also shows that the statistical key concept of data (e.g., number of variables and measurement level, Chapter 2, see https://youtu.be/zpRHhixoYmg and https://youtu.be/5od2uB908PI) should be extended with an understanding of on which axis the data values are depicted—being mostly the horizontal axis in a histogram, with very few exceptions (e.g., age-gender-pyramid). Not fully understanding the key concepts of data and distribution can lead to several misinterpretations (Chapter 2). When students applied a case-value plot interpretation strategy to a histogram, this related to several misinterpretations, including that the frequency is mistakenly seen as the measured value (Chance et al., 2004), that the number of bars is confused with the number of cases (Dabos, 2014), and that the mean of the measured value is mixed with the mean of the heights of the bars or frequencies in a histogram (Lem et al., 2014b). As these are all related to how and where the data and the distribution of these data are depicted in histograms, future interventions should aim to support students in understanding these concepts in histograms. In addition, teachers could ask students to explain their strategies to promote reflection.

3.5.2 Eye-tracking

A first methodological aspect of eye-tracking is that—compared to students' verbal reports—more details of their problem-solving processes are visible. For example, an imaginary horizontal line was visible in many more students' gaze data on single graph items than was reported by students. Furthermore, some students were unable to correctly report their strategy in retrospect, even when cued with their gaze. Hence, gaze data can reveal approaches that students are not aware of or are unable to articulate. The subtle differences in scanpaths for double graph items as well as the item and question-dependent scanpaths emphasize both the need for triangulation and a domain-specific interpretation of gaze data (e.g., Schindler & Lilienthal, 2019). For the double graph items, we did not find a consistent way to distinguish between a strategy involving estimating means, on the one hand, and a distribution-informed strategy, on the other, based solely on gaze data. A possible future line of research is to explore whether these variations can be identified with other analysis methods.

Second, eye-tracking may also influence students' thinking processes less than thinking aloud. We noted, for example, that in retrospect, when explaining—thus reflecting on—their strategy, several students realized that they took an incorrect approach.

A third methodological aspect of our study is that we used spatial eye-tracking measures: scanpaths on one AOI (graph area). Scanpaths are complex data that usually require qualitative and labor-intensive analysis. Nevertheless,

our study shows that spatial measures can reveal task-specific strategies that would have stayed hidden using more traditional measures such as time on—or count of transitions between—AOIs only (cf. Hyönä, 2010). Cued recall data revealed that a correct vertical scanpath is not only associated with estimating the arithmetic mean but also with the median and mode. Future research is needed to investigate whether these can be distinguished, for example, through using machine learning algorithms to analyze raw gaze data or heatmaps.

Fourth, gaze data—scanpaths in particular—can potentially shed new light on tenacious didactical problems in mathematics teaching—including students consistently misinterpreting histograms. We speculate that this holds not only for other misinterpreted graphs, including boxplots (Lem et al., 2013c, 2014a), violin plots, scatterplots, density curves (Nolan & Perrett, 2016), stacked-dotplots (Lyford, 2017), increase diagrams, network topologies and function graphs (Leinhardt et al., 1990) but also for other mathematical topics where scanpaths may play an important role: line and point symmetry in functions, congruency of triangles, the relation between a straight line, axis scales (logarithmic, linear, normally distributed), and functions, and maybe even the representation of a cubic and hexagon.

Appendix A   Background of the eye-tracking and participants' data

A.1   Details of the eye-tracker

A Tobii XII-60 with a sampling rate of 60 Hz was placed on an HP ProBook 6360b laptop with a magnetic strip between the laptop's 13-inch screen (refresh rate 59 Hz) and keyboard. A chin rest was used for reduced data loss and improved accuracy of the gaze data. The Tobii XII-60 uses Pupil-Corneal Reflection (see Tobii's user manual, n.d.-a and Holmqvist et al., 2023). Tobii's eye-tracking software automatically uses both bright and dark pupil methods during calibration and, according to the product specifications, the software automatically chooses the most suitable method.

By using harmless infrared light, the Tobii can detect where people look. In this study, the Tobii Pro Studio software (Tobii, n.d.-a) recorded students' gazes on the screen in real-time. The distance between the screen and the participant was 55–60 cm [mode and mean: 59 cm].

A Røde microphone was used to record the cued recall (verbal data). More details of the set-up are shown in Figure A.1. The participant used a height-adjustable office chair to ensure a comfortable position on the fixed chin rest.

Figure A.1 Set up of the laptop, Tobii eye-tracker, chin rest and microphone

The live viewer mode was enabled and used on a second screen that was turned to the researcher and not visible to the participant. The raw gaze data were exported with the data export function of the Tobii Studio Pro 3.4.5 version. Each participant's data were exported in a separate .tsv-file (tab-separated, comparable to a .csv-file) readable in, for example, Excel.

A.2   Gaze data quality

Getting good data quality can be hard in an eye-tracking study. The quality is influenced by the design of the experiment (e.g., a chin rest usually improves data quality) and by the characteristics of the participants (e.g., wearing glasses). We used a chin rest and asked participants not to wear mascara, although a few of the participants did.

The calibration procedure consisted of a 9-point calibration on the screen provided by the Tobii Pro Studio 3.4.5 software. The calibration procedure was repeated for specific points on the screen if the gaze was too far off the required spot. As the Tobii software does not supply specific quality measures for the calibration, this was based on expert decision.

The Tobii software has no built-in validation procedure. A validation screen was, therefore, included in the set-up at the start, after each item, and at the end. Directly after the calibration by Tobii, students were instructed to look at the middle of a series of four single dots appearing on the screen. Each of these marked important positions in the item (Figure A.2), and each time, only one dot was visible at the same time in a fixed order.

Figure A.2 A 4-point validation procedure was applied before and after the series of items, left: position of points, right: letters indicating the order of points on the screens
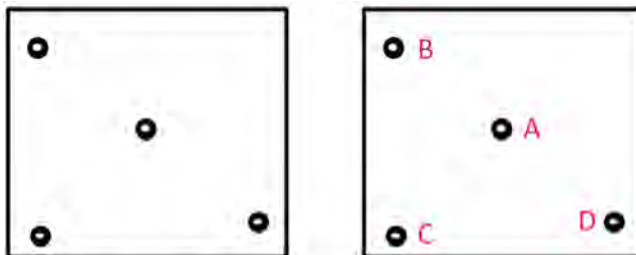
After each item, a validation cross appeared on the screen. Participants were instructed to look at the middle of this cross (Figure A.3). This validation cross was positioned at the right-hand side of the screen, in different positions each time. This ensures that the first eye movement is never accidentally on the graph area.

Figure A.3 Fixation cross that appeared somewhere on the right-hand side of the screen

Data loss is normal during an eye-tracking experiment and will appear during blinking or looking away from the screen. Furthermore, epicanthic folds (almond eyes), wearing glasses, contact lenses and makeup can also lead to extra data loss. Although some of these conditions applied to some students, we did not exclude any student from the data set as the data loss per trial (averaged over all participants) and the data loss per participant (averaged over all items) were below the exclusion point (34% or more; Table A.1). In some cases, there was too much data loss for a specific trial, for example, because the student had accidentally clicked through the item or because the track of the gaze data got lost. If this happened, an attempt was made to identify the strategy. If this was impossible, it was indicated that the strategy was unclear and marked as unknown (type of strategy) and incorrect (correctness of strategy). Two students reported wearing glasses, three students reported wearing contact lenses. In total, 25 trials per participant were recorded; only the first twelve trials are reported in this study.

Table A.1 Data loss

	Average per participant over all 25 trials	Average per trial over all 50 participants, first twelve trials
Average data loss	7.21%	6.58%
Minimum data loss	1.00%	2.48%
Maximum data loss	27.78%	12.04%

In response to recent calls from researchers to provide more insight into the quality of eye-tracking data, we provide measures of accuracy and precision of the gaze data (Holmqvist et al., 2012, 2023; Strohmaier et al., 2020). Accuracy and precision are especially important when using temporal or count measures, as—for example—a low accuracy can indicate that fixations are classified to a different AOI than where a participant was actually looking. For a qualitative study like ours, where scanpaths from videos of the eye movements are used, accuracy and precision have less influence on the final results.

Furthermore, we placed the most important part of the task (graph area) near the center of the screen where accuracy was expected to be higher than in the corners. The corners of the screen were mostly left empty or used for less important parts of the tasks where accuracy does not play a role (e.g., the 'next' button).

The four validation screens at the beginning and end were used to calculate accuracy and precision. To determine accuracy, the offset was calculated: the distance from the closest fixation (gaze-point) to a validation point (Figure A.2 and Table A.2). "The offset gives an indication of the discrepancy between gaze position as reported by the eye-tracker and the assumed gaze position of the participant." (Van der Stigchel et al., 2017, p. 3588). We used the raw x- and y-coordinates of a fixation (gaze-point) closest to the validation point (x- and y-coordinates averaged over the two eyes by the Tobii software, which can be found in the columns named GazePointX (ADCSpx) and GazePointY (ADCSpx)). Only fixations within 170 pixels (3.5°) of the center of the validation point were considered. The mean accuracy is 56.6 pixels (1.16°). As can be calculated from Table A.2, the mean accuracy for the validation points at the start does not differ much from the end (4.9 pixels or 0.10°). As expected, the accuracy of the middle point of the screen (A, 11.4 and 15.3 pixels or 0.23° and 0.31°) is better than for the points in the corners of the screen (B–D).

Table A.2 Accuracy and precision, averaged over all students except studentL50

Validation point	X position (px)	Y position (px)	Accuracy px	deg	Precision px	deg
A – start	683.0	384.0	11.4	0.23	16.8	0.34
B – start	228.5	175.5	55.9	1.15	27.7	0.57
C – start	178.5	592.5	74.3	1.52	31.1	0.64
D – start	1187.5	542.5	74.9	1.53	33.9	0.69
A – end	683.0	384.0	15.3	0.31	17.1	0.35
B – end	228.5	175.5	69.0	1.41	25.9	0.53
C – end	178.5	592.5	79.2	1.62	35.1	0.72
D – end	1187.5	542.5	72.7	1.49	38.4	0.79
Mean			56.6	1.16	28.3	0.58

Precision was calculated as follows. The fixation (gaze-point) closest to the validation point—see the calculation of accuracy—was used to determine which fixations the Tobii Studio software considers as belonging to that point; Tobii software column names FixationPointX (MCSpx) and FixationPointY (MCSpx) were used as filters for the gaze-points. The root-mean-square sample-to-sample (RMS-S2S, Holmqvist et al., 2023) distance (deviation) across all fixations (coordinates of the gaze-points, columns GazePointX (ADCSpx) and

GazePointY (ADCSpx)) was taken over the raw *x*- and *y*-coordinates of these gaze-points to give an indication of the noise. This was done per student for each validation point. This RMS-S2S was then averaged over all students per validation point. StudentL50 was excluded from these calculations due to technical problems. Under optimal conditions, the Tobii XII-60 has 0.4–0.5° accuracy (with artificial eyes) and 0.32° precision Tobii (n.d.-b). As can be expected, the offset for human eyes is higher (1.16°) but is nevertheless within the acceptable accuracy limits for this type of eye-tracker given that the most important part of the stimulus is near the middle of the screen. The precision (0.58°) is considered good. The RStudio-code for these calculations can be found in the data repository.

A.3   Self-reported grades

Students reported their grades (Table A.3).

Table A.3 Arithmetic mean of self-reported mathematics scores (scale 1–10; lowest–highest) for choice of mathematics and grade levels

Mathematics	Grade 10	Grade 11	Grade 12	Unknown	Total
A	6.0	6.4	5.4	6.2	6.1
A and B			7.1		7.1
B	6.1	7.0	7.8		6.8
B and D	7.5	6.7	7.7		7.3
C		6.0			6.0
Total	6.2	6.6	7.0	6.2	6.5

Note. In Dutch schools, a score of 5.5 or above is regarded sufficient.

A.4 Items and students' answers

A.4.1 Single graph items and students' answers
In table A.4 an example of students' answers can be found.

Table A.4 Students' answers for Item01

Answer	Number of students (percentage)
1	1 (2.0%)
2	2 (4.1%)
3	**5 (10.2%)**
3.5	**5 (10.2%)**
4	**9 (18.4%)**
4.25	**1 (2.0%)**
4.5	2 (4.1%)
5	2 (4.1%)
6	3 (6.1%)
7	4 (8.2%)
8	8 (16.3%)
8.5	2 (4.1%)
9	1 (2.0%)
9.5	1 (2.0%)
10	2 (4.1%)
11	1 (2.0%)
Total	49 (100%)

Note. The mean is 3.3; **bold** answers are noted as correct as they are within the range mean +/- 1.1; here: [2.2, 4.4]. One student accidentally skipped this item.
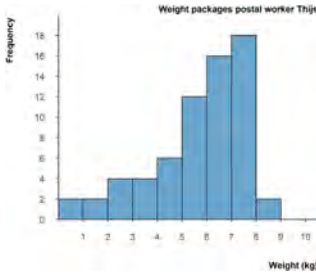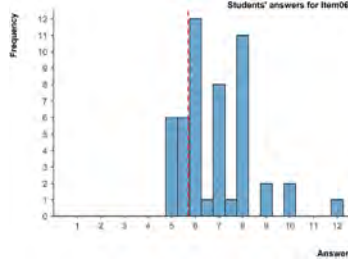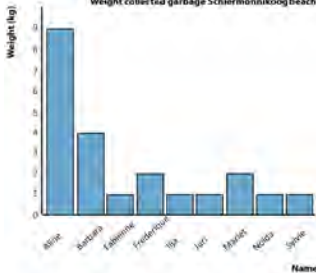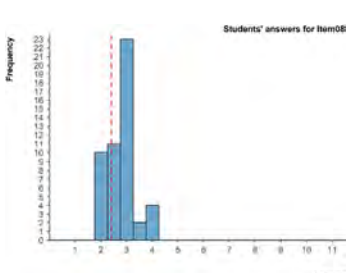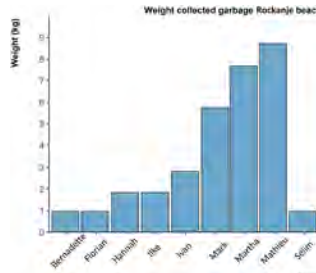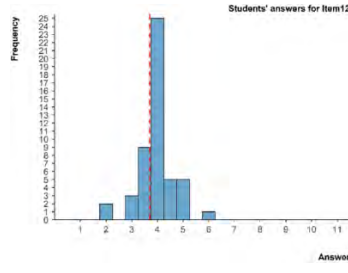
Table A.5 provides an overview of all the answers given by the 50 secondary school students during the eye-tracking study for the single graph items. The original graph for which the students were asked to estimate the arithmetic mean is given in the column *Graph in item*. The items were constructed in line with the recommendations from Orquin and Holmqvist (2017) so that the "stimuli […] differ systematically on one or more features" (p. 6). The last column shows histograms of students' answers. Most students rounded to half and whole numbers. Table A.8 provides an overview of the number of correct answers per item for all items. The range for correct answers in single graph items was based on the following criteria. First, we noted that students' estimations were most often whole numbers and numbers rounded to the nearest half (e.g., Table A.4). Second, we noted that the range in answers of three experts in a small expert pilot was within +/-1.2 of the exact answer. Unlike the students, these experts tended to round their answers to one decimal. Third, we chose the range so that answers found by applying an incorrect strategy (e.g., a case-value plot strategy applied to a histogram)

would not be part of the range. As a result, all answer ranges for correct answers were set to the mean +/- 1.1.

Table A.5 Students' answers on single graphs (open answers)

Item	Graph type	Graph in item	Correct mean (answer range: mean +/-1.1)	Histogram of students' answers (bin width: 0.5). The dotted red line indicates the correct mean of the graph in the item
Item 01	Histogram	Weight packages postal worker René	3.3	Students' answers for Item01
Item02	Histogram	Weight packages postal worker Anton	2.7	Students' answers for Item02
Item04	Case-value plot	Weight collected garbage Terschelling beach	3.7	Students' answers for Item04

Item	Graph type	Graph in item	Correct mean (answer range: mean +/-1.1)	Histogram of students' answers (bin width: 0.5). The dotted red line indicates the correct mean of the graph in the item
Item06	Histogram		5.7	
Item08	Case-value plot		2.4	
Item12	Case-value plot		3.7	

A.4.2   Double graph items and students' answers

Table A.6 provides an overview of all the double graph items. Table A.7 provides an overview of all students' answers for these items.

Table A.6 Overview of double graphs items (multiple-choice)

Item	Graph type	Graphs in item
Item03	Case-value plot	
Item05	Histogram	
Item07	Case-value plot	

Item	Graph type	Graphs in item
Item09	Histogram	Weight packages postal worker Kees / Weight packages postal worker Rianne
Item10	Case-value plot	Weight collected garbage Renesse beach / Weight collected garbage Scheveningen beach
Item11	Histogram	Weight packages postal worker Ellen / Weight packages postal worker Titia

Table A.7 Overview of students' multiple-choice answers on double graphs items (correct answers in **bold**)

Item	Graphs type	Answer (count)	Answer (count)	Answer (count)
Item03	Case-value plots	Texel (0)	**Cadzand (47)**	Same (3)
Item05	Histograms	Julia (11)	Willem (24)	**Same (15)**
Item07	Case-value plots	Kijkduin (1)	Zandvoort (5)	**Same (44)**
Item09	Histograms	**Kees (25)**	Rianne (5)	Same (20)
Item10	Case-value plots	**Renesse (20)**	Scheveningen (24)	Same (5)
Item11	Histograms	Ellen (2)	Titia (30)	**Same (18)**

A.4.3 Item order, item type, and number of correct answers

In Table A.8, the order of the items, item, and graph type, as well as the number of students answering correctly or incorrectly, are given. A fixed item order was used in this study, with never more than two graphs of the same type (histogram or case-value plot) in succession. If no answer was given (two students each accidentally skipped one question), this was noted as incorrect. As we expected most students who confuse case-value plots with histograms to apply a case-value plot strategy to a histogram, we started with two single left-skewed histograms. This was done to avoid priming (e.g., Lashley, 1951). Graphs with the same *most noticeable* features (e.g., Item02 and Item08) never directly followed one another.

Table A.8 Students' answers correctness per item

Item	Item type	Graph type	Correct	Incorrect	Total
Item01	Single	Histogram	20	30	50
Item02	Single	Histogram	19	31	50
Item03	Double	Case-value plots	47	3	50
Item04	Single	Case-value plot	38	12	50
Item05	Double	Histograms	15	35	50
Item06	Single	Histogram	25	25	50
Item07	Double	Case-value plots	44	6	50
Item08	Single	Case-value plot	45	5	50
Item09	Double	Histograms	25	25	50
Item10	Double	Case-value plots	20	30	50
Item11	Double	Histograms	18	32	50
Item12	Single	Case-value plot	42	8	50
Total			358	256	600

A.4.4   Percentage of correct answers per item type and occurrence of 'zero' bars

In Table A.9, the items are clustered by item type. For each item, the percentage of correct answers is given, as well as the occurrence of bars with a measured weight or frequency zero. For Item10, the low scores are due to not understanding the role of the two bars with measured value zero, see the Results section.

Table A.9 Students' answers correctness per item

Item	Item type	Bars with frequency or measured value zero	Correct
Item01	Single histogram	No	40%
Item02	Single histogram	No	38%
Item06	Single histogram	No	50%
Item05	Double histograms	No	30%
Item09	Double histograms	Yes	50%
Item11	Double histograms	Yes	36%
Item04	Single case-value plot	No	76%
Item08	Single case-value plot	No	90%
Item12	Single case-value plot	No	84%
Item03	Double case-value plots	No	94%
Item07	Double case-value plots	Yes	88%
Item10	Double case-value plots	Yes	40%

A.5   Codebooks and detailed results coding

Note that all codebooks and coding results are available in a data repository.

A.5.1   Codebook general

The gaze and verbal data were coded separately. Next, these codes were combined into a final code using the verbal data—if available—as a triangulation of the gaze data, see the Data analysis section for more explanation. From the pilot study, it appeared students did not change their strategies over the first twelve trials (Boels et al., 2018). Therefore, items of the same type (e.g., all single case-value plot items) were coded directly one after another in the same coding session per participant so that the coding per participant was kept as consistent as possible. For example, items 1, 2, and 6 (all with single histograms) were coded for one participant and then for the next participant, and so on. In the next coding round, items 4, 8, and 12 (all with single case-value plots) were coded for one participant, and so on. Participants where the coding indicated a switch in strategy between items but

within a series of one type of item were reconsidered after all participants were coded for this type of item to make sure strategy switches between items were not due to coding inconsistency. In the case of a switch of strategy during a trial, only the code of the strategy that was used just before the answer was given was used. The students volunteered for the cued recall, and after collecting gaze data from the first 27 students, we completed the 25 cued recalls on several items, depending on the time left before the 50 minutes (one lesson) had passed. The gazes were coded from videos (available on request from the first author). A still of such a video as well as the static gaze-plot for these gazes are shown in Figure A.4.

Figure A.4 Video still (left) and static gaze-plot (right) of studentL49's gazes on Item06
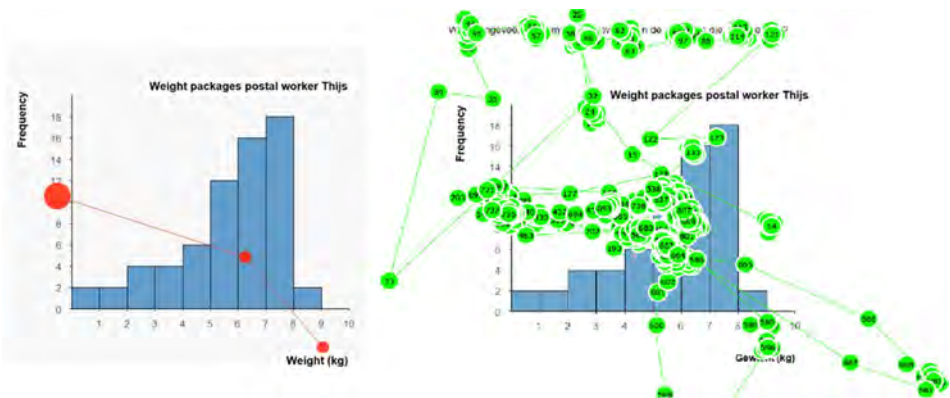
Note. The full video of L49's gazes on Item06 can be found on the publisher's website. The static gaze-plot (right) contains all gazes of this student on this item. Although the mean weight of the packages is approximately 5.7 kg, the student answers 9 [kg] even though this student looked at the word 'Gewicht' (Weight) along the horizontal scale. We removed sound (e.g., a student saying the answer) from the original video for privacy reasons.

A.5.2   Codebook double graph items
Codebook for the gazes on double graph items
In the double graph items, the differences between a case-value plot strategy and a histogram strategy are subtle. Therefore, the answer is considered (Table A.10). When answer and gaze data lead to different codes, the gaze data are leading.

Table A.10 Codebook for the gazes on double graph items. Dh = double histogram, dcvp = double case-value plot

Code	Assign this code if	
	The gazes are	The answer is
Histogram strategy	- on the top of bars in both graphs, indicating comparing double heights - look at (max) frequency 8 and 16 in the relevant graphs - fixations on or around the origin of graphs - going back and forth along the horizontal axis (range) - on the last bars or outer bars of each graph - see the codebook for single graph items	Same or Texel (Item03 – dcvp) Willem (Item05 – dh) Zandvoort or Kijkduin (Item07 – dcvp) Kees (Item09 – dcvp) Same (Item10 – dcvp) Same (Item11 – dh)
Case-value plot strategy	- back and forth between the top of the bars - on the bars with value or frequency zero - on the top of last bars or outer bars of each graph - are not on the bars with value or frequency zero where this is relevant - see the codebook for single graph items Do not assign this code if one of the distribution-informed histogram strategy options hold	Cadzand (Item03-dcvp) Julia or same (Item05-dh) Same (Item07 – dcvp) Same or Rianne (Item09 – dcvp) Renesse or Scheveningen (Item10 – dcvp) Titia or Ellen (Item11 – dh)
Count-and-compute strategy	- see the codebook for single graph items There is a sound of counting during trial	n.a.

Codebook for the verbal data for double graph items
The codebook for the verbal data for double graph items is given in Table A.11.

Table A.11 Codebook for the verbal data on double graph items

Code	Assign this code if
	a student talks about
Histogram strategy	- extra bars outside compensate each other - higher mean through extra bar on the left (overlooking the extra bar on the right in the other graph) - same range, double frequency - one graph is shifted to the right compared to the other graph - see the codebook for single graph items

Case-value plot strategy	- the two zero bars or missing bars lower the mean
	- comparing the heights of the bars
	- higher bars in one graph thus a higher mean
	- less bars thus a higher mean
	- more bars thus a higher mean
	- more area thus a higher mean
	- same number of bars and same heights
	- similar bars but reordered
	- see the codebook for single graph items
Count-and-compute strategy	
	- see the codebook for single graph items

A.5.3 Results final coding strategies

The qualitative analysis of the coded data led to a decision of what kind of strategy was used by each student (Table A.12), and whether the student used a correct or an incorrect strategy (Table A.12). This is not necessarily the same as giving a correct answer. A student may, for example, use a correct strategy to estimate the mean but then make an error in the estimation itself which might lead to an incorrect answer. The codes of the gaze data were combined with the verbal data. If the verbal and gaze data did not align, the gaze data prevailed. When the data were unclear, the coder decided on the strategy's correctness. Strategies were coded unclear when a student accidentally skipped a question (two students each skipped one item), when there were not enough gaze data (two students each on one item), and when the verbal data and gaze data did not align and the first coder could not reach a decision (two students, one item).

Table A.12 Final coding strategy type. In **bold**: number of trials (N = 50) in which students use a correct strategy for this item

Item	Histogram strategy	Case-value plot strategy	Count-and-compute strategy	Unclear	Total
Item01	**24**	15	10	1	50
Item02	**23**	18	7	2	50
Item03	2	**45**	**2**	1	50
Item04		**38**	**12**		50
Item05	**23**	27			50
Item06	**19**	21	10		50
Item07	5	**43**	**2**		50
Item08		**36**	**14**		50
Item09	**30**	19		1	50
Item10	6	**42**	**1**	1	50
Item11	**22**	28			50
Item12		**33**	**17**		50
Total	153	366	75	6	600

Note. For the single graph items, most students used a 'typical' strategy; for double graph items, most students used a 'distribution-informed' strategy. We did not find a consistent way for coding these variations in scanpaths qualitatively, so we only distinguished whether they used a strategy interpreting the graph as a histogram or as a case-value plot.

Table A.13 Strategy correctness of students – number and percentage of trials (N = 50)

Item	Item type	Graph type	Correct	Incorrect	Unclear	Total
Item01	Single	Histogram	24 (48%)	25	1	50
Item02	Single	Histogram	23 (46%)	25	2	50
Item03	Double	Case-value plots	47 (94%)	2	1	50
Item04	Single	Case-value plots	50 (100%)			50
Item05	Double	Histograms	23 (46%)	27		50
Item06	Single	Histogram	19 (38%)	31		50
Item07	Double	Case-value plots	45 (90%)	5		50
Item08	Single	Case-value plot	50 (100%)			50
Item09	Double	Histograms	30 (60%)	19	1	50
Item10	Double	Case-value plot	43 (86%)	6	1	50
Item11	Double	Histograms	22 (44%)	28		50
Item12	Single	Case-value plot	50 (100%)			50

A.5.4 Static gaze-plots examples for specific coding categories

In the gaze-plots below, some examples of typical, distribution-informed and count-and-compute strategies for double graph items can be found (Figures A.5, A.6, A.7, A.8).

Figure A.5 Correct typical strategy for comparing the means of two case-value plots (horizontal line segment superimposed for convenience of the reader)
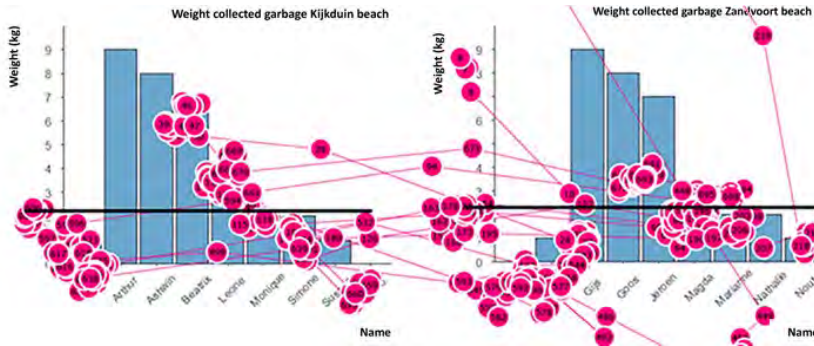
Below, a distribution-informed histogram strategy: similar shape and double heights thus same mean (Figure A.6).

Figure A.6 A correct distribution-informed strategy for comparing the means of two histograms: using similar shape and double heights
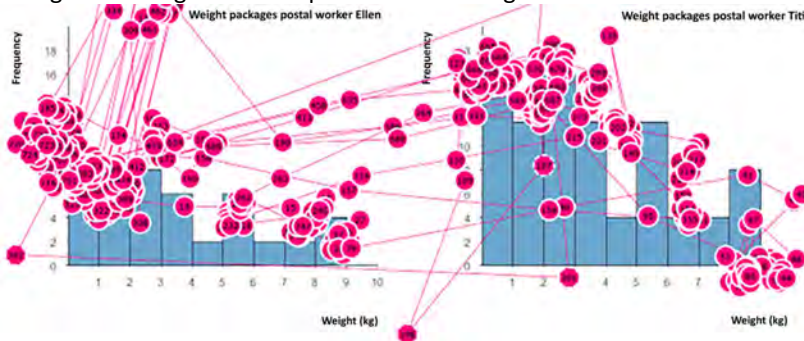
Note. This distribution-informed histogram strategy is only correct if students conclude from the double heights that means are the same.

Figure A.7 A correct distribution-informed strategy for comparing the means of two case-value plots: comparing the heights of the (outer) bars in and between the graphs
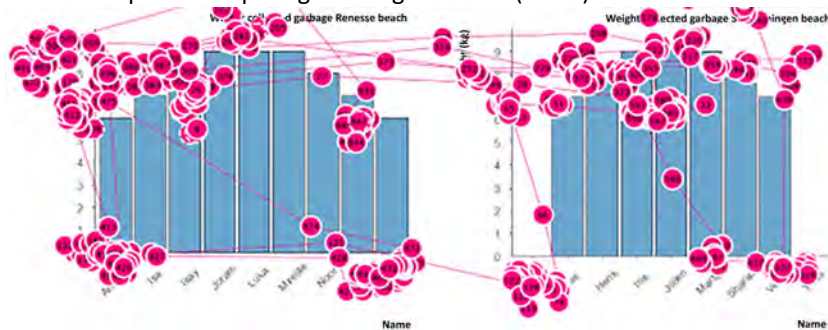
Figure A.8 A correct count-and-compute strategy applied to case-value plots
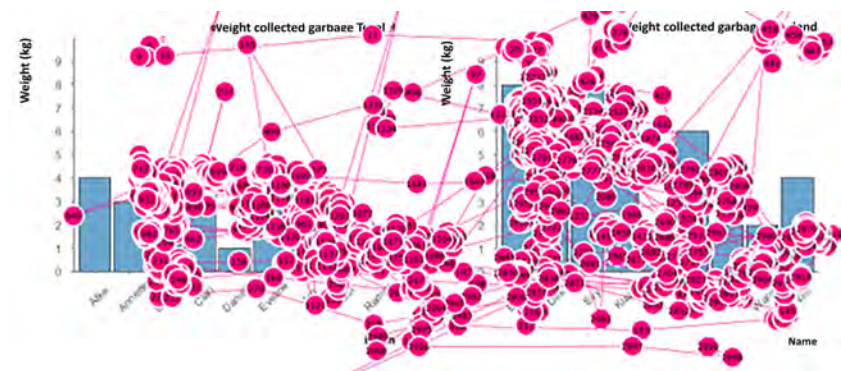
Note. Characteristic for this scanpath is the zigzag form on each graph area.

A.5.5 Examples of transcripts

Below are some transcripts with a short explanation. During the cued recall, several students noted that they thought that weight was on the vertical axis in the histograms. Some students talked about a horizontal line, making all bars equal or the same area above and underneath an imagined line—all case-value plot strategies. Some gestured a horizontal line, for example for Item01:

StudentL15:	Then I think I flipped the axes in my head [...]
Researcher1:	Okay, well you came to the answer 'four', how did you come to that answer 'four'?
StudentL15:	That was about in the middle.
Researcher1:	In the middle of what?
StudentL15:	In the middle of the graph.
Researcher1:	Can you point to where you are pointing if you mean that?
StudentL15:	These bars.
Researcher1:	Yes, okay so you are now pointing to a horizontal line at the height of four.

From the gaze data, this horizontal line was frequently visible in the eye movements on the graph area of the single graphs for both histograms and case-value plots (see Figure 3.2 in this chapter for an example from the pilot study).

For single histogram Item02 a student stated:

StudentL10:   Yes, just read the question first and then I did the same as before. So try to make everything the same length.

Many students who answered incorrectly thought that the outer two bars in the Renesse graph (a case-value plot) lowered the mean, see for example the transcript of studentL03:

StudentL03:   Well, you saw that they were both the same height in the middle. Only that Scheveningen had no lowers at the beginning and at the end, so then it is logical that it is a bit higher because two lowers are taken off. From the average.

A.6   Instruction to the participants

The following instruction was part of the letter accompanying the informed consent:

What do we study?
Graphs are everywhere—in newspapers, on television, in school textbooks, and in surveys or, for example, to present school exam results in a clear manner. We would like to know how we can support students to better understand these graphs. In addition, we can then explain to teachers how they can help students and the makers of textbooks (i.e., publishers) to understand graphs even better. In addition, we can explain to journalists, textbook authors, and researchers how they can make even better graphs.

Therefore, it is important that we discover how people read graphs. We do this by following the eye movements of students and experts (teachers). The device with which we do this is called an eye-tracker. This device is placed on the laptop and has a camera and infrared lights. The light from these lamps is invisible and harmless. The device measures exactly what you are looking at without you noticing it. Your face is not filmed. We do make audio recordings of what you say. We also ask the school what the scores are from the CITO examination that the school takes in ninth Grade. We do this to rule out that the results of our investigation are due to something other than what we are investigating.

What are you going to do in the study?

First, we will ask you a questionnaire with some general questions and a questionnaire that measures what you already know about certain subjects. Then, we must calibrate the eye-tracker. This calibration means that you will be asked to look at certain points on the screen. Sometimes it doesn't work immediately; then we just repeat it. It is also possible that it does not work at all.

Calibration is often more difficult if you are wearing mascara or if you have glasses you cannot remove. Therefore, please come without mascara. If you are wearing glasses, you can usually do this, but occasionally it will not work. Contact lenses are usually no problem.

After the calibration, you will be asked 25 questions on the computer, each of which contains graphics. You determine the speed at which you answer the questions. You must answer all questions out loud. Afterward, we will ask you about the tasks you have had.

At the start of the session, the participant handed over the informed consent. The researcher stressed that participation was voluntary and that the participant could stop at any time without any consequences (as was also written in the informed consent). Then, the participant would fill in both questionnaires. Next, the participant was taken to the chair behind the eye-tracker. The procedure of calibration, validation (looking at certain points on the screen), answering questions, and looking at fixation crosses between items was explained. The participant was asked to say the answer out loud for single graph items. For double graph items, the participant was asked to click the correct answer and was given the option to say the answer as well. After the 25 trials, the video data of the participant were shown to the participant. The participant was asked how s/he solved the item of which the gaze data were shown. Only clarifying questions were asked. The participant was—if necessary—prompted to look at her/his own gazes when explaining how the item was solved. Due to technical problems, looking at one's own gazes was not available for a few items (double histograms) for the first participant, who then clarified the strategy with only the item on the screen.

A.7   Lessons learned on doing eye-tracking research with our items

Although eye-tracking is not new, using gaze data is still rare in mathematics education research (Lilienthal & Schindler, 2019; Strohmaier et al., 2020). Therefore, we provide a list of lessons we learned when doing this study for those who are novices in eye-tracking research. This list is unordered and not complete. We also provide two lessons for those wanting to execute a similar study. Note that Holmqvist et al. (2011) also provide a list with advice.

Lessons learned about eye-tracking:

- Consider whether eye-tracking is really necessary. It is time-consuming to learn how to use the eye-tracker, to design items suitable for such research, and it is not always successful.
- Find a group of people with experience using eye-trackers. The first author is very grateful that the eye-tracking seminar of the Faculty of Social Sciences provided the opportunity for researchers from there and other faculties to join and learn from senior researchers in this field.
- Use a chin rest if you use a remote eye-tracker whenever possible. It does not make much difference for the participant, but it makes the accuracy of the trials much better.
- Check if there is both a calibration procedure (to make sure that the eye-tracker knows where a participant is looking) and a validation procedure (to verify how successful this calibration was). Make sure that the calibration and validation backgrounds are the same as the stimulus (items). Some advice that came too late for us was to take a screenshot of the calibration screen results if the software does not provide it.
- Do a pilot with only a few participants. During our pilot, we discovered that: (1) we made an impractical design for the eye-tracker software we used, which made it impossible to know where a new trial started. This happened because we wanted to use graphs in combination with multiple-choice answers (or open answers). See the data repository for our final setup; (2) therefore, we could also not use AOIs from the software during the pilot study; (3) using multiple-choice answers for estimating the mean items guided students too much toward the correct answer; (4) students were much faster in answering the items than expected. This opened the possibility of a larger set of items.
- Multiple-choice items combined with graphics are not possible in the Tobii software if you also want the answers automatically scored. It would be possible to use a separate screen for clicking the answer options. We did not use this option because of the risk of influencing students' gazes on the items.
- During the eye-tracking experiment, note down relevant information, such as the geometry of the set-up (take a picture, measure distances such as between participant and screen), whether a participant has mascara, contact lenses, glasses, epicanthic eye folds, drooping eyelids

(right/left/both). Keep distances the same for each participant as much as possible. For example, we used a desk chair that we could raise depending on the participants' heights so that the chin rest was always in the same position relative to the screen.

- Cover windows if curtains let through too much light (some people have used dark plastic waste bags) and other sources of light that might create shadows in the room. Avoid light sources directly on/above the screen and eye-tracker. Place the eye-tracker (and screen or tablet) in between fluorescent tubes. More advice on factors influencing gaze data quality can be found in the work of Holmqvist et al. (2012).
- Anonymize/pseudonymize participants from the start (e.g., we used L01 for the first participant on the questionnaires, as the name for the trial and so on). Otherwise, you have to change that afterward (before or after exporting) your data export from Tobii to guarantee anonymity.
- Test the items on paper before you use them in the eye-tracking software.
- Have someone check for typos in your items as well as small deviations. For example, it was hard to get the axis lengths of case-value plots and histograms the same. Furthermore, in the graph of Item03 a small mistake was made. As this was discovered only after several trials and did not influence the correct answer, we did not correct this.
- Make sure that the items only differ where relevant and keep the rest as constant as possible (colors, background, letter size, position, and so on). See the work of Orquin and Holmqvist (2017) for more advice.
- Have some small talk with the participant before you start. Have the participant do something (e.g., fill in a questionnaire) before you put them behind the eye-tracker. This gives them time to become at ease with you, the setting, and so on.
- Make a plan for what you say when, what to do in certain situations, and where the equipment is placed. Use tape to mark places for equipment if you must remove it between participants.
- Try to avoid gazes hovering over a relevant area (e.g., the graph area) when going from the question to answer options. In our double graph items, we avoided this by giving the answer options directly

underneath the question and above the graphs. In the single graph items, we used open answers.

- Think carefully about what measures are relevant to your stimulus and research question. See the section on which metrics to use for gaze data. Apart from the metrics mentioned there and in the Data analysis section, also consider machine learning (with raw gazes, fixations on AOIs, vectors for each saccade, and so on) or latent profile analysis (see Hickendorff et al., 2018 for an introduction). Latent profile analysis could, for example, be used to cluster AOI-based gaze data into groups.
- Take a screenshot of the settings of your eye-tracker.
- Before you calculate any metric, make sure that you have seen enough gaze data (trials) to get a sense of what is going on. Take the time to make yourself acquainted with these data. The advice the first author received from colleagues was to watch the gazes, and when you are done, watch even more.
- Besides analyzing the video of the gazes (sometimes also called dynamic gaze-plots) or AOIs, you can also think of analyzing static gaze-plots or heatmaps. Make sure to save them all for a data paper or later analysis if necessary.
- If you make html files for your stimuli, use fixed places (instead of dynamic) on the webpages as much as possible so that the webpages can easily be re-used on another device using the same size and position on the screen.
- For machine learning purposes and power (for hypothesis testing), it may be necessary to collect data from many students and on many items; the Tobii Studio software may become very slow with large amounts of data. This was a bit annoying during the recall, but most time was lost during data export (one export file with all data was not possible due to its size) and visualization creation (heatmaps, gaze-plots). This is a combination of the software and the processor of the laptop used.

Lessons learned about the items:
- The single graph case-value plot items are, in a way, self-correcting, as it is impossible to use a histogram strategy on these. This was not the case for the double case-value plot items.
- In a next experiment, we would consider positioning the double graph items on a diagonal on the screen. This will make it easier to

distinguish the typical (horizontal or vertical) gazes within the graph area from gazes that go between two graphs. As this positioning has some other disadvantages (e.g., you need a bigger screen for the same size of the stimulus and the comparison of the graphs by the students might be hindered by this positioning), another possibility might be to place the graphs underneath each other in some items and next to each other in others (although that might require a bigger screen).

4

Automated gaze-based identification of students' strategies in histogram tasks through an interpretable mathematical model and a machine learning algorithm

"If you haven't got it, you can't show it. If you have got it, you can't hide it." [38]
Zora Neale Hurston

This chapter is based on
Boels, L., Garcia Moreno-Esteva, E., Bakker, A., & Drijvers, P. (accepted). Automated gaze-based identification of students' strategies in histogram tasks through an interpretable mathematical model and a machine learning algorithm. *International Journal of Artificial Intelligence in Education*.

---

[38] Hurston, Z. N. (1942). *Dust tracks on a road,* p. 143.
https://en.wikiquote.org/wiki/Zora_Neale_Hurston

Abstract As a first step toward automatic feedback based on students' strategies for solving histograms tasks we investigated how strategy recognition can be automated based on students' gazes. In a previous study, we showed how students' *task-specific* strategies can be inferred from their gazes. The research question addressed in the present chapter is how data science tools (interpretable mathematical models and machine learning analyses) can be used to automatically identify students' task-specific strategies from students' gazes on single histograms. We report on a study of cognitive behavior that uses data science methods to analyze its data. The study consisted of three phases: (1) using a supervised machine learning algorithm (MLA) that provided a baseline for the next step, (2) designing an interpretable mathematical model (IMM), and (3) comparing the results. For the first phase, we used random forest as a classification method implemented in a software package (Wolfram Research Mathematica 'Classify Function') that automates many aspects of the data handling, including creating features and initially choosing the MLA for this classification. The results of the random forests (1) provided a baseline to which we compared the results of our IMM (2). The previous study revealed that students' horizontal or vertical gaze pattern on the graph area were indicative of most students' strategies on single histograms and the IMM captures these in a model. The MLA (1) performed well but is a black box. The IMM (2) is transparent, performed well, and is theoretically meaningful. The comparison (3) showed that the MLA and IMM identified the same task-solving strategies. The results allow for the future design of teacher dashboards that report which students use what strategy, or for immediate, personalized feedback during online learning, homework, or massive open online courses (MOOCs) through measuring eye movements, for example, with a webcam.

Keywords Eye-tracking; Computer in education; Histograms; Mathematica Classify Function; Problem-solving strategy; Graphs.

4.1 The challenge of gaze-based strategy identification in statistics education

Imagine students learning statistics on a laptop. They interpret histograms to solve a task. Assume that the webcam camera is good enough to track their eye movements when thinking about the task. It is imaginable that the eye movements of all students can be automatically recorded online[39] in the near future to document students' strategies. With available techniques, it is in principle possible to give automated feedback on students' task-specific solution strategies. In this imaginary situation, feedback on students' strategies can even be given before students answer. We see this chapter as a first step toward using gaze data as a learning analytics source in, for example, an intelligent tutoring system (ITS) in statistics education.

Although techniques are available, there are still several challenges regarding the use of gaze data as a learning analytics source in statistics education, before an ITS can be considered. The first is the availability of gaze data, as the use of eye-tracking in statistics education is rare (e.g., Strohmaier et al., 2020). Recent reviews of eye-tracking studies in mathematics education found only four studies in statistics education (one out of 33 included studies, Lilienthal & Schindler, 2019; three out of 161, Strohmaier et al., 2020).

The second challenge is that current usage of gaze data often addresses general pedagogical themes (e.g., metacognitive skills; Lai et al., 2013) instead of task-specific strategies teachers in statistics education are interested in (sometimes called didactics or domain-specific pedagogy). Most studies investigating students' strategies look at general strategies including planning and evaluation (e.g., Eivazi & Bednarik, 2010) or global scanning followed by local viewing (Van der Gijp et al., 2017). Other studies look at cognitive models such as visual working memory (e.g., Epelboim & Suppes, 2001). The number of studies that uncover task-specific strategies in mathematics in primary and secondary education (e.g., Lilienthal & Schindler, 2019; Strohmaier et al., 2020) and science (e.g., Garcia Moreno-Esteva et al., 2020; Klein et al., 2021; Kragten et al., 2015) is still relatively small but growing. For example, patterns in students' gazes indicating strategies have already been found in mathematical domains such as numbers (Schindler et al., 2021), arithmetic (Green et al., 2007), fractions (Obersteiner & Tumpek, 2016), proportional reasoning (Shayan et al., 2017), area and perimeter (Shvarts, 2017), Cartesian coordinates (Chumachemko et al., 2014), geometry (Schindler

---

[39] For most people, online means on a website. In eye-tracking research, however, online often refers to: real-time, hence, during the task solving process of the student. Here we refer to both meanings of online.

& Lilienthal, 2019), trigonometry (Alberto et al., 2019) and functions (e.g., parabola; Shvarts & Abrahamson, 2019). For mathematics and statistics teachers, such strategies are important as they can reveal students' knowledge of and deficiencies in this specific topic (cf. Gal, 1995).

Third, a challenge is that automation of strategy identification by using interpretable models or machine learning techniques in combination with gaze data, is even rarer in statistics education: Only one of the four studies in the previously mentioned review studies used a machine learning approach (Garcia Moreno-Esteva et al., 2016). None of these studies used an interpretable mathematical model. Our present study aims to address this third challenge by investigating how these two data science tools—an interpretable mathematical model and machine learning algorithms—can be used to automatically identify students' strategies on histograms based on gaze data.

Fourth, although the use of gaze data in ITSs is not new, the majority of studies on ITSs that use gaze data seem to focus on general skills such as engagement (e.g., D'Mello et al., 2012). This is in line with a review of research articles on artificial intelligence in education (AIED) in which an independent cluster of recent eye-tracking articles emerged that "include 'collaborative learning', 'engagement', 'video-based learning' and 'recommender system'" (Feng & Law, 2021, p. 293).

Fifth, many ITSs in mathematics and statistics education seem to focus on *procedural* knowledge—problems that can be solved by following a stepwise solving procedure such as solving a linear equation—although ITSs that focus on students' task-specific *strategies* do exist, also in statistics education (e.g., Tacoma et al., 2019). To the best of our knowledge, none of these seem to use gaze data as a learning analytic source.

That said, ITSs that use gaze data for identifying visual-based task-specific strategies, as far as we are aware, do not exist yet in statistics education. Before such a gaze-based ITS can be considered and developed, we not only need to be able to link students' mathematical task-solving strategies to specific gaze patterns but also to automate the identification (or classification, as data scientists would say) of such strategies. In our previous, qualitative study, we inferred students' strategies from their gaze data. In the current study, we concentrate on automatization through the research question: *How can gaze data be used to automatically identify students' task-specific strategies on single histograms?*

The potential of automated identification of such strategies is to make large-scale, personalized feedback possible for online learning both in the initial stages of learning and during expertise development (Ashraf et al., 2018; Brunyé et al., 2019; Jarodzka, et al., 2017; Hwang & Tu, 2021). This can make

feedback in online courses or during homework more efficient and more accurate.

The aim of our chapter is to show how the *identification* of students' task-specific strategies on histograms can be automated. We expect that this work can nurture the dialogue between experts in the field of data science algorithms—more specifically experts regarding interpretable mathematical models (IMM) and machine learning algorithms (MLAs)—and educational researchers. IMM and MLA experts may be more interested in how the IMM or MLA was or could be tailored to the specific application. Educational researchers may be more interested in using an MLA as it is, as a black box, and wonder what it provides them and how well it works. The advantage of an IMM for educational researchers, is that it is transparent in how exactly it came to its decisions for individuals. We think this chapter can fuel the dialogue between IMM and MLA experts and educational researchers to keep the boundaries between disciplines permeable. At such boundaries, exciting new research can emerge.

In this chapter, we developed an interpretable mathematical model (IMM) and compared its results with a machine learning algorithm (MLA). We used these two methods from data science along with theories and insights from psychology research and neurosciences (e.g., on eye-tracking, what gaze data can and cannot tell us and the sensorimotor system), theories and insights from mathematics and statistics education research (e.g., on averages and histograms) and bring this to the world of human-computer interaction (in which, for example, the usability of an IMM or MLA is important). This means that we sometimes need to bridge worlds in terms of terminology, expectations, and explanations.

Our study is in line with the call for research focusing on methods for using measures of micro-level learning processes—including gaze data (Harteis et al., 2018). For the specific topic of histograms, our study also provides the level of detail that Peebles and Cheng (2001) referred to: "From [...] eye-movement studies it is argued that there is a missing level of detail in current task analytic models of graph-based reasoning." (p. 1069). Yuan et al. (2019) showed that there is a need for searching for "visual cues that mediate the patterns that we can see in data, across visualization types and tasks" (p. 1).

4.2 Theoretical background of tasks and identification methods

4.2.1 Estimating the arithmetic mean from histograms

Developing students' statistical literacy, reasoning, and thinking is an important goal of education (Ben-Zvi et al., 2017). Statistical literacy is especially important in the world of "big data" and alternative truths (Burrill, 2020). Most adults will be data consumers, making decisions based on data collected by others (Gal, 2002). Statistical data in tables are not always clear. Messages can be clearer if these data are presented in more aggregated forms in graphical representations—including dotplots, boxplots, and histograms— that stress some aspects of the data (e.g., variability) and leave out other information (e.g., the exact measurements). Students, however, find it difficult to correctly interpret histograms.

A review of students misinterpreting histograms revealed that many of their difficulties stem from not understanding the statistical key concept of data (see Chapter 2). The key concept of data includes an understanding of what, how many, and how variables and their values are depicted in a histogram. Despite many carefully designed interventions to tackle misinterpretations (e.g., Kaplan et al., 2014), students' difficulties with histograms remain (e.g., Cooper, 2018). We, therefore, decided to use eye-tracking to study in depth how students interpret histograms (see Chapter 3; Boels et al., 2018, 2019a, 2023).

Strengths and caveats in students' knowledge can be revealed by asking them to estimate averages from data in different representations (e.g., histogram, dotplot, case-value plot; cf. Gal, 1995). Estimating the mean can be seen as a prerequisite for assessing variability, as the variation in data is compared to a measure of center (e.g., standard deviation from the mean). Furthermore, our students are familiar with the mean, but not so much with variability. Therefore, in a previous eye-tracking study, students were asked to estimate the mean from various—but univariate—statistical graphs in 25 items (e.g., see Chapter 3). In the present chapter, we re-use gaze data from a subset of this previous study containing all five single histogram items.

Historical examples show that the mean has emerged from estimating representative values for a dataset through compensation and balance (Bakker & Gravemeijer, 2006). Students exhibit minimal difficulty in estimating the mean from case-value plots (Cai et al., 1999), unless zero is one of the measured values (see Chapter 3). Most students know how to calculate the arithmetic mean from raw data (e.g., Konold & Pollatsek, 2004). In a study with various items—including finding the "average" allowance from a histogram—

five approaches were found for solving the items: average as (1) mode, (2) algorithm, (3) reasonable, (4) midpoint, and (5) balancing point (Mokros & Russell, 1995). Students often (implicitly) use the mean of frequencies in a histogram (cf. Cooper, 2018). The latter is incorrect when applied to histograms but correct for finding the mean from a case-value plot, and can be seen as finding the horizontal line that makes all bars of equal height by using compensation.

The weighted estimation of the mean in a histogram is the balance or gravity point of the graph (e.g., Mokros & Russell, 1995). This mean can be found by taking the range or spread of the data in the histogram into account together with the height of the bars. For this approach, it is not necessary to read off frequencies on the vertical axis. We call this approach a histogram (interpretation) strategy or correct strategy. An estimation of the mean in a histogram *with equal bin widths* can also be computed by multiplying the frequency or percentage (height of the bar) with the middle value of that bar, adding the results over all bars, and dividing this by the *sum* of the frequencies. No students in the previous study used this approach. Instead, all that used a computational approach added all frequencies and divided this sum by the number of bars. This would be a correct strategy if the height of each bar was representing weight and the number of bars was the number of measured weights (as in a case-value plot). Therefore, this count-and-compute strategy is incorrect for histograms.

In our previous study, we found several strategies for estimating the mean from a histogram based on students' visual search strategies (cf. Goldberg & Helfman, 2011) inferred from their gaze patterns (see the Empirical background of the re-used data section). A visual search strategy can be part of a task-specific strategy. People use these strategies to get "from an initial problem state to a desired goal state, without knowing exactly what actions are required to get there (Newell & Simon, 1972)" (Van Gog et al., 2005, p. 237). As the debate between Lawson (1990) and Sweller (1990) illustrates, there are different opinions on what strategies are. In our study on graph interpretation, students' strategies typically consist of (1) visually searching for the relevant information, (2) making inferences based on this information, and in some cases (3) verifying the inference; see also the section Theoretical interpretation of students' gaze patterns. Given our focus on what eye-tracking and data science tools (IMM, MLA) can provide to educational researchers, we now first discuss the theoretical background of using eye-tracking. Next, we discuss the background of our methodological choices including eye-tracking, IMM and MLA, and provide a short introduction to supervised MLAs. How we tailored these methods for our purpose, is discussed in the research approach section.

4.2.2 Use of gaze data

There are multiple reasons for using gaze data to identify students' strategies. First, eye-movement patterns (e.g., order of fixations[40] or saccades) are online, real-time measures that may allow for more adequate feedback than feedback on answers only. Moreover, feedback on strategies can be provided earlier during the task-solving process (e.g., Gerard et al., 2015; Mitev et al., 2018), although strategy feedback can also be on answers (e.g., Tacoma et al., 2019). Second, low-accuracy eye-tracking—for example, through webcams—is expected to be a standard option for computers in several years' time (e.g., Kok & Knoop-Van Campen, 2022), which would make it possible to give feedback to large groups of students. Third, gaze data are direct motor data that are almost impossible to manipulate. This makes measuring eye movements more reliable than, for example, thinking-aloud protocols (e.g., Van Gog et al., 2005). In addition, younger students, novices, and sometimes even experts find it difficult to articulate their thinking process, are sometimes not aware of their thinking (e.g., Green et al., 2007) or might respond to what they think the interviewer expects or what is easily accessible (e.g., Wilson, 1994).

The implicit assumption here is that eye movements reflect cognitive processes. Spivey and Dale (2011) state: "Our most frequent motor movements—eye movements—are sure to play an important role in our cognitive processes. [...they] provide the experimenter with a special window into these cognitive processes." (p. 551). It is indeed generally assumed that gaze data can provide evidence of conceptual actions, however with some caveats (e.g., Radford, 2010). First, the relationship between eye movements and cognitive processes is not straightforward (e.g., Kok & Jarodzka, 2017; Russo, 2010). In addition, not every eye movement is part of a student's strategy (e.g., Anderson et al., 2004; Schindler & Lilienthal, 2019). Furthermore, one could argue that students' fixations on the screen do not indicate where they looked, as people also observe through their peripheral vision (Lai et al., 2013). Nevertheless, in our items, focused vision is needed for locating detailed information (e.g., locating a bar, reading a specific number on the horizontal axis). As the fovea has the greatest acuity (sharpness; Wade & Tatler, 2011), locating a number on an axis is only possible with foveal vision, and peripheral vision most likely guides our attention to it and to the bars (cf.

---

[40] "A fixation is a period of time during which a specific part of [we use *graph on the computer screen* here] is looked at and thereby projected to a relatively constant location on the retina. This is operationalized as a relatively still gaze position in the eye-tracker signal implemented using the [Tobii] algorithm." (Hessels et al., 2018, p. 22). A period of time during which a relatively fast switch of gaze between two fixations occurs is called a saccade.

Kok & Jarodzka, 2017). Therefore, we can infer that the fixation on the screen is what the student is looking at.

Choosing what eye-movement measures to use is a methodological decision. According to a review study (Lai et al., 2013), the measures used most often were temporal (e.g., total fixation duration, time to first fixation, total reading time), followed by count (e.g., fixation count). The least used measures were spatial (e.g., scanpath, fixation position, order of AOIs). Goldberg and Helfman (2010) stated that "with appropriate task design and targeted analysis metrics, eye-tracking techniques can illuminate visual scanning patterns hidden by more traditional time and accuracy results" (p. 71). Scanpaths can reveal learning in more detail (Hyönä, 2010). Tai et al. (2006), therefore, advise using spatial measures such as scanpaths in problem-solving research.

Studies using students' scanpaths for identifying strategies are rare, and often use the *sequence* of AOIs (e.g., Garcia Moreno-Esteva et al., 2018) or scanpaths that are aggregated over time and fixations (e.g., in heatmaps, Schindler et al., 2021). Up to now, scanpaths mostly require qualitative inspection or analysis of the eye-movement data (e.g., Alemdag & Cagiltay, 2018; Susac et al., 2014)—especially when looking for task-specific strategies. Figure 4.1 provides an example of such a scanpath (a sequence of fixations and saccades). Qualitative analysis is both time-consuming and harder to objectify. In this chapter, we, therefore, use the raw scanpath data to identify students' strategies.

In studies using angles and direction of saccades in educational settings (e.g., Dewhurst et al., 2018), scanpaths are often compared on multiple or all AOIs[41]. In our previous, qualitative study, we took a new approach in using the perceptual form (e.g., vertical gaze pattern) of the gazes on *one* AOI only (graph area)—the one that was found was particularly relevant for students' task-specific strategies (see also the following section, 'inferring an attentional anchor from gaze data'). This perceptual form consists of angles and direction of saccades that are roughly *aligned*. So far, we have not found any other study in education that uses alignment of saccades. For more details, see the Research approach section (students' strategies). A possible advantage of looking at saccades over fixations or order of AOIs is that it may be less sensitive to spatial offsets (e.g., Jarodzka et al., 2010).

---

[41] In addition, this study is about the influence of task difficulty on scanpaths. It does not consider the kind of task-specific strategies we are aiming for.

4.2.3   Inferring an attentional anchor from gaze data

For our theoretical interpretation of the *perceptual form* of students' gaze patterns on the graph area (horizontal or vertical line segments), we draw upon insights from theories on enactivism and embodied cognition. According to these theories, cognition arises from interaction with the environment (e.g., Rowlands, 2010). The focus of an actor's interaction with this environment is called an attentional anchor (AA) (AA; Hutto & Sánchez-García, 2015). An AA is "a real or imagined object, area, or other […] behavior of the perceptual manifold that emerges to facilitate motor-action coordination" (Abrahamson & Sánchez-García, 2016, p. 203). Other behavior of the perceptual manifold, for example, includes students gesturing a horizontal line when explaining how they made all bars equally high in a case-value plot strategy. The AAs found in our previous research (see Chapter 3) facilitated students' imagined actions (strategies for finding the mean)—regardless of the strategies' correctness.

Examples of AAs in motor action can be found in research on high school trigonometry where students coordinate the movement of the left hand to describe a circle and their right hand to describe a sine graph (Alberto et al., 2019). Another example is the manipulation of two bars that are proportional to each other, see Figure 4.1 (e.g., Shayan et al., 2017). Students needed to keep the bars green, which occurred when the bars had a fixed ratio of, for example, 1 : 2 (unknown to the students). They dragged both bars up to find various points where both are green. Students had different strategies for finding these points. In one strategy, gaze fixation is on the right-hand bar in the middle, which is mathematically relevant, as this bar is twice as high as the left-hand bar (Figure 4.1). As this imagined triangle emerges to facilitate the coordination of the motor action, it is an example of an AA.

Figure 4.1 Stable triangular scanpath that was interpreted as an AA
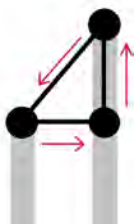
Note. Circles are fixations (places where students looked), arrows indicate the direction of saccades (fast transitions between two fixations), redrawn after Shayan et al. (2017, p. 175).

4.2.4 Theoretical interpretation of students' gaze patterns

Although enactivism often assumes manipulation of an environment, there are indications that people's sensorimotor systems are also activated in situations without physical manipulation (e.g., Fabbri et al., 2016; Lakoff & Núñez, 2000; Molenberghs et al., 2012). In the retrospective stimulated recall interviews, students talked about the graphs as if manipulation were possible. For example, studentL10 refers to chopping and flattening all bars (hence, using compensation, Bakker & Gravemeijer, 2006), see the excerpts below. This student describes sensorimotor actions, namely: breaking up the longest bar into pieces that are then divided over the shorter bars, resulting in a horizontal line along the top of the now equally high bars. This would be a correct strategy for finding the arithmetical mean in a different type of graph (namely, a case-value plot, e.g., Cai et al., 1999; Yuan et al., 2019). An imaginary horizontal line segment is used for coordinating this imagined action. Gaze data show this imaginary segment in the form of a stable scanpath indicating the focus of interaction of this student, see Figure 4.2. We, therefore, interpret this segment as an AA. Gazes on Item06 indicate that this AA was also visible before item20.

StudentL10:	I looked at the graph itself [Item06] first and then I kind of looked at the axes, how is it constructed and then I looked at the question, and then I looked again at the frequency, how to group it. That was it in my opinion.
Researcher1:	And were you doing that here the same way you did with those other [previous] questions? Chop it into pieces?
StudentL10:	Yes.
Researcher1:	You said five here [Item20].
StudentL10:	Yes, because I thought the weight would be on the left side. So, if you flattened it all out, between 4 and 6 would be the imaginary [horizontal] line.

Figure 4.2 Part of the stable scanpath of studentL10 on Item06 (left) and Item20 (right)
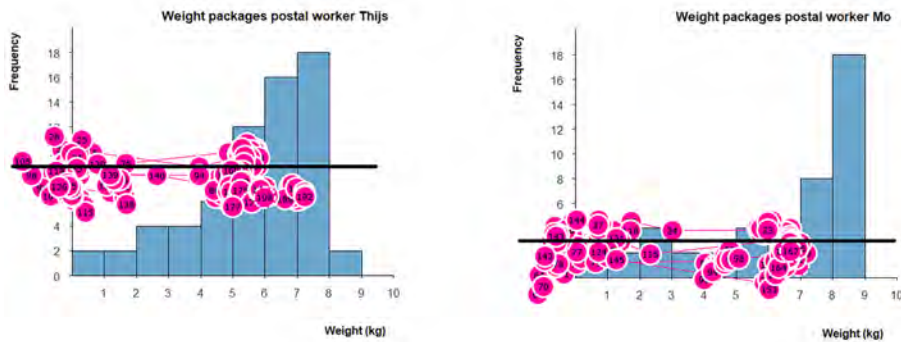
Note. The stable scanpaths reveal the horizontal line segment along which the student looks on Item06 (left; here superimposed on the figure for the reader) and Item20 (right). Circles indicate fixations, thin lines indicate saccades. As the weight value is on the horizontal axis, so is the actual mean. However, from the eye movements of this student, we can conclude that the mean of the frequency is estimated, instead of the mean weight. The interview data support this conclusion. This stable scanpath, therefore, indicates an incorrect (case-value plot interpretation) strategy.

4.2.5   Considerations for strategy identification with machine learning algorithms

For automated strategy identification, an analytic model of this strategy is needed. A machine learning algorithm[42] (MLA) automates building an analytical model. An MLA is a computer program that improves with experience (Kersting, 2018; Mitchell et al., 1990; Molnar, 2019). An MLA is not explicitly programmed to use any particular input features. 'Features' here refers to variables constructed from input data.

An MLA can be supervised or unsupervised. Being supervised means that the training cycle of the program is fed with, for example, the correctness of the strategy; in unsupervised learning, only the gaze data would be given to the program during the training cycle and the program might infer correctness information by itself. As we previously identified two groups in our qualitative study, we wanted to see if the MLA could identify those students correctly. That calls for a supervised MLA. During the training cycle, see Figure 4.3, an MLA is fed with the raw gaze data as well as a classification code for the already identified strategy (0 for incorrect, 1 for correct). After this learning or training cycle, the trained MLA identifies the strategy of *other* students (or trials) that were not part of the training set.

---

[42] In the media, machine learning (ML) and artificial intelligence (AI) are often regarded as synonyms, but there is a difference. Interested readers are referred to Kersting (2018).

Figure 4.3 Training and identification cycles of a supervised MLA
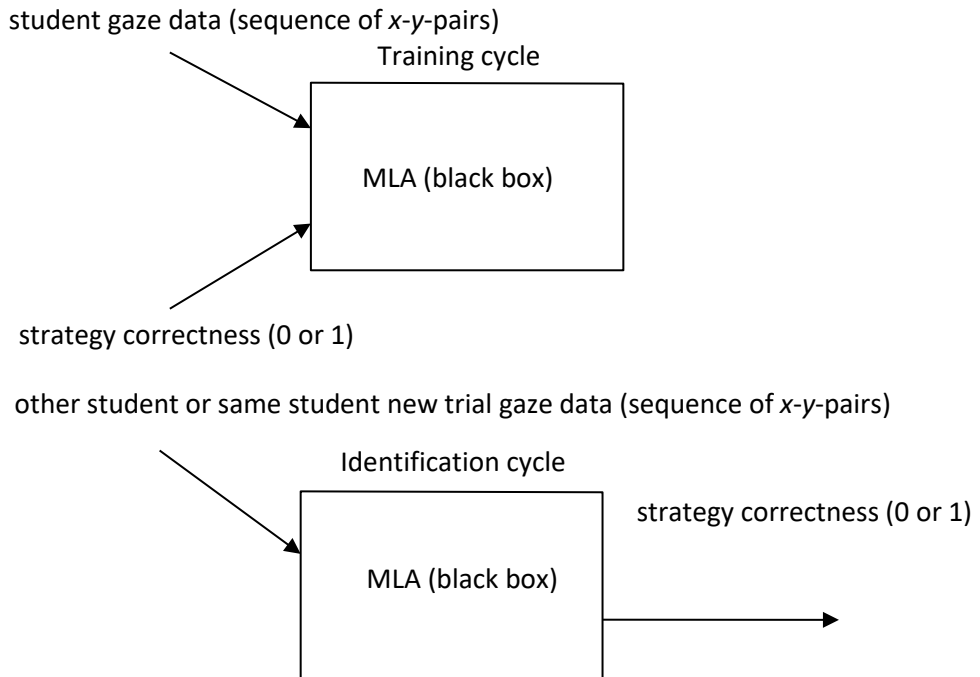
student gaze data (sequence of *x-y*-pairs)

Training cycle

MLA (black box)

strategy correctness (0 or 1)

other student or same student new trial gaze data (sequence of *x-y*-pairs)

Identification cycle

strategy correctness (0 or 1)

MLA (black box)

We used Mathematica's implementation of random forest in the 'Classify Function' (version 13.2.1; WRI, 2020) with default parameters. The Classify Function automates and optimizes the data preparation process. For example, it automatically handles the different lengths of input vectors of *x-y*-pairs (it normalizes input features).

We underline that we used the MLA as a baseline to compare our interpretable mathematical model (IMM) with. We see it as a tool. Educational researchers may be more interested in how well such an MLA performed and what it can provide for them. MLA experts may be more interested in the details of the ML method. It is like users of an electric screwdriver being interested in how well it works and designers of such screwdrivers being interested in the details of how this screwdriver was assembled and could be optimized and whether better ones exist. We would like to emphasize that our article is not a report about research into machine learning methods. It is a report of a study of cognitive behavior which uses machine learning tools to analyze its data. Hence, the purpose of our research was not to conduct an in depth investigation into which machine learning methods would work best for our data but to see how well our IMM performed compared to an MLA.

An important prerequisite for ML analysis is that the dataset is large enough and contains enough information for the MLA to identify (classify) students' strategies. For this purpose, we decided to do a 'sanity check' and see if the MLA would be able to identify students' answer correctness. As 50% of the students answered Item06 correctly (Table 4.1), an MLA with an accuracy of about 50% would not be better than throwing a coin. Fifty percent accuracy could also be reached by identifying all students having a correct strategy. Therefore, for the prerequisite of enough information to be met, the accuracy of this answer identification needs to be well above 50% when using balanced data as in Item06. We decided to set it to 70% or above, as this is a low-stake classification problem. Such a sanity check can be used as a first step before proceeding to the manual determination of students' strategies through qualitative research, or when qualitative research is still in progress. To judge the performance of an MLA, other metrics also need to be considered; see the Methodological evaluation criteria section. Therefore, we first trained the MLA on students' answers (and then identified other students' answers) before we retrained the MLA on students' strategies (and then identified other students' strategies).

4.2.6 Considerations for analyzing eye movements with an MLA

The use of an MLA for analyzing gaze data is still unusual (e.g., Kang et al., 2020) and even more so for educational use (e.g., Brunyé et al., 2019; Mitev et al., 2018). For example, in a review of eye tracking in medical education, the use of an MLA is described in only two out of 33 studies (Ashraf et al., 2018). In many studies, areas of interest (AOIs) are used. AOIs are predefined areas of the item that are judged by the researchers as being distinct from each other and relevant to the strategy. Eye movements on these AOIs or between them are recorded. Typically, the information is often reduced to a single measure for each AOI—for example, whether an AOI was visited or not. In our study, we used raw gaze data on one large AOI to study the perceptual form of the scanpath. Examples of AOI usage in ML studies are the order in which AOIs are visited, or the number of fixations on the areas of interest (e.g., Garcia Moreno-Esteva et al., 2020; Najar et al., 2014). In other studies, temporal measures are used, such as the total duration for fixations on an AOI or mean duration per fixation (e.g., Voisin et al., 2013). Schindler et al. (2021) used heatmaps of students' gazes in an MLA, thus discarding the order of fixations in the scanpath pattern. However, in most eye-tracking studies, no MLA is used at all (e.g., Strohmaier et al., 2020; Van Gog & Jarodzka, 2013). In some studies, an interpretable (vector) model was made from the raw data (e.g., Dewhurst et al., 2012). The use of multimodal data—including eye-movement data—in combination with an MLA is an emerging line in educational research (Järvelä

et al., 2019). What is new in our study, to the best of our knowledge, is that we feed the MLA (and IMM) with the raw gaze data to identify students' task-specific strategies.

An advantage of supervised MLAs for analyzing gaze data is that they provide a rather generic approach. However, the disadvantage of many MLAs is that they are effectively like black boxes that do not reveal how they identify the results from the data, and, hence, what analytical model emerges (e.g., Guidotti et al., 2018; Kuhn & Johnson, 2013; Lakkaraju et al., 2019; Rudin, 2019). Therefore, it is unknown what gaze data patterns are used by the MLA for data classification.

Several solutions are suggested in the literature to overcome this disadvantage. More transparency can be created by (1) model or global explanations, (2) outcome or local explanations, or (3) model inspection or differential explanations (e.g., Guidotti et al., 2018; Lakkaraju et al., 2019). Explainable means that humans can understand how the MLA reached its decision (e.g., Doshi-Velez & Kim, 2017). An alternative is to create an (4) interpretable model directly from the data, sometimes after first using MLAs to understand what is relevant in the data (e.g., Rudin, 2019), or (5) use white-box techniques such as models that are made *a priori.* The disadvantage is that (5) is based on human assumptions, not on data; this risks relevant information in the eye-movement data being overlooked (e.g., Villagrá-Arnedo et al., 2017).

In our case, (1) model explanation could involve trying to extract the general rules that the MLA uses to decide what strategy a student uses (for examples from weather forecasting, see McGovern et al., 2019). Outcome explanations (2) might involve trying to extract why student A is identified as having strategy *z* (for clinical examples, see Krause et al., 2016; for an example with birds, see Rudin, 2019). Model inspection (3) could be understood as finding out how sensitive the model is to variations in the data.

As (2) is even more complex to achieve, we decided to make an interpretable model instead (4). An interpretable model is a model that captures the most important characteristics of the strategy, in a way that can be understood by human beings and that is transparent (Rudin, 2019). An interpretable model is often a mathematical and logical model (e.g., Hancox-Li, 2020; Lakkaraju et al., 2019; Molnar, 2019); in our case, it consisted of a set of rules that approximately describes the stable scanpath of the gazes. We call this our interpretable mathematical model (IMM).

To provide a baseline for this IMM, we compared it with a random forests MLA. We used Mathematica's implementation of random forest in the 'Classify Function' (version 13.2.1; WRI, 2020) with default parameters. The Classify Function automates and optimizes the data preparation process. For

example, it automatically handles the different lengths of input vectors of *x-y*-pairs (it normalizes input features).

The Classify Function initially choose random forest as the best MLA for each of all five items in a *previous* version of the software that we started our analyses with. In the newest version (13.2.1) another MLA (logistic regression) provided slightly better results for several of our items and random forest for others. We report the results for random forests for several reasons, including consistency, and that the MLA is only a baseline for our IMM. For all further analysis we seeded these random forests to make sure that our results are reproducible. An advantage of the Classify Function in the Mathematica software is that users do not need to deal with the details of machine learning methods. The Classify Function is used "as is" and automates many aspects of the methodological stack. Specifically, to give an example, it automates the selection of the machine learning method and the preparation of data (consisting of only the—temporally ordered—*x*- and *y*-coordinates of fixations on the AOI graph area) so the selected method can be applied to it. Note that timestamps are not provided to the MLA, so the MLA does not 'know' that the data is *temporally* ordered or that data from other AOIs are removed. The downside of our approach is that we know little about how the Classify Function is handling our data. For example, data preparation and feature selection is all hidden in the software and we, therefore, consider it a black box, even though the MLA it initially selected —random forest—is itself known for its possibilities for feature extraction (e.g., through a feature importance plot). Our data, which are continuous, are by themselves difficult to interpret (*x*- and *y*-coordinates). Ultimately, the problem is that it is impossible to know whether the classification is based on gazes that are typical for the task-specific strategy at hand (cf. Kuhn & Johnson, 2013). In addition, the first step of (3) was performed by using the random forest MLA trained on one item for classifying strategies on other items and vice versa for all item pairs.

4.2.7 Considerations for the construction of an interpretable mathematical model

An (IMM) is transparent in what and how characteristics of students' gaze patterns are used to identify students' strategies. To detect task-specific strategies and ensure that the IMM is usable—an evaluation criterion, see the next section—we use the idea of an attentional anchor (AA) as described earlier to search for a task-specific perceptual form of the gazes (stable scanpath; see the section on Theoretical interpretation of students' gaze patterns). The advantage of using an AA for constructing an IMM is that it is both task-specific (for each topic and task a different perceptual form is

expected) and generalizable (it has been found already for topics and tasks in various mathematical domains).

4.2.8   Methodological evaluation criteria

Validity, reliability, and causality are important methodological evaluation criteria in educational research. Different terminology and definitions are used in data science and human-computer interaction research for evaluating MLA results. In this section, we compare these terminologies.

First, in educational research, validity "concerns whether we really measure what we intend to measure" (Bakker & Van Eerde, 2015, p. 443). In educational research, threats to the validity are internal and external, for example, maturation of subjects between measurements, subject selection effects on results, loss of subjects, changes in instrumentation, and so on (Eisenhart & Howe, 1992). This mostly applies to how the data were collected and what can be inferred from them. Regarding the data collection, as all data were collected in one session, maturation does not apply. Also, we did not exclude any subjects from the dataset, hence, there is no loss of subjects. We did not change our instrumentation, and so on. What does apply is that we selected subjects from pre-university track students in Grades 10–12 only, from one school, and only those who volunteered (which is inevitable). However, we observed the same phenomena (strategies) in different subjects in previous studies with secondary school teachers (Boels et al., 2019b) and with university students (Boels et al., 2018). Moreover, qualitative research does not seek to generalize from sample to population but from variation in the data to the phenomenon (Levitt, 2021). Regarding what can be inferred from the data: we combined the gaze data with the results of cued recall. Cued recall means that students were shown their gazes (the cue) and asked to explain what strategy they used (cf. Van Gog et al., 2005). This ensures that the data collection was valid. For the MLA and IMM, validity can be understood as whether these actually measure the phenomenon (strategies). As we discuss in the Research approach section, this is true for the IMM by its design, but we cannot be sure about the MLA. However, the results of the IMM suggest that the phenomenon can be measured from the gaze data.

Second, reliability in educational research is about "independence of the researcher[s' judgment]" (Bakker & Van Eerde, 2015, p. 443) or small variation in outcomes. Reliability of a method entails that its results can be reproduced with the same population with comparable items (e.g., Golafshani, 2003). Reliability, in this sense, can be understood as the MLA results having about the same and sufficient accuracy as the results of the qualitatively identified strategies and of those of an IMM. Another way to look at the reliability of MLAs in this sense is by comparing results on different items (e.g.,

train the MLA with data of one item and identify students' strategies for another). When there is sufficient overlap, all are considered to identify the same phenomenon. This, in turn, makes the MLA results more reliable. Reliability here refers to what data scientists sometimes call the performance of the MLA. In human-computer interaction research, reliability (also) involves the safety of the system, downtime, and consistency in the results (e.g., Bosnić & Kononenko, 2009; Webb et al., 2020), which is not relevant to us as we are not building an application.

Data scientists, however, define "reliability of classification as an estimated probability that the (single) classification is in fact the correct one" (Kukar & Kononenko, 2002, p. 219). To avoid confusion, we will, therefore, use *performance* when evaluating our results. Performance is also checked through cross-validation (e.g., Berrar, 2019) which involves applying the trained MLA to unseen data. We used several cross-validation procedures. First, we used a procedure often applied in statistical research—jackknife— which is a form of resampling (e.g., Efron & Stein, 1981) which means that the answers or strategies for all 50 students are identified in an iterative process by the MLA based on learning from the other 49 students. Since there are 50 students who can be left out one at a time, there are 50 ways to do this and the 50 results are averaged. Second, we performed a leave-one-out cross-validation (LOOCV) which means that data from 49 students are used as training data, and the strategy of the 50th student is classified. This is done 50 times until all students' data are used as test data once. Furthermore, we used a stratified 5-fold cross-validation which means that the data are split into groups of ten students. Stratified means that in each group of ten students, the number of students with a correct strategy is roughly the same. Then, the MLA is trained with 40 students and classifies the strategies of the remaining 10. This is repeated five times until data from all groups are used once as test data.

An MLA's performance can be measured in different ways; through accuracy, through a confusion matrix (e.g., comparing the results of the MLA with the results of the qualitative coding), and through a ROC[43] plot (Fawcett, 2006; see the Results of applying an MLA and IMM section) that gives an idea of the true positives and false negatives rates. Accuracy is expressed as a percentage of correctly predicted or identified cases (e.g., Afonja, 2017). As explained earlier, for our supervised MLA in this low-stake situation, we regard an accuracy of 70% or higher as good. In addition, we consider 80% or above as

---

[43] ROC stands for Receiver Operating Characteristic. Originally it was used to judge how well a specific Receiver Operated meaning how well it was picking up enemy signals (radar). In our case, the plot visualizes how well the signal (in our case: true positives) is detected compared to false negatives.

very good, and 90% or more as excellent. However, accuracy can be misleading. For example, if only ten percent of the students used the correct strategy, and the MLA identifies all strategies as incorrect, the accuracy would be 90%, but this identification would not be valid. Therefore, accuracy should be used with precaution. In a confusion matrix, counts for true and false positives and negatives are reported separately (see the Results section and Appendix A of this chapter). From this matrix, sensitivity and specificity can be calculated using the following formulas (cf. Kuhn & Johnson, 2013), which give a better idea of how the MLA is performing:

$$Sensitivity = \frac{\text{\#samples qualitatively coded correct } and \text{ by MLA identified as correct (strategy)}}{\text{\#samples qualitatively coded having correct strategy}}$$

This formula is often shortened to (e.g., Fawcett, 2006):

$$Sensitivity = \frac{\text{true positives}}{\text{true positives } + \text{ false negatives}} = \frac{\text{true positives}}{\text{total positives in true class}} = \frac{TP}{P}$$

In this second formula it is not immediately clear what is considered to be the 'true' class, whereas in the first it is clear that we took the results of the qualitative study as the 'true' results and the MLA results as the hypothesized class. For example, for Item01 and the IMM, $TP = 14$ and $P = 14 + 10$, (Table 4.6), therefore, $sensitivity = \frac{14}{24} = 0.583$ … which is rounded to 0.58 (Table 4.5, Results of applying an MLA and IMM section).

$$Specificity = \frac{\text{\#samples qualitatively coded incorrect } and \text{ MLA identified as incorrect (strategy)}}{\text{\#samples qualitatively coded having incorrect strategy}}$$

Similarly to the above formula for sensitivity, the formula for specificity is often shortened, to:

$$Specificity = \frac{\text{true negatives}}{\text{true negatives } + \text{ false positives}} = \frac{\text{true negatives}}{\text{total negatives in true class}} = \frac{TN}{N}$$

Third, in educational research, measuring is only valid "if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al., 2004, p. 1061). Data science does not use the word validity. With the MLA, we intend to measure students' strategy correctness based on their gazes. Therefore, variations in students' gazes should produce variations in the classification by the MLA. Furthermore, if an IMM can be understood by human beings—and describes the observed phenomenon accurately—it adds to the validity of both the model and the MLA (e.g., Doshi-Velez & Kim, 2017; Rudin, 2019) as well as

to its usability (e.g., Guidotti et al., 2018). Usability is another criterion from human-computer interaction literature. In the previous section, we described how an IMM that meets these criteria can be constructed.

Fourth, causality implies that a change in the results of the MLA is due to a change in the real system (Doshi-Velez & Kim, 2017). Although not every eye movement is part of the task-solving strategy (e.g., Schindler & Lilienthal, 2019), eye movements and strategies do associate (e.g., Kok & Jarodzka, 2017). Therefore, instead of causality, we use association, meaning that if students perform a specific pattern of gazes, they use a specific strategy. If there is an association between a gaze pattern and a strategy, the accuracy of identifying this strategy by the IMM and the MLA will be sufficient or better.

4.3   Research approach

In a previous study, we collected and qualitatively analyzed students' gaze and stimulated recall data and classified these into groups (Table 4.2) of students using the same strategy (see Chapter 3). The present study consists of three phases: (1) analysis of gaze data on *one* AOI (that contains the stable scanpath) through a random forest MLA; (2) construction of a separate interpretable mathematical model (IMM) based on the gaze data and using insights from qualitative research; and (3) comparison of the results of the MLA with the IMM and with the results of the previous qualitative study as the MLA is a baseline to compare our IMM to. The most important information from the previous study is presented in the following section. Next, the first two phases of the present study are explained in more detail. A comparison of the results is made in the Results section.

4.3.1   Empirical background of the re-used data

Participants
This study re-uses data from an eye-tracking study with 50 Grades 10–12 students of a Dutch public[44] city high school (see Chapter 3). All participants are Dutch pre-university track students (15–19 years old; mean age 16.3; all with normal or correct-to-normal vision; 23 males, 27 females).

---

[44] In the Netherlands, private schools are rare and in general, there is no difference between state schools in terms of students' results.

Eye-tracking apparatus

The gaze data that are used as input for the IMM and the MLA were collected with a Tobii XII-60 eye-tracker with a sampling rate of 60 Hz that was placed on an HP ProBook laptop between the laptop's 13-inch screen and keyboard (Figure 4.4). A chin rest was used to reduce data loss and improve the accuracy of the gaze data. Furthermore, a 9-point calibration on the screen was used. The Tobii Pro Studio 3.4.5 software recorded in real-time where people looked at the screen by using harmless infrared light to detect their gaze. Data loss was minimal (7.2% on average) and none of the students (averaged over all trials[45]) or items (averaged over all students) went over the exclusion point of 34%. The mean accuracy is 56.6 pixels (1.16°) with the highest accuracy on the graph area (mean 13.4 pixels or 0.27°). The average precision (0.58°; RMS-S2S; Holmqvist et al., 2023) is considered good. For other measures of gaze data quality, see Chapter 3. We, therefore, did not exclude any student, although for some specific trials, data loss could come close to or even over this exclusion point (e.g., studentL39 and studentL32 had 27.5% and 46.0% data loss respectively on their trial of Item01). Some data loss is normal, due to blinking, wearing glasses or make-up, epicanthic eyes, or students looking above or below the screen while thinking. Another reason for not excluding students is that this would not be representative of a future gaze-based feedback application where real-time data collection and processing would occur.

Figure 4.4 Set-up of the experiment
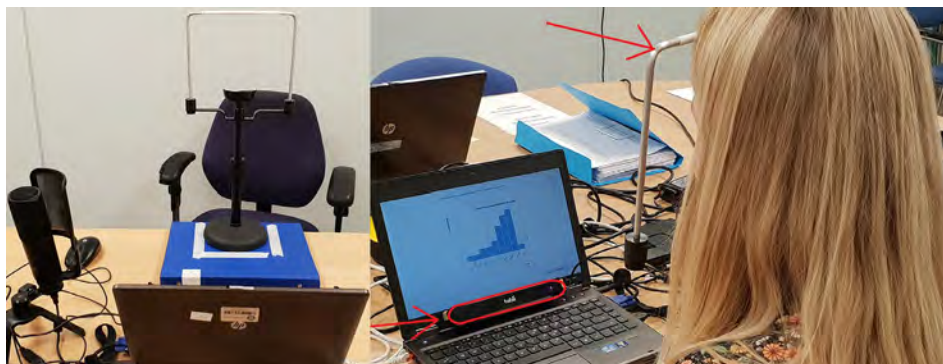
Note. The red arrows in the right-hand picture point at the eye tracker (bottom left, see the red oval) and chin rest apparatus (top middle). The person in the picture is not a participant.

---

[45] In mathematics education, we usually talk about an item, task or problem. In eye-tracking research, a series of gazes of one student solving one such item is called a trial.

Tasks

The data on five histograms items are used and analyzed in the present chapter, see Figure 4.5. The question for all five items was: "What is approximately the mean weight of the packages [name of postal worker] delivers?"

The students verbally estimated the mean (Table 4.1); their answer was coded as correct or incorrect. Answer correctness can differ from strategy correctness, due to, for example, underestimation of the mean, even if a correct strategy was used to locate the mean.

Students' strategies

Qualitative data were collected through expert judgment on students' strategies on the items (Table 4.2), which in turn was based on (1) videos of students' gaze data on the items; (2) interview data when available; (3) students' answers. Three common strategies were identified (Table 4.2): a histogram strategy (Figure 4.6—a correct strategy that reads off the estimation on the horizontal weight axis), a case-value plot strategy (Figure 4.2—a strategy that would be correct for a case-value plot but is incorrect for finding the mean from a histogram as it returns the mean frequency, read on the vertical frequency axis), and a count-and-compute strategy (an incorrect strategy[46] that, for example, adds the height of the bars, hence the frequencies, and divides by the number of bars—resulting in a kind of zig-zag pattern of horizontal and vertical gazes, see Chapter 3 for more details). Both the case-value plot strategy and count-and-compute strategy relate to the same misinterpretation: interpreting the histogram as a case-value plot (Boels et al., 2019a; Cooper, 2018); the difference is whether students estimated (case-value plot strategy) or calculated (count-and-compute strategy) the mean. Hence, almost all strategies can be attributed to one of two classes: one in which students correctly interpreted the graph as a histogram and one in which students incorrectly interpreted the graph as a case-value plot.

---

[46] Although a correct variant of this strategy is, in theory, possible, we did not find such a correct variant in the gaze data, nor in students' explanations.

Figure 4.5 Graphs (all single histograms) used in Item01 (upper left), 02 (upper right), 06 (middle left), 19 (middle right) and 20 (bottom)
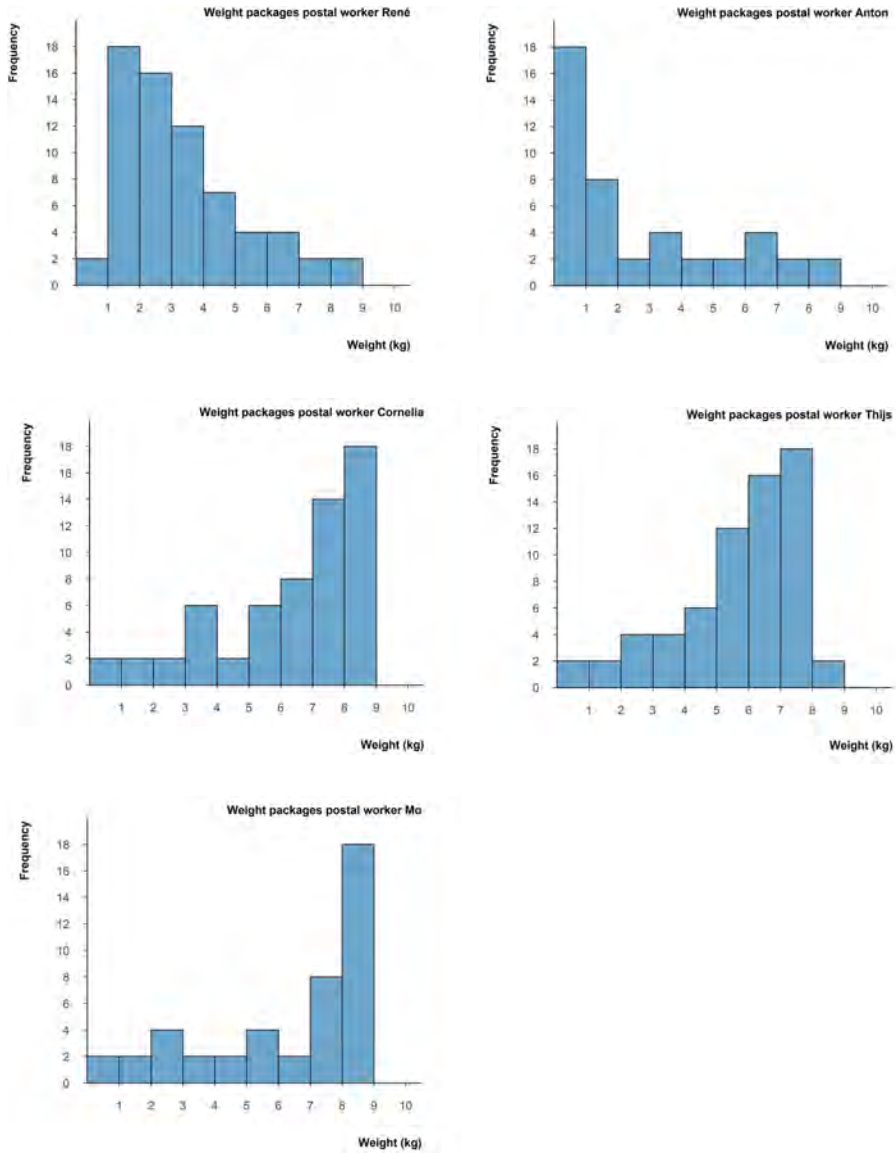
Table 4.1 Answers given by the students, $N = 50$ (see also Chapter 3, Table A.9)

Item	Correct answer	Answer range correct answers [a]	Average of given answers	Number of students correct	Percentage of students correct
Item01	3.3	2.2–4.4	5.6	20	40%
Item02	2.7	1.6–3.8	3.8	19	38%
Item06	5.7	4.6–6.8	6.9	25	50%
Item19	6.4	5.3–7.5	6.9	34	68%
Item20	6.3	5.2–7.4 [b]	6.7	17	34%

Note. [a] Experts were asked to give an answer to these items as well. Based on these results as well as students' preference for whole numbers, the answer range was set to +/-1.1 for all items.
[b] If the answer 7.5 had been included, 18 students would have answered correctly. Furthermore, 10 students answered 5 for Item20.

Table 4.2 Strategies, percentage of trials (correct strategy in **bold**), $N = 50$ per item (see also Chapter 3, Table A.12)

Item	Histogram strategy	Case-value plot strategy	Count-and-compute strategy	Unclear
Item01	**48%**	30%	20%	2%
Item02	**46%**	38%	12%	4%
Item06	**38%**	42%	20%	0%
Item19	**44%**	42%	14%	0%
Item20	**46%**	34%	16%	4%

A second coder coded 10% (25 trials). The interrater reliability of the coding in Table 4.2 measured with a Cohen's Kappa of .62 is considered substantial (Landis & Koch, 1977). Four out of five disagreements involved the second coder choosing a count-and-compute strategy and the first coder choosing one of the other strategies. If this coding is aggregated to correct (histogram strategy) and incorrect (all others)—as used as input for training the machine learning algorithm—agreement goes to 22 out of 25 trials, which corresponds to a Cohen's Kappa of 0.73 (substantial).

We used a code for a correct (bold) or incorrect strategy (all other, mostly misinterpreting the histogram as a case-value plot) as input in the training phase of the MLA and compared these to the results of the IMM and of the testing phase of the MLA (we explain later how the testing was done relative to the training). The eye movements belonging to correct strategies were mainly vertical and answers were read on the horizontal axis, as the data in a histogram are positioned along this axis (Figure 4.6).

In contrast to most eye-tracking studies, in the qualitative study, we looked at the *perceptual form* of the scanpath, for example, the vertical gaze pattern (Figure 4.6), and refer to this as a stable scanpath if it includes multiple aligned fixations and saccades along this scanpath and was explicitly mentioned by at least some students as being relevant for their strategy (e.g., Boels et al., 2018, 2019a; Chapter 3). This vertical line is formed by looking back and forth between the balance point of the graph on the horizontal axis and the height of the bars as a weighting factor. Incorrect strategies contained mainly horizontal gaze patterns and searching for the answer on the vertical (frequency) axis—hence using incorrect data—and leveling all bars (Figure 4.1 in the section on the Theoretical interpretation of students' gazes).

Figure 4.6 Part of a stable vertical scanpath (left) and all gazes (right) on Item06
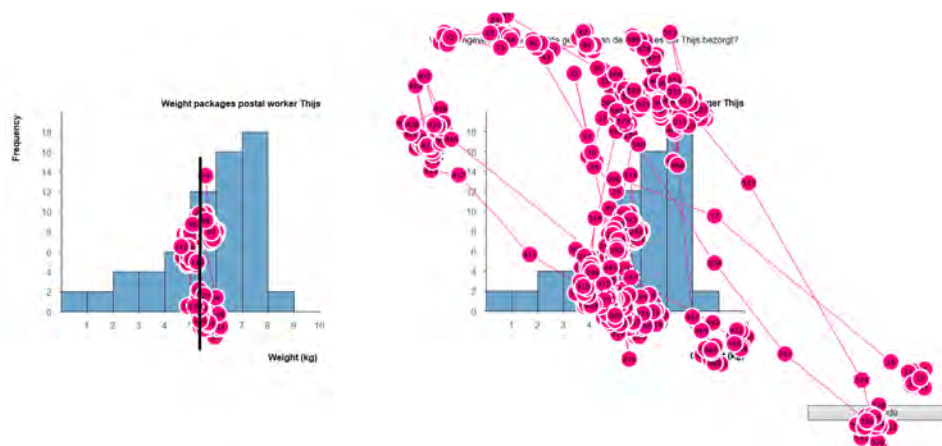
Note. The left figure reveals the vertical line segment (left: superimposed for the reader) in studentL26's gazes on Item06 (right: all gazes). Circles indicate fixations, thin lines indicate saccades: fast transitions between two fixations. The left figure is translated for the reader's convenience. These gazes indicate a correct (histogram interpretation) strategy.

4.3.2   Random forests machine learning analysis

For the machine learning analysis we used the random forest MLA that is implemented in the Mathematica software as described in a previous section, both for training the MLA—with a subset of students—and for the classification of the remaining students by the trained MLA. The aim of our ML analyses was to set a baseline for an IMM (see next section). We started our analysis with Item06 from the original 25 items in the qualitative study (e.g., Chapter 3). The first reason for choosing this item is that 50% of the students answered this item correctly (Table 4.1) which is—in theory—an ideal situation for MLA. If it did not work with this item, we would not expect the MLA to

work on other items. The second reason was that we expected students to have settled on a strategy by the sixth item of the original study—in line with our observations from the qualitative study—possibly making the identification of strategies easier than for the first two items. The third reason was that this graph is skewed to the left, making large horizontal eye movements—part of the incorrect strategy—more likely to be explicit than in graphs that are skewed to the right (e.g., Figure 4.7).

Figure 4.7 Example of all gaze data on Item02 (left) and Item06 (right) of studentL27
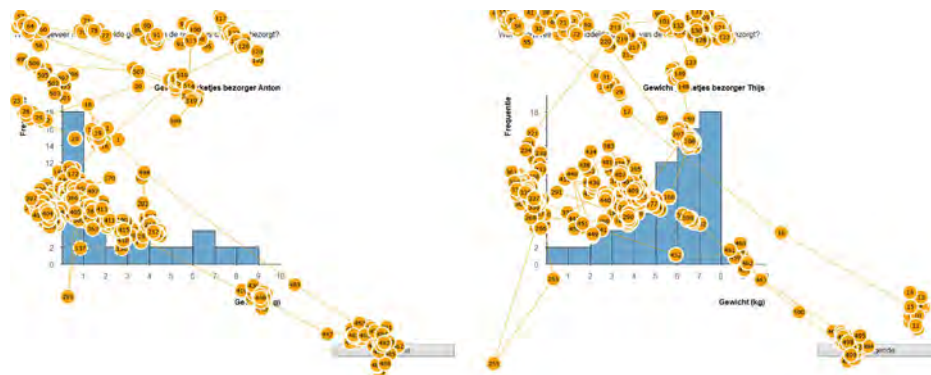
Note. This student incorrectly answered six and eight, respectively. The incorrect strategy is visible by the many horizontal saccades going from the left-hand side of the graph to the middle or right-hand side and the absence of vertical gazes going from the top or middle of bars to the bottom of the graph (see also Fig. 4.2 and 4.5 for comparison). The horizontal saccades in the gaze cloud on the graph area in Item02 (left) are smaller than in Item06 (right). On both graphs, this student applied an incorrect strategy, even though for the first items (e.g., Item02) this student looked at the titles of the axes[47].

In the present study, we used a supervised MLA. The MLA is fed with the raw data of the gazes: the *x*- and *y*-coordinates of the eyes for selected timestamps and the correctness of the answers (0 = incorrect, 1 = correct). As the stable scanpath (see students' strategies section) occurred only in the graph area, we did not use other AOIs here. Note that we are interested in *task-specific* strategies, not in general reading or viewing strategies. Both the previous qualitative study and another study (Lyford & Boels, 2022) suggested that reading axes was not a relevant part of such a strategy and would add noise when used in an MLA. Examples of other AOIs were the horizontal label (title of the horizontal axis), vertical label, horizontal axis, vertical axis, graph title,

---

[47] Looking at statistical graphs (e.g., concentrations of greenhouse gases from 0 to 2005), experts tend to spend more time on AOIs that help them understand the data in the graph (title, legend, axes) compared to novices (Harsh et al., 2019). The gaze pattern of studentL27 on Item02 indicates that attending axes and graph titles might not be enough.

question, and 'next' button. We filtered and prepared the data as follows. First[48], we removed all data that fell outside the computer screen (all rows with *x*- and *y*-coordinates outside the range 0–1366 horizontally, and 0–768 vertically). Then, we calculated new *y*-coordinates as 768 minus the original values, as the coordinate system is upside down in the Tobii software compared to the Cartesian plane commonly used in mathematics and preferred in Mathematica[49] so that a *y*-coordinate of 700 in Tobii indicates a position close to the bottom of the screen. In Mathematica, we selected the data that Tobii indicated as being within the graph area (signposted by a 1 in the column 'AOI graph' in the dataset) and then selected the coordinates that were in the graph area (pixels 500 to 1100, horizontally, and pixels that were originally between 190 to 525, vertically, for all items, Figure 4.8). We also removed few bad data (start and end line of the gazes on an item, as well as incomplete data due to data loss as described elsewhere). The order of the gaze data was kept in the input file but without timestamps.

The tool we used (the Classify function in Mathematica version 12.1) initially automatically chose random forest (considered to be a high-performance model, Kuhn & Johnson, 2013) as the best MLA for our continuous gaze data (WRI, 2020). During the review process, we made once more all analyses with the newest version of Mathematica, 13.2.1. Instead of the random forests, the software now suggested that logistic regression performs slightly better in most cases, and random forests in some. However, our analyses that are based on random forests still hold as the conclusions that can be drawn from these analyses with logistic regression are the same. As the MLA is intended as a baseline for our IMM, we, therefore, report the results of random forests for all items in the remainder of this article. We seeded the random forests and prescribed it as method in follow-up analyses (see code line below). Although random forest is known for its explainability the way it is embedded in the software made us consider it a black box. This ML-model is not to be confused with the IMM we constructed and is described in the next section. The following code line was used for obtaining a trained classifier function in follow-up analyses (some detailed codes can be found in the Appendix A of this chapter):

---

[48] Even before that, some data cleaning was partially done by hand, as, due to the huge amount of data, the Tobii software could not deliver the data in one database with all students and items together. Instead, all data were delivered per student (25 trials per student), whereas we wanted to have all data per item (50 trials per item). Moreover, the original dataset contained few empty lines that needed to be removed.

[49] In version 13.2.1, the vertically flipped coordinate system used by the Tobii software now can also be used.

```
Classify[{datarow1 -> class1, datarow2 -> class2 ... datarow50 ->
class50}, Method -> "RandomForest"]
```

Here, each data row is a sequence of *x-y* pairs—the coordinates of the gaze locations on the screen, in pixel units, keeping the original order of the fixations—and each class is either 0 (incorrect strategy) or 1 (correct strategy). Classify returns a classifying function. After this training cycle, a list of data rows is fed to this Classify Function to obtain the algorithm's classification of the data. It returns a list of zeros and ones: the algorithm's identification of students' strategy correctness.

Figure 4.8 Example of the AOI graph area (yellow rectangle) in Item06
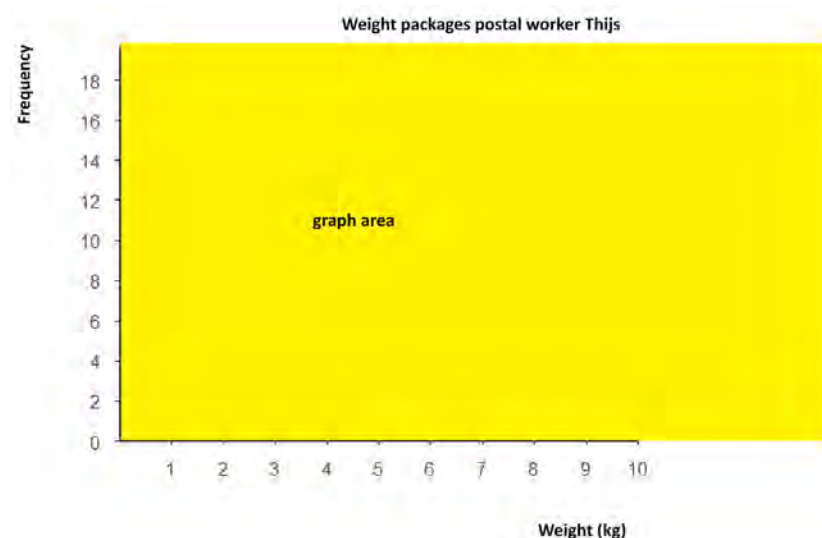
Note. The size and place of this AOI are the same for all five items.

The ML analysis consisted of two steps: (1) verification of whether prerequisites were met (see the section Considerations for strategy identification with machine learning algorithms) and (2) identification of students' strategies. At each step, the MLA was re-trained. We repeated these two steps for all five items. As the results of the second step were above our expectations, we decided to take an extra step: training the MLA with one item and testing the MLA with another; see the Results of applying an MLA and IMM section. We did this for all item pairs. In addition, we performed several cross-validation procedures as described in the Methodological evaluation criteria section.

4.3.3 Construction of an interpretable mathematical model

In the second phase, we constructed an interpretable mathematical model (IMM) based on attentional anchors (AA) found in a qualitative analysis of students' strategies in a previous study (see Chapter 3). Two AAs were found: an imaginary horizontal line and an imaginary vertical line. We tried several ways of capturing these two strategies mathematically, with varying success. The best model we found relies on the following algorithm, which is based on saccade lengths and angles. The cut-off values below were found empirically by testing many values close to a slope of 1. The starting point 1 for the slope followed from insights about the two AAs from the qualitative study. For each participant, we transform the sequence of saccades on the graph area into a sequence of -1, 0, and 1 values (Table 4.3):

- If a saccade is less than 200 pixels long (Euclidean distance), map it to 0.
- If a saccade is at least 200 pixels long and the absolute value of the slope of the saccade line is greater than or equal to 0.875 (or ⅞), map it to 1 (these saccades are considered vertical).
- If a saccade is at least 200 pixels long and the absolute value of the slope of the saccade is less than 0.875, map it to -1 (these saccades are considered horizontal).

Our algorithm continues as follows:

- Split the -1, 0, 1 valued sequence into subsequences of identical consecutive values.
- Delete the duplicates in each subsequence of the sequences and join the subsequences.
- Remove the "0" cases; this is equivalent to disregarding saccades that are "short".
- Add up the elements of our sequence.
- If the total is negative or 0, we replace the sequence with 0; if the total is positive, we replace the sequence with 1. This counts the number of runs of consecutive "long" horizontal or vertical scans, and based on the total, determines whether there were more horizontal or vertical sets of long scans. Replacing the sequence with a value of 0 indicates that the scanning is "mostly horizontal," if we disregard short saccades and regard consecutive long scans with similar slopes as a single "scanning run".

Table 4.3 Example of applying the algorithm for the IMM

Step in algorithm	Result
1–3	{-1, -1, 0, 0, -1, 1, 1, 1}
4	{{-1, -1}, {0, 0}, {-1}, {1, 1, 1}}
5	{-1, 0, -1, 1}
6	{-1, -1, 1}
7	-1
8	0

An illustration of what we do is given in the following graphs, in which the scans that are considered horizontal are in red and the vertical ones in blue (Figure 4.9). The green saccades are shorter than 200 pixels. The color becomes lighter as the sequence progresses, to get a sense of the order in which the saccades occurred. In the left-hand graph, there are more blue lines than red ones, indicating a correct strategy for finding the mean in a histogram. In the right-hand graph, there are only red lines, so this scanning is regarded as horizontal. The 200 pixels saccade length cut-off point was fixed, thus not scaled to the width and height of the graph area (AOI). The size of the AOI in Figure 4.9 is indicated with a black rectangle (not present in the item) and is the same for all histogram items. As the AAs are similar on all five items, we constructed one IMM for all items. This means that the IMM is a more general model compared to the random forest models as the latter are different for each item.

Figure 4.9 Examples of horizontal and vertical eye movements that were counted in the IMM
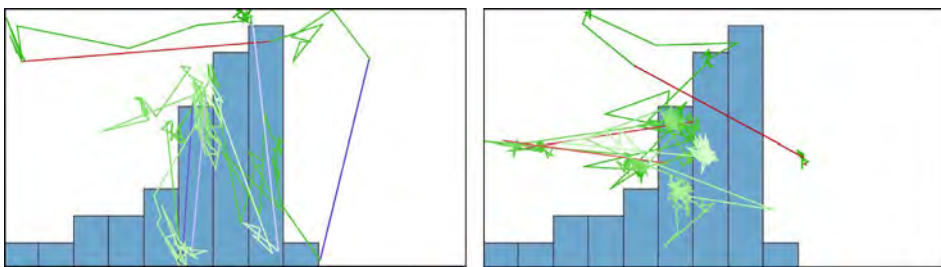
Note. Red—long horizontal—gazes correspond to value 1 in step 7 of the IMM, blue—long vertical—gazes correspond to 0 in step 7. Lighter colors indicate later occurrence. Green saccades are less than 200 pixels (hence disregarded). Left is an example of a correct strategy for Item06 (more blue vertical gazes), right an example of an incorrect strategy (more red horizontal gazes). Readers are referred to the online enlargement of this figure for subtle differences in coloring.

4.4 Results of applying an MLA and IMM

The details of the results, including confusion matrices, can be found in the Appendix A of this chapter (Tables A.1–A.8).

4.4.1 Machine learning algorithm results

Supervised machine learning with students' answers' correctness
The identification accuracy of the MLA for students' answers for the first item we looked at, Item06, turned out to be 88% (very good) with a jackknife cross-validation procedure. As we aimed to identify strategies, not students' answers, we needed accuracy to be at least 70%. This criterion is met for all items when using version 13.2.1 (Table 4.4). Based on the confusion matrices (Tables A.2–A.3) sensitivity and specificity were calculated (Table A.1).

Next, we trained the MLA with gaze data on one item and then identified answers on another item (Table A.6) for the results. Accuracies varied between chance level (32%) to well above (70%).

Table 4.4 Accuracies of the IMM and of the random forests MLA after cross-validation

Validation procedure	Classification of:	Item01	Item02	Item06	Item19	Item20
Jackknife	Answers (mean)	83%	80%	88%	86%	82%
Jackknife	Strategies (mean)	71%	71%	86%	88%	83%
Leave one out	Answers	66%	68%	74%	60%	58%
Leave one out	Strategies	60%	38%	64%	78%	48%
5-fold	Answers	62%	62%	62%	58%	68%
5-fold	Strategies	56%	56%	64%	74%	62%
IMM	Strategies	62%	70%	84%	70%	72%[a]

Note. [a] With a small adjustment of the slope, this could go up to 74%. Accuracy is expressed as a percentage of correctly predicted or identified cases (e.g., Afonja, 2017). The results of the qualitative study are treated as the positive case.

Supervised machine learning with students' strategies' correctness
The accuracy of the random forests MLA for identifying students' strategies in the first item we looked at, Item06, turned out to be 86% with a jackknife cross-validation procedure. This result is considered very good. Consistency of the MLA was tested through various procedures such as jackknife, leave-one-out cross-validation, and 5-fold cross-validation (see section Methodological evaluation criteria). Overall, the MLA correctly identifies 71% to 88% of students' strategies (jackknife cross validation) for five different (but all histogram) items (Table 4.4). Percentages for correct strategies in the qualitative study (Table 4.2) varied between 38% and 48% (or 52%– 62% when reversed) and strategies identification results are all well above these chance

levels with jackknife. When applying other cross-validation procedures, results drop and vary from around chance level (38%) to good (78%) for leave-one-out cross-validation to around chance level (56%) to good (74%) for 5-fold cross-validation. We consider these results a baseline for the IMM.

Next, we trained the MLA with gaze data on one item and then identified strategies on all other items. Accuracies varied between around chance level (52%) to quite well above that level (80%; Table A.7). Specifically interesting are the results when gazes on Item01 are used as training data (Figure 4.10). Testing this trained random forest MLA for identifying strategies on other items resulted in accuracies that vary between good (72%) and very good (80%) which suggests that the MLA has the potential to generalize beyond a specific item, even when shapes and skewness of the histograms differ. See also the ROC plot (Fig. 4.10).

Figure 4.10 ROC plot for strategies with train-test item pairs for random forest
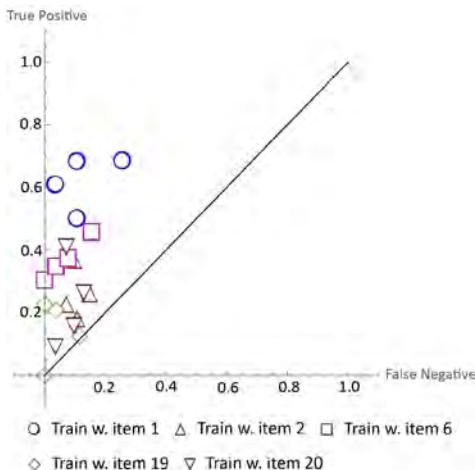
Note. Ideally, for educational use, points should be concentrated in the upper left corner of the plot and close together for all items.

Based on the confusion matrices (Tables A.4–A.5), specificity and sensitivity (e.g., Kuhn & Johnson, 2013) were calculated (Table 4.5). Sensitivity (identification of correct strategies) is low to acceptable and lower than the low to excellent specificity (identification of incorrect strategies) after cross-validation. An excellent specificity is favorable for a future application that seeks to provide feedback to this particular group of learners. In practice this could mean that only a few students who used an incorrect strategy will be missed (Q-incorrect, MLA-correct, type I error) which we consider most important for feedback. In addition, some more students who used a correct strategy would get feedback implying that they used an incorrect strategy (Q-

correct, MLA-incorrect, type II error) while they were not. As such feedback would hint at the correct strategy, this feedback could make them think once more and then conclude that their strategy was correct, which is not a problem.

Table 4.5 Sensitivity and specificity of the MLA and the IMM when classifying strategies

Cross-validation procedure	Metric	Item01	Item02	Item06	Item19	Item20
Leave one out CV	Sensitivity	0.58	0.45	0.21	0.64	0.39
Leave one out CV	Specificity	0.62	0.32	0.90	0.89	0.56
5-fold CV	Sensitivity	0.50	0.50	0.21	0.59	0.57
5-fold CV	Specificity	0.62	0.61	0.90	0.86	0.67
IMM	Sensitivity	0.58	0.46	0.74	0.36	0.43
IMM	Specificity	0.65	0.89	0.90	0.96	0.96

Note. Calculation of sensitivity and specificity is not possible for jackknife.

The confusion matrices (see Appendix A of this chapter), provide further insight into how well the results of the MLA align with the results of the qualitative coding. Differences between the qualitative coding of the strategies and the MLA results can be due to what data scientists call ground truth noise in the data. Ground truth, here, is what the strategies actually "are" (in the real world, independent of coding). Educational researchers would explain this noise as inconsistencies, inaccuracies (e.g., due to merging two different but similar strategies in one code) or errors in the qualitative coding (as coders usually do not fully agree on the qualitative codes), or as noise in the gaze data (e.g., not every fixation or saccade on the graph area being part of the strategy). Differences can also be used to reconsider qualitative coding.

4.4.2   Results from the interpretable mathematical model

With the IMM described earlier, we can correctly identify 62% to 84% of students' strategies (Table 4.4). From the confusion matrices (Tables 4.6–4.7) that compare the results of the IMM with the results of the qualitative study (Q), sensitivity and specificity can be calculated (Table 4.5). Sensitivity varies between low and good; specificity varies between acceptable and excellent, see also Figure 4.11.

Table 4.6 Confusion matrices for Item01, 02, and 06 (IMM)

	Item01		Item02		Item06	
	Q-correct	Q-incorrect	Q-correct	Q-incorrect	Q-correct	Q-incorrect
IMM-correct	14	9	10	3	14	3
IMM-incorrect	10	17	12	25	5	28

Note. N = 50 per item.

Table 4.7 Confusion matrices for Item19 and 20 (IMM)

	Item19		Item20	
	Q-correct	Q-incorrect	Q-correct	Q-incorrect
IMM-correct	8	1	10	1
IMM-incorrect	14	27	13	26

Note. N = 50 per item.

4.4.3   Comparison of IMM and random forests MLA results

The accuracy results of the IMM are quite close to the results of the random forests MLA with the jackknife cross-validation. In addition, the overlap between the IMM and MLA in identifying students' strategies before cross-validation was good and varied between 66% and 82% (Table A.8). The results of the IMM are better than the MLA after cross-validation. Moreover, the results of both the IMM and the MLA indicate that strategies might be clearer in Item06, which is in line with what we qualitatively found. In the ROC plot (Figure 4.10) the MLA results after cross-validation are compared to the IMM results. The plot shows that the IMM performs better. Altogether, this is considered a very good result, as it is not possible to know what parameters our MLA is using. As the accuracy of strategy identification of the MLA and the IMM is sufficient or above, the association is also sufficient or above.

The IMM is a more general model than the random forest models as the latter is different for each item. Given the MLA as a baseline for the IMM, we consider the current IMM likely to be a main component of a more general and precise model explaining gaze behavior during the kind of cognitive tasks considered in this article. The fact that the MLA, when trained on one item, relatively successfully predicts performance in a different item can be considered as evidence that the gaze data contain information about cognitive behavior on this task, even though this behavior is not explained by the MLA.

Figure 4.11 ROC plot for strategy identification (Fawcett, 2006) in which "the point (0, 1) represents perfect classification" (p. 862). Random forest is the MLA the IMM is compared with
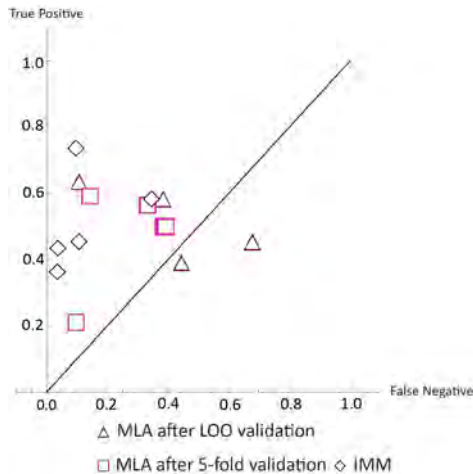
Note. Ideally, for educational use, points should be concentrated in the upper left corner of the plot and close together for all items. One triangle is hidden behind the square in the lower left corner. The MLA provides a baseline for the IMM. Although the IMM worked well, the plot shows there is room for improvement.

As this chapter aims to provide proof of principle, we did not further optimize the IMM. We chose the IMM that had the best overall performance for all items among the models we tried. When aiming to use the model in an application, one way to refine the IMM could be to split the model into two models: one only for identifying incorrect strategies (with two possible outcomes: the strategy is incorrect[50] or unknown) and one only for correct strategies. Combined, both models would have four options for strategy identification: correct strategy, incorrect strategy, unknown, or contradicting outcomes. Unknown means that the strategy is unknown, contradicting outcomes would require an extra rule for deciding what strategy it is. Another idea for optimizing these two models—besides adjusting the slope that distinguishes between horizontal and vertical gazes to obtain better results for specific items—is to adjust the saccade length in the model. In the current model, only saccades of at least 200 pixels are considered for both horizontal and vertical gazes. The saccade length for the vertical gazes could be scaled (shortened) to the size of the AOI. The relatively small height of the graph area (335 pixels) as opposed to the width (600 pixels) would justify such an

---

[50] This incorrect strategy could also be further divided into a case-value plot interpretation strategy and a count-and-compute strategy not further discussed here.

adjustment. A possible improvement of the IMM could also be to incorporate alignment of saccades (relevant for the most common strategies). As the aim of the IMM in this research is not to find the best model but to show that the perceptual form of the stable scanpath can be captured by a model, we did not try to further optimize this model.

In addition, the performance of both IMM and machine learning might be influenced by students switching their strategy during or in between trials, making strategy identification less clear. There is some evidence in the interview data (e.g., Chapter 3 and below) that students' strategies were influenced by item13 to item18 (with dotplots) which were designed to scaffold students in applying a correct strategy (e.g., Lyford, 2017). Although this did not result in a higher number of correct strategies, it might have changed the strategies on a more subtle level.

StudentL22:   Yes then [Item19] I was going to change my approach a little bit. Then I started doing it a little bit similar to what I did with the dots. So then here is about more and there is again less, so then it will be somewhere here in between.

4.5   Conclusions and discussion

Automated identification of students' strategies is a prerequisite for targeted intelligent feedback. The present study took on this challenge by providing an example of automated identification of students' *task-specific* strategies on single histograms based on gaze data. This could allow for future applications such as real-time feedback based on gaze data, for example, collected through webcams (e.g., Knoop-Van Kampen et al., 2021).

We analyzed a set of gaze data in three phases: (1) an analysis of raw gaze data on *one* AOI (that contained the relevant scanpath) through an MLA that provided a baseline for the second step; (2) the construction of a separate interpretable mathematical model (IMM) using the same gaze data and insights on the perceptual form of this stable scanpath from previous qualitative research; and (3) an evaluation of the results by comparing the performance of the MLA, the IMM and the overlap between the two. The IMM outperformed the MLA in several cases.

The MLA (phase 1, random forests as implemented in the software Mathematica Classify Function) has the advantage that it can process raw gaze data (*x*- and *y*-coordinates). It has the disadvantage that it is a black box in that it does not explain how it reached its decision for an individual student. The results of the random forests MLA after cross-validation provided our baseline for the IMM (phase 2). The IMM performs well (62% to 84% accuracy) with the

advantage of being transparent for individual decisions and theoretically meaningful. The overlap between the results of the IMM and MLA is sufficient (phase 3).

Although these results are encouraging, an issue in data science is whether an MLA trained on one item is able to identify students' strategies for similar but different items. We, therefore, trained the MLA on one item and then tested it on all other items, and repeated this until all items were used once as training items. The accuracy results varied between around chance level and well above. The latter indicates very good performance. Combined, these results indicate that the IMM and MLA do describe the same phenomena—strategies—and that these strategies can be derived from students' gaze data.

What is new in our approach is that the filtering and preparation of the gaze data for the IMM and MLA are based on one AOI that contains the perceptual form of the gazes, instead of, for example, the number of transitions between AOIs. This perceptual form is a stable scanpath indicating the student's focus of interaction and is interpreted as an attentional anchor (AA, e.g., Abrahamson & Sánchez-García, 2016). In our previous research (Chapter 3) we found that this perceptual form is indicative of students' strategies.

A prerequisite for the approach used is that a stable scanpath has been found in the gaze data. Furthermore, it requires gaze data to be classified into groups of students using the same strategy. Both prerequisites were met for our study. A further prerequisite for ML analysis is that the dataset is large enough and contains enough information for the MLA to identify (classify) students' strategies.

A limitation of our study is that we had gaze data on only fifty students. We alleviated this limitation by showing it worked for five items with differently shaped histograms, by using a resampling approach (jackknife cross-validation), and by training the MLA with gaze data from one item and then having it identify strategies for all other items. For future research, collecting data from a larger and different population is recommended. Another limitation of our study is that we use one item type (single graphs) and five variants of one graphical representation (histograms). It would be interesting to apply our approach to other domains, following the three phases described above for each new topic. Once the IMM is optimized and the MLA is trained, both an IMM and an MLA could be implemented. The MLA might be more accurate (see results of the jackknife cross-validation) but is computationally more complex. The IMM is easier, faster, provides insight into the relevant part of the gaze pattern that its decision was based on, and can deal with partial data. This allows feedback on strategies before an answer is given. The

agreement between the IMM and MLA can be used as an indication of how certain the strategy identification was.

From a methodological perspective, a first contribution of this study is that our approach is both item-specific and generalizable. It is important that both IMM and MLA are item-specific, as the mathematical strategies of students are specific to an item type. Our IMM and MLA meet this requirement since the stable scanpath is specific to a given item; a stable scanpath refers to the *perceptual form* of this scanpath (e.g., a horizontal line, triangle, or point), not the sequence of AOIs. We expect that a stable scanpath can be found in gaze data on items in various domains (e.g., Strohmaier et al., 2020). It is also important that this method is generalizable, and, therefore, suitable for other mathematical domains. First, IMMs and MLAs (e.g., Rudin, 2019)—such as our set of rules for the IMM—are general methods. Moreover, our approach is also theoretically generalizable in the sense that educational researchers aim for "how and why the studied events occurred (or not)" (Yin, 2013, p. 326). The studied events are students' strategies and are observed as stable scanpaths that indicate students' focus of interaction with the item.

It could be argued that our approach is not very generalizable as *x*- and *y*-coordinates are sensitive to, for example, scaling, position of the graph on the screen, and shape of the histogram. However, the same could be argued for AOIs, as AOIs are *x*- and *y*-coordinates binned into categories by researchers and, therefore, are also item-specific. Furthermore, coordinates can be rescaled which adds to their generalizability. In addition, we showed that for five differently shaped histograms, the IMM performed above chance level to good and the MLA performed at chance level to very good after cross-validation. In addition, the same IMM was used for all five items which makes it a more generable model than the MLA that was initially retrained for each item. Also adding to the this generalizability is that, for several items, we have successfully trained the random forest MLA with this one item and then tested it for all other items.

A second methodological contribution is that *all* gaze data on one AOI are used, in contrast to methods that use aggregated data such as the *transition* from one AOI to another (e.g., Garcia Moreno-Esteva et al., 2020). Unlike heatmaps that are produced afterward (e.g., Schindler et al., 2021), raw gaze data offer the possibility of real-time feedback.

Third, we show that it is possible to *automatically* identify students' task-specific strategies from their gaze patterns. As soon as a stable scanpath is found, we think this can be grasped in an IMM as well as through an MLA, hence, be automated. Scanpaths are found for tasks in various mathematic domains: numbers (Schindler, et al., 2021), arithmetic (Green et al., 2007), proportional reasoning (Shayan et al., 2017), area and perimeter (Shvarts,

2017), Cartesian coordinates (Chumachemko et al., 2014), geometry (Schindler & Lilienthal, 2019), trigonometry (Alberto et al., 2019), parabola (Shvarts & Abrahamson, 2019), statistical graphs (Chapter 3) and more (Lilienthal & Schindler, 2019; Strohmaier et al., 2020). Therefore, we believe that a similar approach can be used for other topics.

Fourth, using raw gaze data opens the possibility of implementation in an online (e.g., Cavalcanti et al., 2021) and adaptive tutoring system (e.g., Scheiter et al., 2019) with real-time feedback. The IMM can process raw data directly without the (black-box) preprocessing that the MLA performs and the results are straightforward to interpret.

Fifth, studying the differences between the results of the IMM and MLA on the one hand, and the qualitative coding on the other hand have the potential to improve the qualitative coding. Whenever the three methods lead to a different outcome, a closer inspection of the gaze patterns on the item, combined with interview data (if available), may lead to new insights for qualitative coding. This would combine the best capacities of people and machines, as suggested by Van de Schoot (2020). Sixth and final, our approach offers a new road for replicating results from a qualitative study.

From a theoretical perspective, this study shows that an AA can be used as a theoretical lens to search for a stable scanpath that reflects a mathematical strategy that is meaningful to the students. These stable scanpaths can be linked to the idea of an AA as follows. In a retrospective recall, students talked about an imagined action. This imagined action is—according to students—coordinated by an imaginary mathematical object: a horizontal or vertical line. As an AA is an existing or imagined object or area that emerges to facilitate or coordinate sensorimotor actions (Abrahamson & Sánchez-García, 2016), we also interpret these lines as an AA.

In addition, the manipulation of an imaginary object manifest in the gaze data could suggest links between theories of mental processes and embodied cognition. The AA was previously found when students interact with the environment. In our items, students cannot physically manipulate the graph. Nevertheless, gaze data show a stable scanpath indicating scanning this imaginary object corresponding to a concrete location on the graph area on the screen (Chapter 3). We believe this reveals that cognitive processes can also be embodied and that eye movements can be a manifestation of both the perception and the action.

This chapter may fuel the dialogue between educational researchers and data science experts. An advantage of our IMM is its interpretability. MLAs may be experienced as a black box and educational researchers may focus on how well it performs rather than how it performs. As educational researchers, we wondered what the application of data science tools to our data would

bring us. It is important to promote the dialogue between educational researchers and MLA-experts, to keep boundaries between disciplines permeable. At such boundaries, exciting new research can emerge.

Future research into finding stable scanpaths for applying this method might, for example, concern geometry such as the Pythagoras theorem or the cosine rule. In calculus, one might consider interpreting the slope or direction field when learning to solve differential equations. To allow task-specific gaze patterns to emerge an alternative way of introducing a topic could be considered (e.g., Janßen et al., 2020). Finally, the domain of graphical or diagrammatic literacy could be a future line of research. Several examples can be found in the literature of students having difficulties with graphs in mathematics (e.g., difficulties with complex line graphs, Carpenter & Shah, 1998; overgeneralization of linearity, Leinhardt et al., 1990; misreading of graphs, Roth & Bowen, 2001; inadequate strategies, Tai et al., 2006), but also in science education (e.g., Kragten et al., 2015). All these domains have in common that spatial patterns may play a role.

Another direction for future research could be to improve the IMM and to investigate the apparent trade-off between its sensitivity and specificity. We know that students used several strategies but it is unclear whether and how this is visible in this trade-off. In addition, a possible improvement of the IMM could be to tailor it to each item. Furthermore, the alignment of saccades could be included in the IMM (important for the most common strategies, see Chapter 3).

Future research might also focus on the appearances and changes of students' strategies over time. By using an IMM and an MLA, online automated feedback becomes possible on students' strategy, in some cases maybe even before students give their answers. This might make online feedback in massive online courses, online teaching, and homework more accurate and efficient. Another possibility would be to provide teachers with a dashboard on students' strategies (e.g., Knoop-Van Kampen et al., 2021). The agreement between an MLA and an IMM could then be used to provide a measure for how reliable the strategy identification is. A prerequisite is the availability of cheap equipment for measuring eye movements. We expect more exact measuring of eye movements will be available for consumer computers in the near future, for example, through webcams. Whether this will be implemented in software and used by consumers will also depend on ethical discussions about privacy, fairness, bias, et cetera.

Appendix A Additional code and results

In this Appendix we provide additional code and all results of the IMM, random forests ML-analysis, and cross-validation procedures.

Mathematica code used to find the best method for the data

```
pickMethod[answerTrainingData[[item]],answerData[[item]],met
hod, seed, performanceGoal]


tob = DateObject[]
beforeValidationResults =
Monitor[Table[{pickMethod[answerTrainingData 〚item〛 ,
answerData 〚item〛 ,
Automatic, "1234", Automatic], pickMethod[strategyTrainingData
   〚item〛 ,
strategyData 〚item〛 , Automatic, "1234", Automatic]},
    {item, 1, 5}], item]
DateObject[] - tob
Clear[tob]
```

Overview of the results for sensitivity and specificity
For the results of strategy classification, see Table 4.5 in the article. Below, the results for answer classification can be found.

Table A.1 Sensitivity and specificity of the random forests of answers

Validation procedure	What	Item01	Item02	Item06	Item19	Item20
Leave one out	Sensitivity	0.26	0.32	0.72	0.82	0.00
Leave one out	Sensitivity	0.90	0.90	0.76	0.13	0.88
5-fold	Sensitivity	0.11	0.05	0.40	0.68	0.12
5-fold	Sensitivity	0.94	0.97	0.84	0.38	0.97

Confusion matrices of answers

Table A.2 Confusion matrices of answers for all items, random forests MLA, after LOOCV

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	5	3	6	3	18	6	28	14	0	4
MLAi	14	28	13	28	7	19	6	2	17	29

Note. N = 50 per item. The results of the qualitative study (Q) compared with the results of the MLA random forests, c = correct answer, i = incorrect answer. The qualitative study is treated as the positive case. For example, the number 5 in the top-

left corner of Item01 stands for 5 students identified by both the qualitative study and the MLA as having a correct strategy.

Table A.3 Confusion matrices of answers for all items, random forests MLA, after 5-fold cross-validation

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	2	2	1	1	10	4	23	10	2	1
MLAi	17	29	18	30	15	21	11	6	15	32

Note. N = 50 per item. The results of the qualitative study (Q) compared with the results of the MLA random forests, c = correct answer, i = incorrect answer. The qualitative study is treated as the positive case.

Confusion matrices of strategies

Table A.4 Confusion matrices of strategies for all items, random forests MLA, after LOOCV

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	14	10	10	19	4	3	14	3	9	12
MLAi	10	16	12	9	15	28	8	25	14	15

Note. N = 50 per item. The results of the qualitative study (Q) compared with the results of the MLA, c = correct strategy, i = incorrect strategy. LOOCV = leave-one-out cross-validation. The qualitative study is treated as the positive case.

Table A.5 Confusion matrices of strategies for all items, random forests MLA, after 5-fold cross-validation

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	12	10	11	11	4	3	13	4	13	9
MLAi	12	16	11	17	15	28	9	24	10	18

Note. N = 50 per item. The results of the qualitative study (Q) compared with the results of the MLA, c = correct strategy, i = incorrect strategy. The qualitative study is treated as the positive case.

Results of the analyses when using one item as training item

Table A.6 Accuracy of answer prediction for all items for the random forests MLA

	Item01	Item02	Item06	Item19	Item20
Item01	x	64%	68%	46%	62%
Item02	70%	x	52%	38%	64%
Item06	58%	56%	x	62%	60%
Item19	42%	48%	58%	x	38%
Item20	62%	64%	52%	32%	x

Note. N = 50 per item. The item in each row is used as training item. The items in the columns are the test items. For example, when using Item01 as training item, the accuracy of predicting students' answers on Item02 is 64%.

Table A.7 Accuracy of strategy prediction for all items for the random forests MLA

	Item01	Item02	Item06	Item19	Item20
Item01	x	72%	72%	80%	80%
Item02	66%	x	70%	58%	58%
Item06	66%	66%	x	62%	68%
Item19	60%	56%	64%	x	68%
Item20	52%	58%	62%	70%	x

Comparison of the random forests MLA and IMM results

Table A.8 Confusion matrices of strategies for all items for the random forests MLA versus the IMM

	Item01		Item02		Item06		Item19		Item20	
	MLAc	MLAi	MLAc	MLAi	MLAc	MLAi	MLAc	MLAi	MLAc	MLAi
IMMc	20	10	11	13	11	6	7	2	7	4
IMMi	4	16	4	22	3	30	9	32	8	31

Note. N = 50 per item. The results of the MLA random forests compared with the results of the IMM, c = correct strategy, i = incorrect strategy. The MLA before any cross-validation is treated as the positive case.

#

Assessing students' interpretations of histograms before and after interpreting dotplots: A gaze-based machine learning analysis

"The bulk of the world's knowledge is an imaginary construction." [51]
Helen Keller

---

[51] Keller, H. (1910), C*hapter 8, The Five-sensed World*. Quoted in: Keller, H. (2002, p. 289). *Organization & Environment, 15*(3), 285–292. https://www.jstor.org/stable/26162186

Abstract

5.1 Introduction

Statistical literacy includes "people's ability to *interpret and critically evaluate* statistical information, data-related arguments […], which they may encounter in diverse contexts, and when relevant" (Gal, 2002, p. 4; emphasis in original). As data consumers, citizens should be able to correctly interpret various graphical displays. This is particularly important in this era of vague and fake news "that place interpretive and evaluative demands on a reader or viewer" (Gal & Geiger, 2022, p. 2). In this study we specifically focus on the graphical representation of histograms.

Histograms can reveal particular aspects of the distribution of the data often hidden in other graphs (e.g., Pastore et al., 2017). Furthermore, as histograms are ubiquitous in research and education, they need to be learned (cf. Garfield & Ben-Zvi, 2008b). For example, searching for 'histogram' in Google Scholar resulted in more than 3.2 million hits (June 13, 2023). Therefore, the guidelines for assessment and instruction in statistics education II (GAISE II) for all Grades up to Grade 12 contain several examples of histograms and dotplots (for levels A, B, and C), with levels B and C roughly corresponding to middle and high school (Bargagliotti et al., 2020). Moreover, some alternatives for histograms, such as boxplots, are even more complex (e.g., Bakker et al., 2004, Lem et al., 2014a).

However, many people persistently misinterpret histograms (e.g., Cooper, 2018; Kaplan, 2014). For example, Bakker (2004a) found that secondary school students (Grades 7–8) considered the individual heights of bars in a histogram to be the heights of individual people, rather than aggregations of data. Students' conceptual difficulties with histograms are well documented (e.g., Chapter 2), but it is unclear how to support students in learning to interpret histograms.

Figure 5.1 Example of dotplot Item17 for which students were asked to compare two datasets regarding their mean
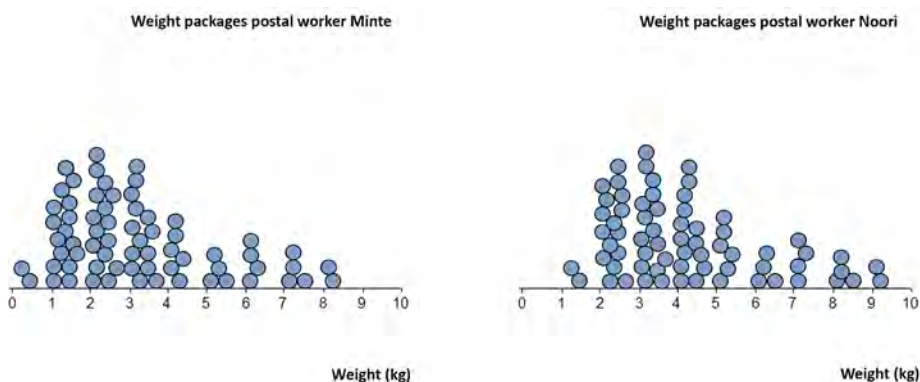
Several studies suggest that having students solve dotplot items can scaffold this learning (e.g., delMas & Liu, 2005; Garfield & Ben-Zvi, 2008b; Makar & Confrey, 2004). In most of these studies, students' answers and verbal reports were the main source of information. Dotplots have the advantage that they show all individual data points as well as their distribution (Figure 5.1). In addition, the absence of a vertical scale in dotplots can turn students' attention toward the horizontal scale, which is where the variable is presented in both graphs. However, little is known about whether solving dotplot items allows students to become aware of aspects of graph representation and statistical variables that are useful for interpreting histograms. The aim of this study is, therefore, to explore how solving dotplot items influences secondary school students' thinking on a detailed level when they interpret histograms. Our overall research question is: *In what way do Grades 10–12 pre-university track students' histogram interpretations change after solving dotplot items?* In the Theoretical background section, we will specify this overall question with three sub-questions.

Figure 5.2 Example of a dotplot (left) and a histogram (right) depicting the same distribution
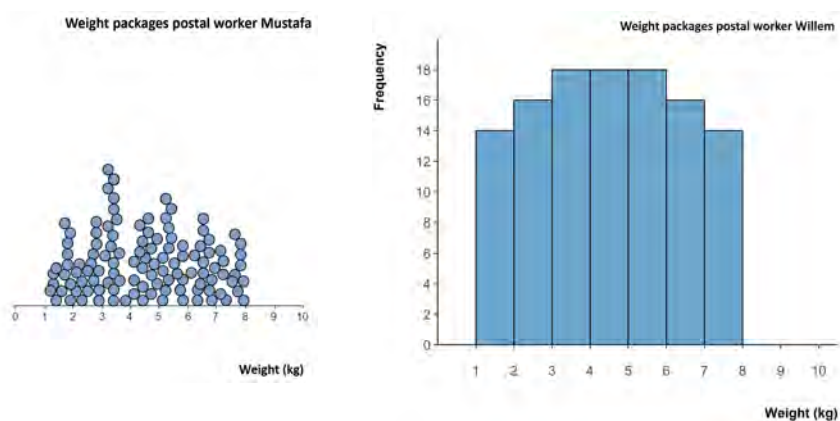
Note The dotplot was part of Item16. The histogram was part of Item05, not further discussed here (for more details, see Chapter 3).

As we elaborate further in the Theoretical background section, gaze data can reveal students' strategies in real-time, and in more detail, compared to concurrent thinking aloud (verbal reports) and without the risk of influencing the thinking process (Van Gog et al., 2005; Van Gog & Jarodzka, 2013). We use students' gaze data when solving four items with histograms before and after solving similar items with dotplots, as well as their answers on these items. The four histogram items were taken from a larger sequence with 25 digital items

in total. Furthermore, we examined transcripts from stimulated recall (Lyle, 2003) verbal reports about students' strategies (for more details see section Data collection methods: Eye-tracking, stimulated recall verbal reports). In the next section, we elaborate on difficulties with histograms and dotplots and discuss how gaze data can be used.

5.2   Theoretical background

5.2.1   Review of statistics education literature

In this section, we review statistics education literature on the problem (many students persistently misinterpreting histograms), a gap in this literature (the variation in results on students' interpretations of dotplots), and graphs that are suggested for supporting students in learning to interpret histograms.

Histograms are persistently misinterpreted
Many people persistently misinterpret histograms (e.g., Cohen, 1996; Setiawan & Sukoco, 2021). Researchers and teachers think that there is no difference between bar graphs and histograms (e.g., Clayden & Croft, 1990; Tiefenbruck, 2007). Dabos (2014) found that some college teachers did not see when students incorrectly counted the number of bars in a histogram to get the total frequency instead of adding the bars' heights. First-year university students in educational sciences had difficulties finding or interpreting the mean, median, variation, and skewness in histograms (Lem et al., 2013c). College students interpreted the horizontal salary scale in a histogram as a timescale (Meletiou, 2000). Middle school students used unequal intervals in a histogram with frequency on the vertical axis—instead of density—hence, not correcting the frequencies for unequal bin widths (McGatha et al., 2002). Other middle school students thought that bars in histograms are connected for easier comparison (e.g., Capraro et al., 2005). Students in Grades 6–12 answered histogram items 17% to 53% correctly on average (Whitaker & Jacobbe, 2017). Many students mistakenly took bars' heights as the measured value. Such students possibly think that only nine packages are depicted in the histogram in Figure 5.3 (the number of bars) instead of 67 (the actual number).

Dotplots are not always correctly interpreted
Generally, dotplots are interpreted better than histograms (e.g., delMas et al., 2005), although stacked dotplots (in the early days also called line plots; e.g., Tiefenbruck, 2007) might still confuse students (e.g., Lyford, 2017). Lem et al. (2013c) found that university students understood dotplots slightly better than histograms (on average, 55% correct responses for dotplots versus 51% for histograms). However, in that study two dotplot items scored worse. University students taking introductory statistics explored variability and

standard deviation through a kind of stacked dotplots (delMas & Liu, 2005). Most of these students did not fully understand how standard deviation was related to the distribution of data in a histogram.

Figure 5.3 An example of a histogram

Weight packages postal worker René
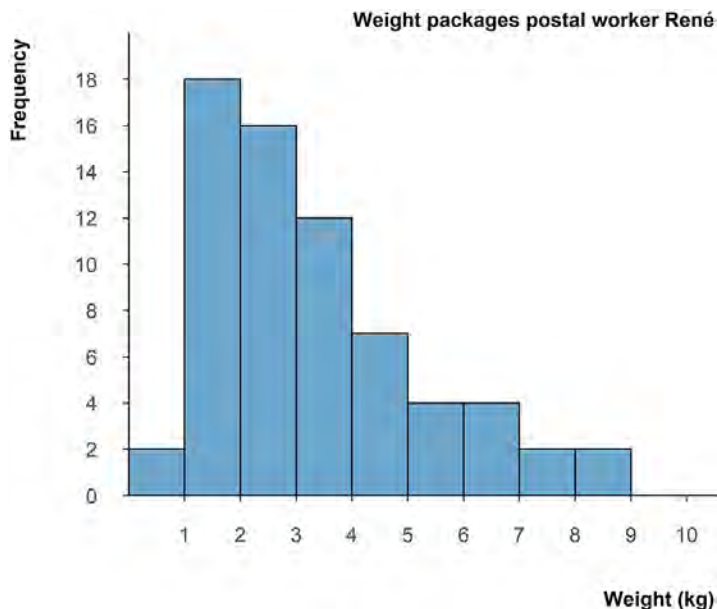
Note. The measured variable (weight) is along the horizontal axis. Weights of 67 packages (sum of frequencies) are depicted in this histogram. The arithmetic mean weight is 3.3 kg.

A local instruction theory in statistics education suggests that dotplots are suitable for supporting students' learning of distribution and variability in data represented in histograms (e.g., Bakker & Gravemeijer, 2004; Garfield, 2002). Garfield & Ben-Zvi (2008b) stated "studies [that] suggest a sequence of activities that leads students from […] dotplots […] to histograms" can support students in "developing the concept of distribution as an entity" (p. 175). An advantage of dotplots over histograms is that dotplots show the distribution of data in a disaggregated form. In addition, dotplots have the possibility to draw students' attention to the variable being depicted along the horizontal axis— similar to histograms—, as dotplots typically have only this axis. A possible disadvantage of dotplots for teaching students to interpret histograms (aggregated data) is that dotplots might invite them to see the data as individual cases (Konold et al., 2015) instead of looking at aggregated measures (including arithmetic mean).

One explanation for dotplots sometimes being misinterpreted is that students do not understand where the measured values are depicted in

stacked
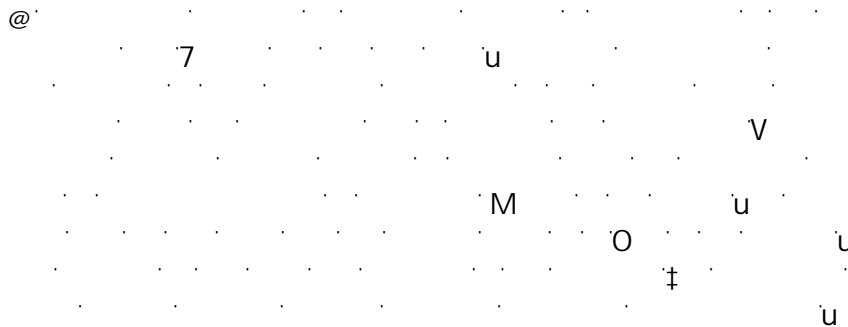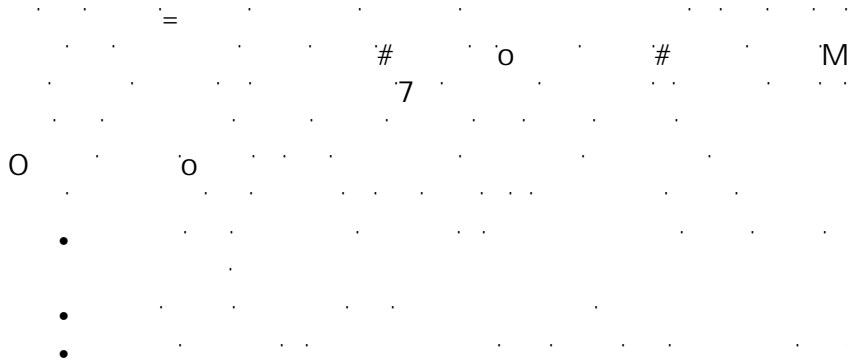
In what way do Grades 10–12 pre-university track students' histogram interpretations change after solving dotplot items?

Figure 5.4 -

Weight packages postal worker Frans

Weight packages postal worker Angela

Weight (kg)

Weight (kg)

Note.

5.2.2 Review of literature on eye-tracking in education

In this section, we review what is already known from gaze data in education and what measures are most suitable for our aim. In addition, we elaborate on how gaze data can be connected to students' strategies. We end each section with a sub-question.

Use of spatial gaze measures to reveal students' strategies for interpreting histograms

The use of gaze data for studying learning is not new (e.g., Strohmaier et al., 2020). For example, Garcia Moreno-Esteva et al. (2018) and Khalil (2005) studied students' visual cognitive behaviors on statistical graphs. A main advantage of eye-tracking "is that it can provide detailed information about the time-course of processing" (Kaakinen, 2021, p. 170). Most studies neglect this level of detail by using gaze data measures that are temporal (e.g., total fixation duration, reaction times), count (fixation count, number of saccades be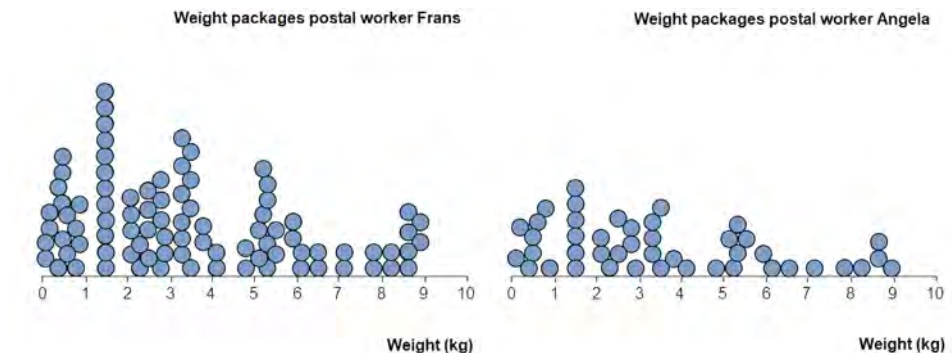tween relevant or irrelevant parts of the stimuli), or both (e.g., Kaakinen, 2021; Lai et al., 2013). Traditional time measures, for example, can hide visual scanning patterns (Goldberg & Helfman, 2010). A similar argumentation can be made for count measures such as percent of fixations on specific parts of the screen (Godau et al., 2014).

Spatial measures, such as a sequence of Areas of Interest (AOIs, e.g., Garcia Moreno-Esteva et al., 2018, 2020) can disclose the kind of detailed information Kaakinen (2021) refers to. Spatial measures, such as scanpaths, seem better suited for providing detailed information about students' thinking (Hyönä, 2010). Dewhurst et al. (2018) were one of the first who studied (simplified) scanpaths using vectors in (scene) viewing tasks. Their vectors include direction and magnitude of saccades.

In a previous study, we qualitatively analyzed students' scanpath patterns (sequence of fixations and saccades) when students estimated the mean from histograms (Boels et al., 2019a). After qualitatively coding 300 videos with students' gazes and verbal reports of 25 students in that study, we found that the *perceptual form* of students' scanpath patterns within *one* AOI—the graph area—was most relevant for students' task-specific strategies on these items, see Figure 5.5 (Chapter 3). This perceptual form can be captured by the direction (angle) and magnitude (length) of students' saccades.

In that study, we found several scanpath patterns that were indicative of students' task-specific strategies. All patterns were found on the graph area only. In one pattern, the *perceptual form* of that pattern was identified as vertical if successive saccades on the graph area were vertical and roughly aligned with each other (Figure 5.5). This vertical scanpath pattern indicates

that this student (correctly) tried to find the balancing point of the graph as an estimation of the mean. Another scanpath pattern was a horizontal gaze pattern indicating that this student (incorrectly) tried to make all bars equally high which results in the mean of the frequencies instead of mean weight. In total, five different scanpath patterns were found for students estimating and comparing means of histograms, each related to a specific strategy (Chapter 3). Other AOIs did not emerge as relevant to these students' task-specific strategies.

Figure 5.5 Example of a vertical scanpath on Item20
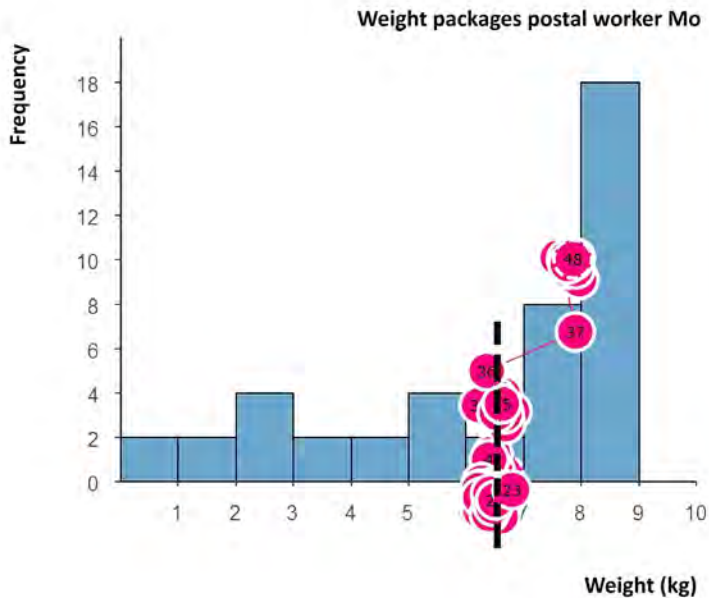
Note. Circles indicate fixations (positions on the screen where students look longer), thin lines between the circles indicate saccades (fast transitions between two fixations). A vertical line segment—indicating a scanpath—is superimposed for the reader's convenience. A scanpath is a sequence of fixations and saccades. "A fixation is a period of time during which a specific part of [the computer screen] is looked at and thereby projected to a relatively constant location on the retina. This is operationalized as a relatively still gaze position in the eye-tracker signal implemented using the [Tobii] algorithm." (Hessels et al., 2018, p. 22). (The figure has been translated into English.)

As the overall research question for this study indicates, we want to explore in what way secondary school students learn from dotplot items. Given that the scanpath patterns on the graph area indicate students' strategies, we examine differences in these patterns on histogram items before and after students solved items with dotplots. We only address the main differences, those being

differences relevant to students' task-specific strategies. The first sub-question for the present study is, therefore:

> *What are the main differences in students' gaze patterns on histogram items before and after solving dotplot items?*

Connecting gaze data to students' strategies

Although scanpaths can reveal students' strategies on a detailed level, there is no simple relation between eye movements and strategies (e.g., Orquin & Holmqvist, 2017; Russo, 2010) as not every eye movement is part of a task-specific strategy (e.g., Schindler & Lilienthal, 2019). Therefore, it is often needed to also ask at least some students what approach they took to solve the items.

Instead of concurrent think-aloud protocols, recalls (retrospective reports) are preferred for complex items (e.g., Van Gog et al., 2005) as concurrent thinking aloud may influence both eye movements and students' thinking (Van Gog & Jarodzka, 2013). The disadvantage of such retrospective think-aloud reports, however, is that students may have forgotten their strategy after completing all items. This risk can be reduced by having students look back at their eye movements (e.g., Guan et al., 2006; Kragten et al., 2015; Van Gog et al., 2005). Therefore, in the stimulated recall (cued retrospective reports), we individually cued each student with their own gazes. How we did that, is explained in the data collection section. The second sub-question for the present study is:

> *What indications can be found in students' verbalizations during stimulated recall that changes in their approaches to histograms occurred?*

5.2.3 Learning from a series of items: the practice effect

Students learning from a sequence of tasks is known as the practice, test-retest, or retesting effect in assessment theories (e.g., Heilbronner et al., 2010; Scharfen et al., 2018). The practice effect refers to improved performance (often scores or answers) due to repeated assessment with the same or similar, equally difficult items. The time interval between two assessments can be very short—5 or 10 minutes—to find such an effect (e.g., Catron, 1978; Falleti et al., 2006). The practice effect was found for several general cognitive function assessments for items that required memorization (e.g., of numbers), change detection (e.g., of changed colors between two items), and matching (e.g., what parts of items are alike). In addition, familiarity with test requirements can cause differences between the test and retest results (e.g., Falleti et al., 2006) and reduce anxiety (e.g., Catron, 1978). Furthermore,

regression to the mean can cause extreme results—high and low performance scores—to come closer to the mean, resulting in both under- and overestimation of improvement (e.g., Temkin et al., 1999). For achievement or knowledge tests, such as formative assessments in secondary education, the practice effect is also associated with *actual* or *true learning* as opposed to most cognitive tests, for example, IQ tests, for which learning is unlikely to occur (e.g., Lievens et al., 2007; Scharfen et al., 2018). Lumsden suggested that the practice effect can also be found *within* a sequence of items (1976). In addition, gaze data have been used to examine the practice effect (e.g., Guerra-Carrillo & Bunge, 2018; Płomecka et al., 2020). Although this is not the focus of our study, to the best of our knowledge, our study is the first that looks at a within-a-sequence-of-items practice effect.

Most research investigating the practice effect uses scores on standardized tests (e.g., in this meta-analysis: Hinton-Bayre, 2010). However, standardized tests often lack instructional relevance (e.g., Hohn, 1992). Practitioners, such as mathematics teachers, are more interested in knowing whether students learn from a low-stake sequence of items. Moreover, teachers are interested in students' strategies, hence "gaining qualitative insight into student understanding" (Bennett, 2011, p. 6). In this study, we, therefore, examine students' changes in strategies during solving items as an indication for potential learning. To exclude several other possible influencing factors—such as peers' or teachers' interventions—we use items from *one* sequence of items with statistical graphs.

For some items, students verbally reported their answer (estimation of the arithmetic mean), while for other items, they chose one of three answer options (comparing means, Figure 5.4). If a change in students' strategies occurred, toward a correct instead of an incorrect strategy, we would expect a difference in students' answers, including answer correctness. Therefore, the third sub-question for this research is:

> *What are the differences in students' answers on histogram items before and after solving dotplot items?*

5.2.4 Rationale for Using a Machine Learning Algorithm

For very small data sets or very short sequences of tasks, the first sub research question could theoretically be answered through the careful, manual study of gaze data. Our study, however, seeks to use machine learning to both augment the effectiveness of identifying differences in students' gaze patterns between items and to identify these differences at a scale that would be impractical to do by hand. To build an analytical model of the gazes, a non-ML-approach could be used. The ones we tried (e.g., logistic regression) performed relatively

poorly (see also Lyford & Boels, 2022). Instead, we use supervised learning, a subset of MLAs that use training data and pattern recognition to predict a well-defined output (Friedman et al., 2001). In particular, the present study makes use of random forests algorithm (Breiman, 2001), which will allow us to effectively and efficiently identify systemic differences in gazes between our two hundred student-item pairings. These random forests can not only be efficiently trained and used to identify patterns in students' gaze data, but they are likely to identify systematic differences in gaze data that are unnoticeable upon manual inspection (James et al., 2013). In addition, through assessing the importance of specific variables (Figure 15) random forests allow for some interpretability so that researchers can better understand what some of the differences in gazes might be (e.g., proportionally more vertical instead of more horizontal gazes could indicate a change from an incorrect to a correct strategy), and postulate about possible mechanisms.

5.3   Materials and methods

Details on participants as well as details on the eye-tracking method and two items (Item02 and Item11) were previously reported in a qualitative study (Chapter 3). Two items were previously used in a machine learning analysis (Item02 and Item20; Chapter 4) but with a different aim, namely, to examine how a machine learning algorithm (MLA) could identify students that used a correct or incorrect strategy—for solving the item—purely based on their gaze data on the graph area of this item. For the reader's convenience, we summarize here all information relevant to the present study.

5.3.1   Participants: pre-university track students Grades 10–12

Participants were 50 Grades 10–12 pre-university track students from a Dutch public secondary school [15–19 years old; mean = 16.31 years]; 23 males, 27 females (more details in Table 5.1). In the Netherlands, secondary school students are in a pre-vocational, pre-college, or pre-university track. Generally speaking, pre-university track implies mostly high-performing students. All participants had statistics in their mathematics curriculum. Each student individually solved the items in a separate room in their own school. Participation was voluntary; permission from the Utrecht University ethical committee was obtained, and informed consent was signed. Participants received a small gift for their participation.

Table 5.1 8 \\

#

8	V		V	

y
u

y
u

Note.)

@ V

h

5.3.2 Materials: histogram and dotplot items requiring comparing and estimating means

Estimating and comparing arithmetic means reveals students' knowledge
u

8

‡

u

U k

\ )

7

@

8 M h

‡

7

before *after*
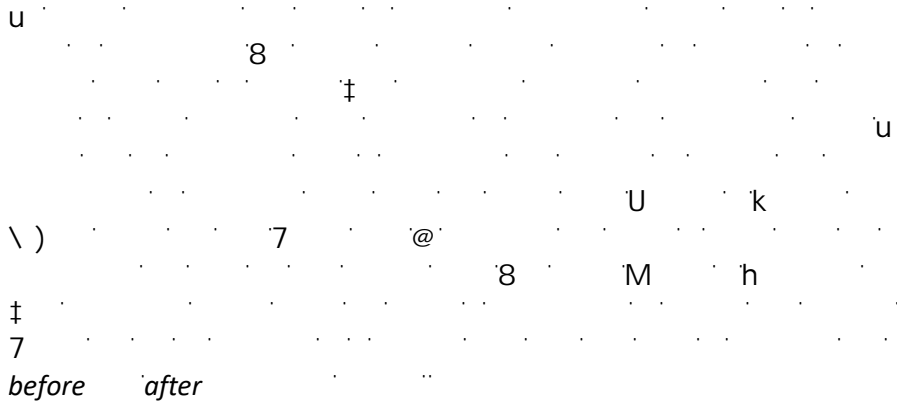
Four histogram items—dotplots items in between
@

‡

7

u

V *after* *before*

u *before* @ *after* @

‡

u

‡

U u *before* @ *after*
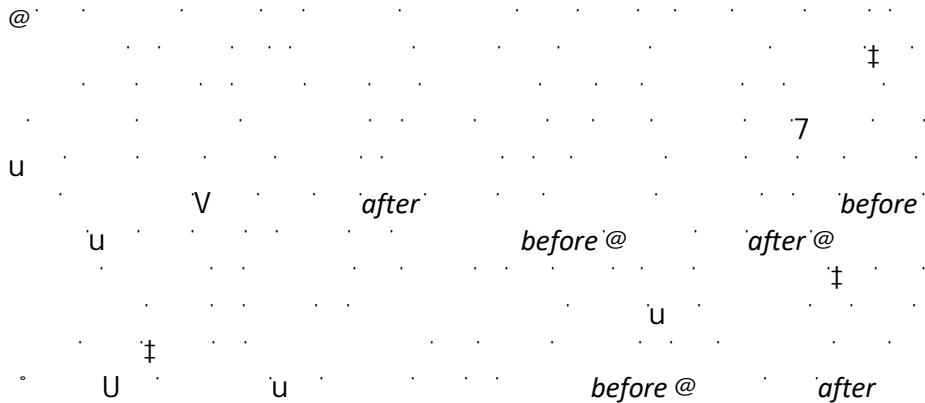
Item21—require students to compare the mean of the data in two histograms. We will henceforth refer to these as 'double-histogram' items. The question for both items was: Which postal worker delivers the heaviest packages on average? For each item, three answer options were given: (a) [Ellen/Elizabeth] delivers the heaviest packages on average, (b) [Titia/Monsif] delivers the heaviest packages on average, and (c) The mean weights for both are approximately the same. The correct answer for both items is (c).

Figure 5.6 Graphs of single-histogram items (left) and double-histogram items (middle and right) in the before (top row) and after (bottom row) versions
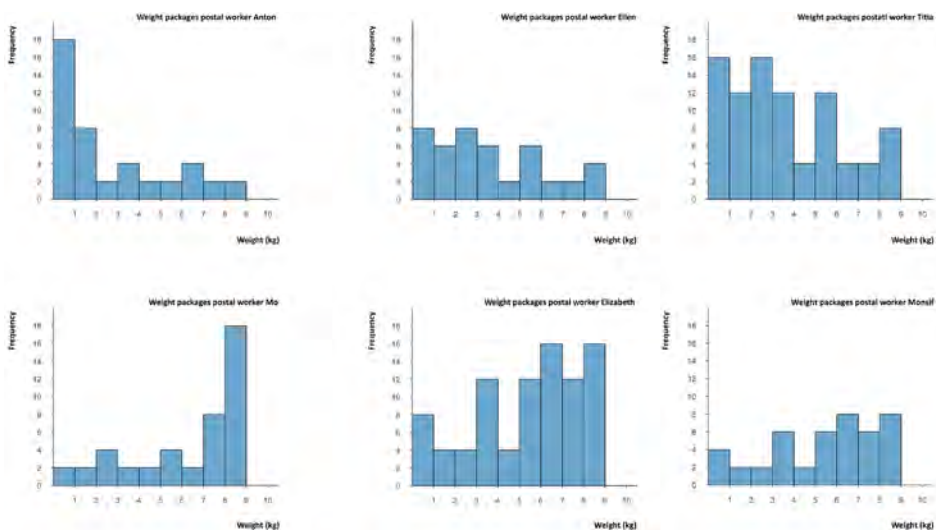
Note. Translated into English and numbering added. The numbering of the items (e.g., Item11) refers to the numbering in the original sequence of 25 digital items (Chapter 3). Each *after* item (bottom row) is a mirrored version of the *before* item (top row).

Six of the items between the items *before* and *after* were non-stacked (messy) dotplots that were specifically designed to scaffold students (Items13–18 from the original data collection, e.g., Figures 5.1, 5.2, and 5.4). As described in the Theoretical background section, we used dotplots to draw students' attention to specific features of the histograms that are important but might have been misunderstood.

5.3.3 Data collection methods: Eye-tracking, stimulated recall verbal reports

Data of a previous qualitative study is re-used for this study (Chapter 3). Data collection included students' answers on each item, *x*- and *y*-coordinates of gaze data on the items through an eye-tracker, and stimulated recall verbal reports. Collection of the gaze data and stimulated recall are shortly described

Data collection with an eye-tracker

ᵒ u ⟨𝒦ⓐ = =h h "
=
7 h 8
u o u u
‡
x- y
\ @ # 7
u ‡
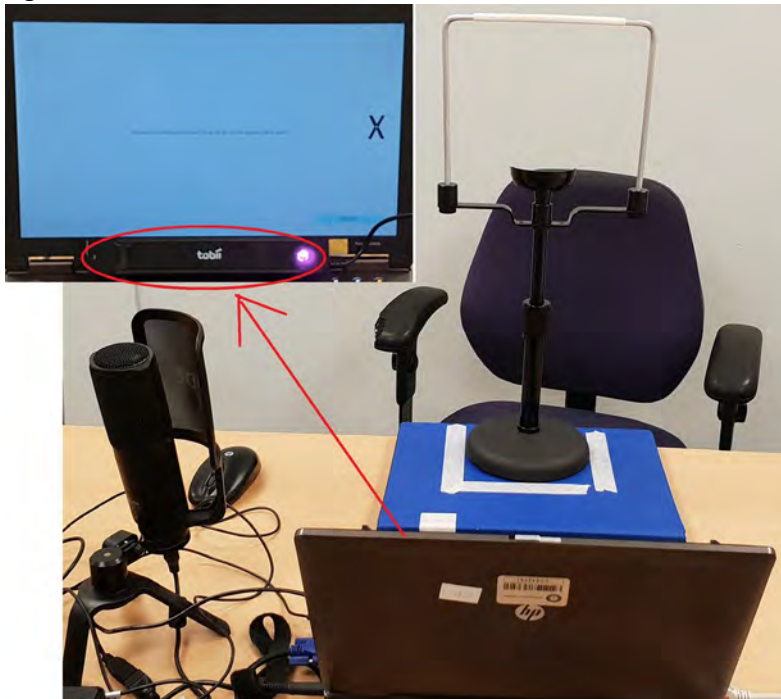
Figure 5.7 ο

Note. 7 u

Figure 5.8 AOIs of before Item11 (top) and Item02 (bottom)
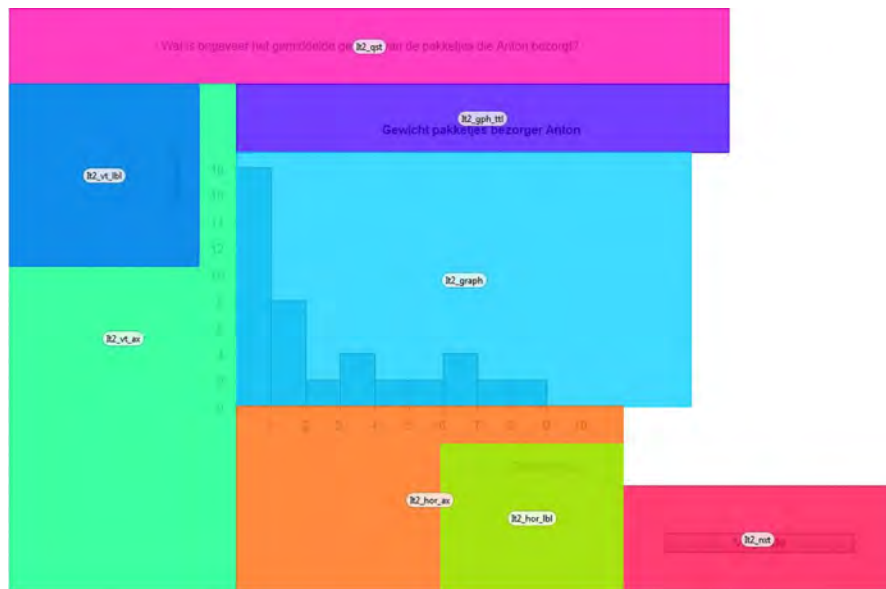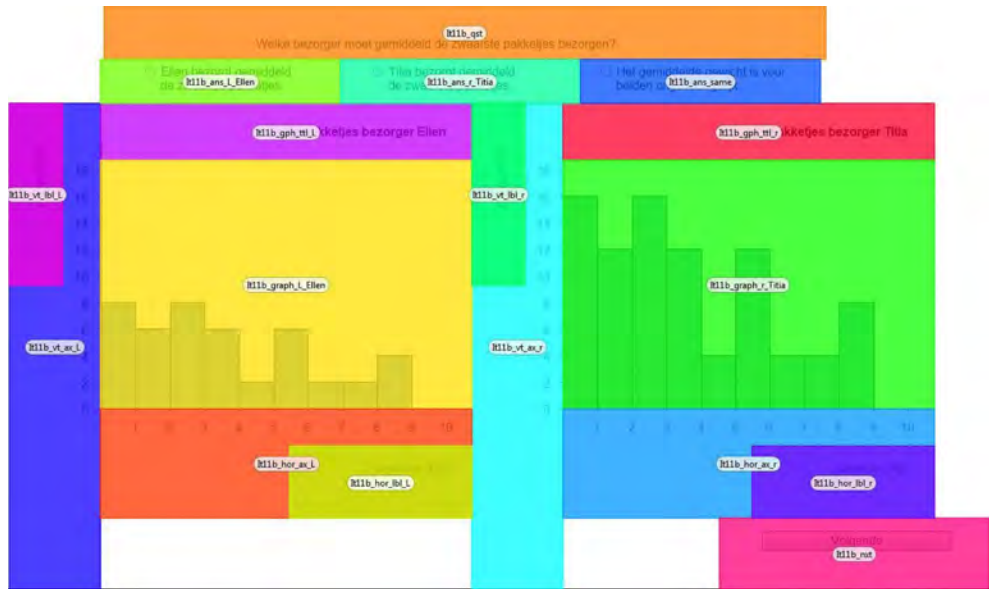
Note. Upper row: the graph area consists of the yellow and green areas named *It11b_graph_L_Ellen* and *It11b_graph_r_Titia*. Bottom row: the graph area is the light blue area named *It2_graph*.

No students were excluded from the data set, as the data loss per trial (averaged over all 50 participants) and the data loss per participant (averaged over all 25 items from the original dataset) were below the exclusion point (34% or more). The mean accuracy is 56.6 pixels (1.16°) with the highest accuracy on the most relevant part for our study: the graph area (middle of the screen; 13.4 pixels or 0.27°). The average precision (0.58°) is considered good. More details on accuracy, precision, and the eye-tracker can be found in Chapter 3 in line with advice from Holmqvist et al. (2023). The design of the complete sequence of items (25 items in total), including files used in the Tobii Studio software, as well as AOI sizes and output are available from a data repository.

Data collection through stimulated recall verbal reports

Stimulated recall (Lyle, 2003) is also known as "cued retrospective reporting" (Van Gog et al., 2005, p. 273). It is called retrospective "own-perspective video think-aloud with eye-tracking" (McIntyre, 2022, p.4) when used with a head-mounted eye-tracker. The first part of the verbal reports consisted of cued retrospective think-aloud. This means that students watched videos of their own gazes laid over the items, while they explained their thinking when they solved the items. This took place after students solved all items of the sequence of 25 items (Boels et al., 2023). During the second part of the verbal reports, clarifying questions were asked such as why they stated that their previously given answer was incorrect. In this second part, participants were also confronted with inconsistencies in their reports, such as differences between the answer given during recall and the answer during item solving. Time constraints influenced how many items could be questioned when students reported verbally. During this stimulated recall, we illuminated the location where students looked—through a kind of spotlight—and made the rest of the graph darker (see also Chapter 3). We preferred this method over having students look back at their fixations (e.g., red dots) for two reasons. First, it prevents students from making different eye movements when looking back—and describing the corresponding strategy—instead of the strategy they initially used. Second, this makes visible the exact information that the learner has looked at, instead of the information being covered by, for example, a red dot (the fixation; e.g., Jarodzka et al., 2013).

5.3.4 Data analysis through a machine learning algorithm

We used different methods for analyzing our data. For the first sub-question about differences in gaze data, we analyzed our data through a machine learning algorithm (MLA). For the second sub-question about changes in students' strategies, we coded transcripts of verbal reports (for the codebooks, see Chapter 3). For the third sub-question about students' answers, we

explored changes in answers and answer correctness. In the remainder of this section, we elaborate on the analysis with a machine learning algorithm. Studies usually only report on successful approaches. As a result, other researchers keep reinventing the wheel. For the first sub-question, we, therefore, decided to report both the MLA approaches we tried: our failed attempt to use time metrics as inputs for the MLA and a successful approach with spatial metrics.

Before applying a machine learning algorithm, we first wanted to get a better understanding of the underlying data. Therefore, first, we plotted a graph using the time metric total fixation time per AOI (also known as total dwell time or total fixation duration). Next, we used this same time metric as input for training our MLA. This approach failed to produce an accurate MLA. Moreover, although this time metric is commonly used, recent literature strongly advises against using total dwell time (Orquin & Holmqvist, 2017). Second, we examined saccade directions and magnitudes (spatial metrics). Finally, using these spatial metrics as inputs for our MLA was successful, which is in line with the results of previous studies (Chapters 3 and 4).

In the next section, we first describe how the MLA we used (random forest) works for those not familiar with MLAs and wanting to roughly understand these. Next, we describe how we applied the MLA in a failing approach using total dwell time and in a successful approach using saccade direction and magnitude.

Gaze-data analysis through a machine learning algorithm (random forest)
A machine learning algorithm (MLA) learns from input data without explicitly being programmed to use certain characteristics of the data. Supervised learning algorithms (see Figure 5.9) are a subset of machine learning algorithms whose training data contain known output values—in our case whether a student's gaze data belonged to a *before* or *after* item. Supervised MLAs are broadly used for pattern recognition and for making predictions (Friedman et al., 2001). Specifically, our work focuses on the use of random forests to identify whether student gaze patterns change substantially between similar items across our sequence of items.

Figure 5.9 u   U ơ

Gaze data (variables)

Training cycle

Ensemble of MLA
(random forest)

Item *before* or *after*

New gaze data (variables)

Identification cycle

Ensemble of MLA
(random forest)

Item *before* or *after*
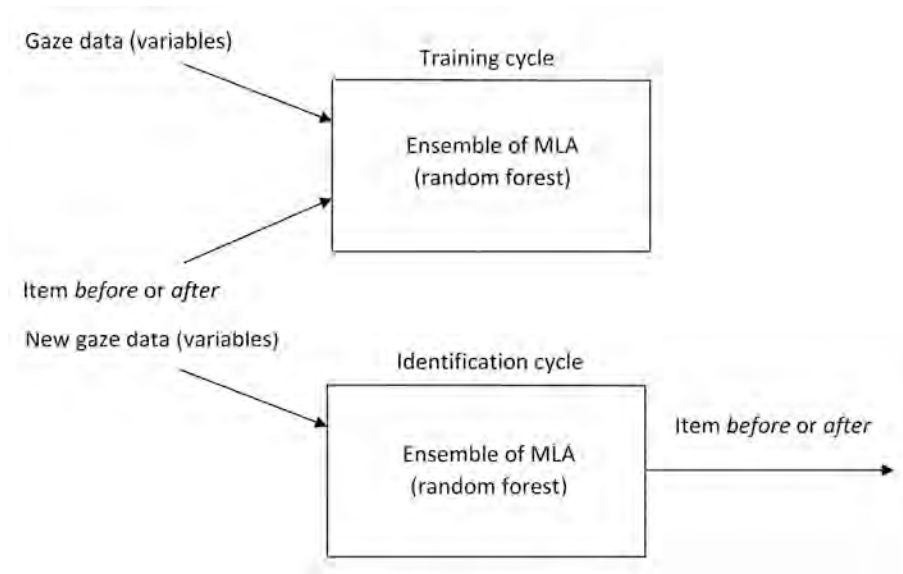
participants' data—with the possibility of sampling the same participant's data multiple times—and each split in the tree uses a small subset of the total number of variables. The exact size of each sample is part of the tuning process and our final values can be seen in the supplementary R-code.

Figure 5.10 Example of a decision tree. ESE means east-south-east direction of the saccade
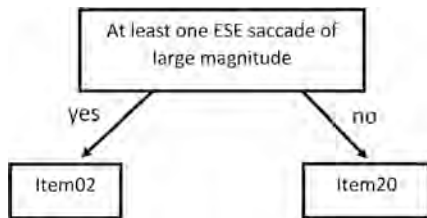
Allowing each tree to be built on only a subset of data and variables typically leads to a worse-performing tree than if all data were available (Dzeroski & Zenko, 2004). However, the risk of building one (the best) tree only, is that this tree might work perfectly for exactly the given data set, but not on data sets that are similar but slightly different. This is called overfitting. Building many trees using independently sampled data and variables stops any individual tree from drastically overfitting the training data and leads to trees that are relatively uncorrelated (Hansen & Salamon, 1990). These uncorrelated trees are then used together in an ensemble to make classifications, known as the random forest. Each tree makes a prediction about the class of the given data—in our case whether the user is seeing the item for the first or second time—and then the votes are totaled. Whichever class receives the most votes is the resulting classification of the random forest (Breiman, 2001).

This technique of simultaneously combining multiple machine learning algorithms—the trees in a random forest—together is known as ensemble learning. This approach is effective since the combined knowledge of many algorithms is often more accurate than any single algorithm (Dzeroski & Zenko, 2004). Here, we train our ensemble using gaze data as inputs and a binary output indicating whether the user is seeing the item for the first time or the second time. We use the *randomForest* package in R (Liaw & Wiener, 2002) to implement our random forest. Our final, fully-tuned model utilized the following hyperparameters: 1,000 trees, 5 variables considered at each split, a minimum node size of 1, and a maximum tree depth of 5.

We identified these optimal hyperparameters using a grid search of size 3^5 (we tried all combinations of three different values for each hyperparameter). Our nested resampling scheme utilized both an outer resampling and inner resampling of 5-fold cross validation. The reported best hyperparameters are the average values used across each of our outer resamplings. We likewise evaluated our model using 5-fold cross validation. To ensure no students' data were part of both the training and testing data when evaluating our model, we split the data into five groups, each group containing 10 students. We then used the 10 students' data (yielding a total of 20 student-item pairings) as testing data, and trained our random forest on the remaining 40 students' data (80 student-item pairings). This process was repeated five times until all students' data have been separately used as training and testing data. The software used for this data analysis is RStudio (RRID:SCR_000432). The full reproducible code to build our random forest as used in RStudio as well as the processed data are available through a data repository. The original data can be found in Boels et al. (2023).

A failed MLA—using dwell time on AOIs

As described in the beginning of this section, we first plotted the data before we analyzed it with an MLA. This plotting is considered part of the data analysis, as the plots can provide indications of what features might be relevant as inputs for the MLA. Differences in where and how long participants looked—fixated—were explored over time throughout each of the four items of interest. As an example, Figure 5.11 shows the fixations for two selected, archetypical participants, L01 and L05, who progressed very differently through the same item (here, the double-histogram Item11). The *x*-axis, time, has been rescaled from 0 to 1 so fixations could be compared between participants who spent different amounts of time on each item. A time of 0.5, for example, indicates the time at which the given participant is halfway through completing Item11. In this figure, points are jittered (shifted a slight amount in a random direction) to better display the density of points in a given AOI at a given time.

StudentL01, like many participants, fixated on several different AOIs throughout their time working on Item11, often moving back and forth between the graphing area and the corresponding axis. StudentL05, however, spent most of their time fixating on the graph area of both the left and right graphs—stopping briefly to look at the right graph's vertical axis and label after having spent a considerable amount of time looking at the graphing area. In addition to these two main archetypes, the remaining gaze patterns varied widely between each of the four items and between participants on a given item.

Figure 5.11 Distribution of gazes of participants L01(left) and L05 (right) over AOIs for Item11
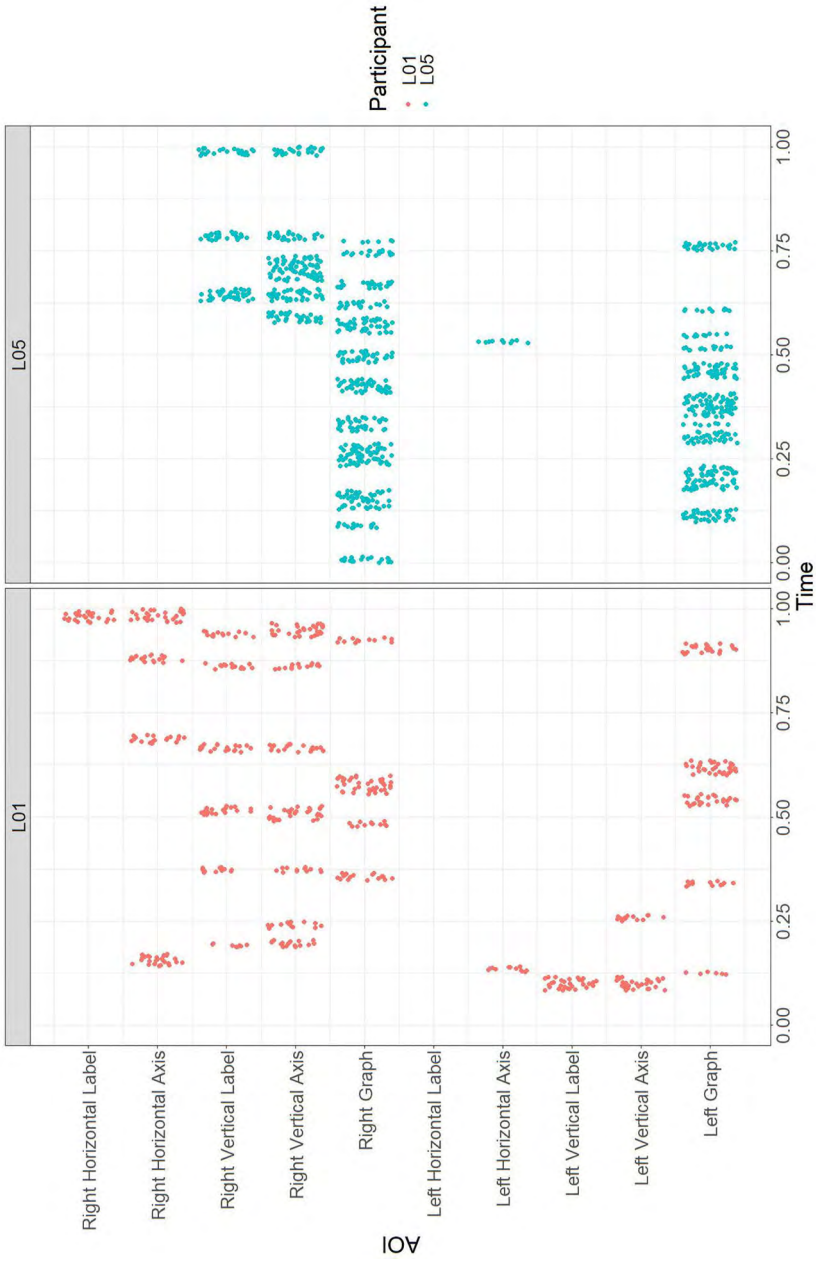
Note. The horizontal axis shows time, which is rescaled from 0 to 1 for each participant individually.

To quantify the differences between student approaches to the *before* and *after* items, we began by identifying features (variables) for training our series of random forest models. If the random forest algorithm can consistently differentiate between gaze data from the *before* and *after* item in each pairing, then some combination of features must exist that is more prevalent in one item when compared to the other, indicating a difference in gaze patterns between the paired items.

For each of the two item pairings (pairing Item02 and Item20, pairing Item11 and Item21), we began by treating each participant-item combination as our unit of observation, yielding a total of 200 data points (50 participants' gaze patterns across two items in each of two pairings). For each data point, we calculated the *proportional* time spent in each AOI, we identified the path each participant took through the AOIs, and converted this information into features for our random forest model.

In short, this initial approach was unsuccessful. To prevent readers from scrolling back and forth, we provide a short description of these results here. There were no discernable differences at the individual level between each pairing of *before* and *after* items. Participants spent roughly the same proportion of time looking at each of the AOIs when they saw the *before* items as when they saw the corresponding *after* item. Though the order in which participants progressed through each of the AOIs differed between *before* and *after* items, no discernable pattern emerged, and the correspondingly trained random forest algorithms were unable to accurately predict whether a participant was viewing a *before* item or an *after* item in a given pairing. We, therefore, do not further elaborate on this approach in the Results section.

A successful approach—exploring saccade direction and magnitude
Based on previous qualitative work (see Chapter 3), we then used directional movements—saccades—first by, again, visually investigating whether differences appeared in saccade patterns between *before* items and *after* items. We noticed a clear difference in the pattern of saccades due to the mirrored orientation of the otherwise-identical graphs in Item11 and Item21. Thus, our subsequent analysis focused on mirrored versions of the *after* items, Item21 and Item20, so that the graph area is made identical to their *before* counterparts. In other words, we took the gaze coordinates for the mirrored *after* items and adjusted them to match the corresponding coordinate of the unmirrored *before* items. Without this un-mirroring, the random forest algorithm could have easily differentiated between gaze data from the *before* and *after* items in each pair. Figure 5.12 shows the patterns of saccades for the same two selected archetypical participants—L01 and L05—on one particular pairing, Item11 and mirrored-Item21. In this figure, all saccades are centered

at the origin and radiate outward based on the direction and magnitude of the saccade. Only saccades of magnitudes greater than 200 pixels are shown since these are the saccades used in our final model. Saccades of less than 200 pixels were generally eye movements that are not indicative of students moving from one fixation point to another.

Given the size of the graph areas (width 489 pixels, height 313 pixels for double graph items, 600 x 335 for single graph items), the maximum possible saccade magnitude is 687 pixels (diagonal). The maximum speed for human saccades is approximately 700 degrees per second (Fuchs, 1967; Oohira et al., 1981) which is 570 pixels per 16.7 ms (a sampling rate of 60 Hz equals to one sample every 16.7 ms). Therefore, the maximum for long saccades was set at 600 pixels; longer saccades were considered to be artifacts. Furthermore, we removed 'saccades' smaller than 50 pixels. Given the accuracy of the eye-tracker (mean 13.4 pixels in the center of the screen and mean 56.6 pixels over all measures), we consider these small 'saccades' part of fixations or noise. Although this meant removing more than ninety percent of the measurements on the graph area, the accuracy of our MLA became slightly better.

We defined the beginning of a saccade to be a movement with a velocity greater than 50 pixels per 16.7ms. We defined the end of the saccade as any two consecutive 16.7ms windows where the participant's gaze had not moved more than 50 pixels. Points of fixation were determined by averaging the x- and y-pixel values of gazes in between two saccades, and each saccade's direction and magnitude were calculated between these points of fixation.

Figure 5.13 shows all saccades of a magnitude of 200 pixels or more for each of the four items. There is a discernable difference in the number of vertically oriented saccades between *before* and *after* items, especially in the Item02 and Item20 pairing. The Item11 and Item21 pairing also shows differences in the orientations of many horizontally facing saccades. Notably, there are several more northwest- and southeast-facing saccades in Item11 and more northeast and southwest-facing saccades in Item21 (after mirroring).

Figure 5.12
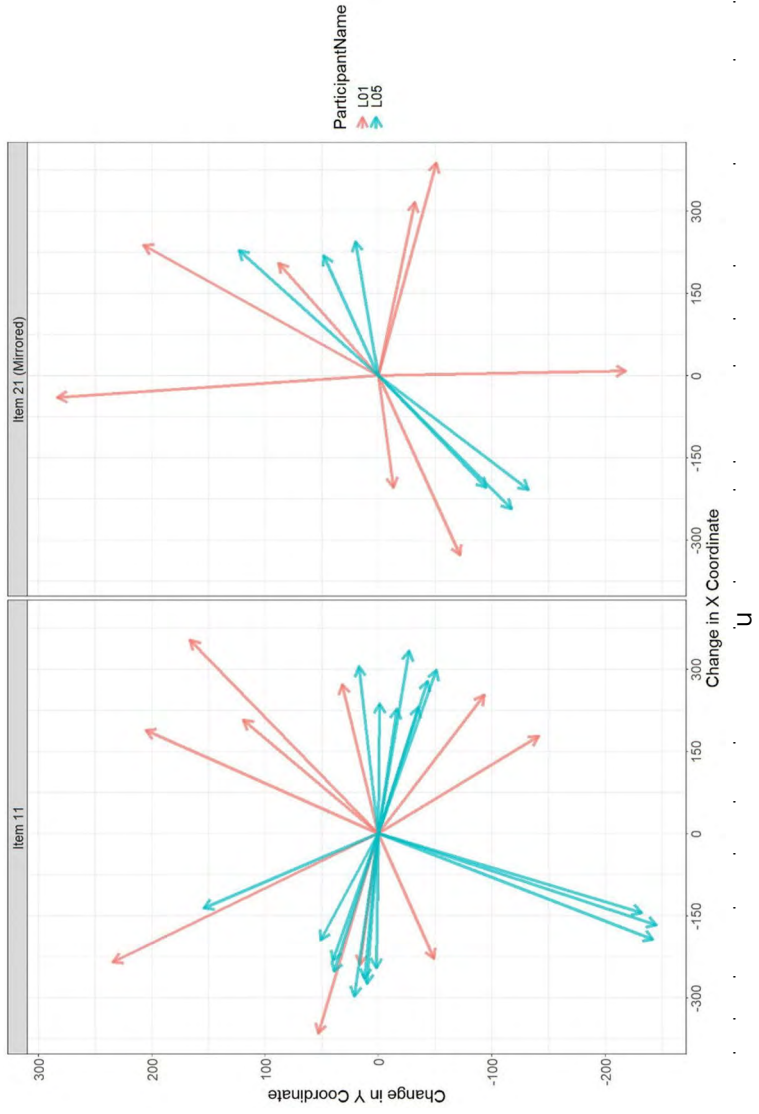
Note.

Figure 5.13 Saccades of magnitude 200 pixels or more of all participants on Item11 and Item21 (double-histograms, top) as well as Item02 and Item20 (single-histogram, bottom)
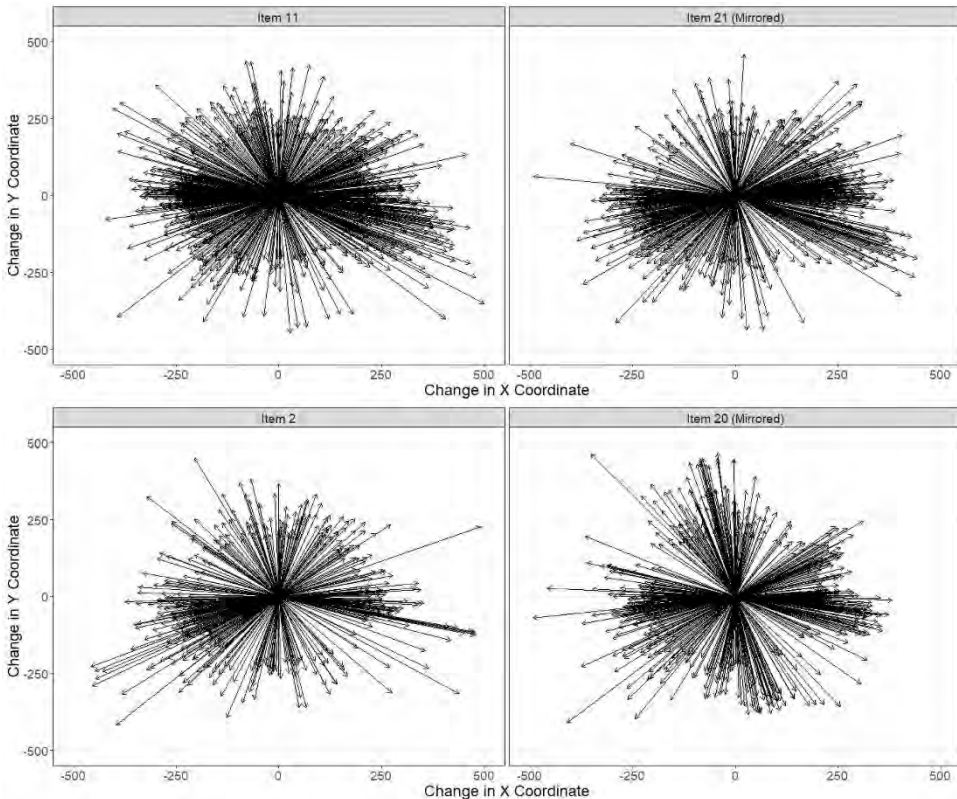
Note. Notice the difference in the density of students' saccade directions between the *before* items (left) and *after* items (right).

We examine differences in students' gaze patterns on histogram items. If the random forest algorithm can consistently differentiate between gaze data from the **before** and **after** **item in each pairing**, then there must be some combination of features (variables) that is more prevalent in one item when compared to the other, indicating a difference in gaze patterns between the paired items. To construct our random forest model, we tried different sets of features—placing each saccade into mutually exclusive bins depending on the direction, magnitude, and phase of the saccade, regardless of the point of origin. We tested two different directional schemes, two magnitude schemes, and three phase adjustment schemes, yielding a total of twelve combinations. Here, a phase adjustment is the angle (in radians) at which the direction bins are shifted, where 0 radians is equivalent to 0 degrees in mathematics—a saccade pointed eastward—and pi radians is equivalent to 180 degrees—a saccade pointed westward. Table 2 shows the details of each scheme. The

7 　　　 u 　　　 -o-

Table 5.2 "　　　　　　　　　　　　　　　 U σ

7	V	#	#
)			V- V‡ o‡ o-
)			7
			-V- VV- VV‡ ‡ V‡
			‡ o‡ oo‡ oo- -o-
U			U
U			†
h °			
h °			
h °			

Note. 7 　　　　　　　　　　　　　　　　　　　　　　@
× ×

Figure 5.14 #

N
NW　　　NE
W　　　E
SW　　　SE
S

‡ 　　　　　　　　　　　　　　　　　　　　　　 ‡
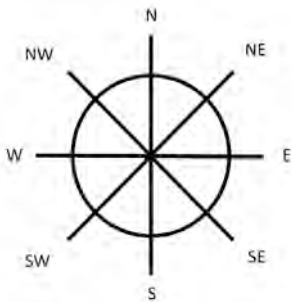
before

before　　　*after*

o

before　　　　　　*before*

o 　　　　　*after*

identified as belonging to the *after* item. Each metric was calculated using 5-fold cross validation.

We categorized the features into bins (see Table 2) for two reasons. First, we wanted to extract the variable importance metrics from our random forest in a way that might better inform us *why* the model was differentiating so well. More specifically, which direction or magnitude of saccade is more present in one item's data and not the other's. If we would have used continuous features, this interpretation would have been much more convoluted to human beings. Second, we also tried a continuous features model. That model performed worse. This might be because saccades that are close in direction and magnitude are functionally identical. In other words, perhaps a saccade of 25 degrees and a saccade of 5 degrees both imply that a student is scanning from left to right, and the difference in angles is either an artifact of data error or a meaningless difference between fixation points.

5.4 Results

Our overall research question is: In what way do secondary school students' histogram interpretations change after solving dotplot items? In this section we answer this question through answering the following three sub-questions:

1) *What are the main differences in students' gaze patterns on histogram items before and after solving dotplot items?*
2) *What indications can be found in students' verbalizations during stimulated recall that changes in their approaches to histograms occurred?*
3) *What are the differences in students' answers on histogram items before and after solving dotplot items?*

5.4.1 Main changes in students' gaze patterns on histograms

The first sub-question is answered by using a random forest model. The twelve combinations of direction, magnitude, and phase schemes (Table 5.2) yielded accuracies, sensitivities, and specificities that varied between 55% and 88% (Table 5.3). The standard deviations for each performance metric are reported in parenthesis using 100 resamples. The most accurate combination for both pairings was direction 2, magnitude 2, and phase adjustment 1, which corresponded to the most granular direction and magnitude bins and no phase adjustment. The details of each combination can be seen in Table 5.2. This best combination yielded a remarkably high 77% accuracy for the single-histogram items (Table 5.3) and 86% accuracy for the double-histogram items (Table 5.4). We note that accuracy alone can be potentially misleading—in our study attributing scanpath patterns randomly would yield 50% accuracy. Therefore,

Table 5.3

)	U	h	o)	o	o)	o	o)

Note o) o

Table 5.4

)	U	h	o)	o	o)	o	o)

Note o) o

Figure 5.15 Plots showing the importance of variables for single-histogram MLA-model (top) and double-histogram MLA-model (bottom)
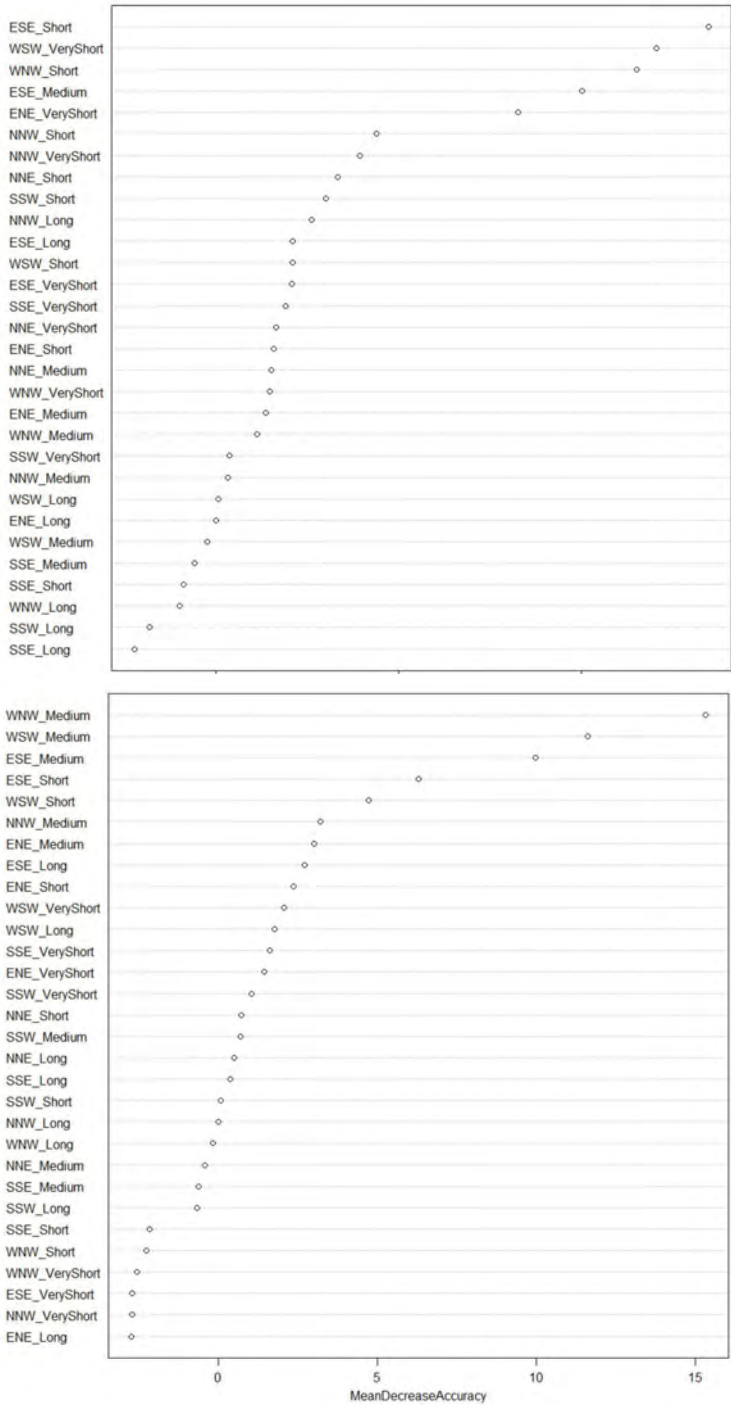
5.4.2 Students' post-activity verbal descriptions of approaches to histogram items

What indications can be found in the verbalizations during stimulated recall that changes in students' approaches to histograms occurred?

$18 + 8 + 5 \cdot 2 + 2 \cdot 4$

StudentL20:   Because there's a lot less than one kilogram, and relatively a lot [of] two kilograms. And then after that it really expands to nine kilograms but those are all very small numbers. So, then you end up with three.

Researcher1:   Yes. So now that you look at it again you think I should have given a completely different answer?

StudentL20:   Yes.

What this transcript also shows is that this understanding of how to estimate the mean from a histogram took place sometime after single-histogram Item02, but it is not clear when exactly this understanding occurred. For some students, it occurred after (at least one item of) the second series of histogram items, as the following student excerpt shows. This student reflects on the chosen approach in (left-skewed, single histogram) Item19 in the stimulated recall:

StudentL01:   The mean will be about between five and nine, because there are a lot of values [measured weights] there. And then around seven, because that's a little bit more to the left to zero from the middle between five and nine.

Researcher1:   Would you like to look at your eye movements again?
                [L01 looks back at eye movements]

Researcher1:   [...] You gave the answer ten. And that's where you looked.

StudentL01:   Ten? [sounds astonished]

Researcher1:   Yes, look at your eye movements again.
                [L01 looks back eye movements]

StudentL01   Oh yes, in that case I misread the axes.

For twenty-six students, there was (almost) no room for improvement, because they already gave answers within or close to the answer range during the before sequence of single-histogram items. Another four to twelve students seem to have learned specifically from the dotplot items. For example, studentL16 answered seven (instead of 2.7) for single-histogram Item02, but starts giving answers within or very close to the answer range for all following single dotplot items, and continues with these correct answers for the second series of single-histogram items after the dotplot items. During the recall, this student first describes a correct strategy for finding the mean from Item02 which is not in line with the given answer:

StudentL16:   Yes, again looking at frequency and weight, and then we see that the one occurred very often and the further you get [to the right] actually the less [frequency]. So, then the mean goes much more to the one than to the high numbers.

Researcher1:   Yes, you then said about seven.

StudentL16:   Yes, then I looked at it wrong again. Then I got weight and frequency flipped again.

5.4.3 Differences in students' answers to histogram items

At first glance, there seems to be no real difference in answer correctness between Item02 and Item20 (Table 5.5). Nevertheless, there are two indications in students' answers that students learned between Item02 and Item20. First, the answer range chosen for correct answers impacts answer correctness and was set the same for all items. In this study, students seem to prefer whole and half numbers. Enlarging the answer range to include the next whole or half numbers would result in (a non-significant) improvement in answer correctness (see note Table 5.5). Answer correctness is, therefore, quite sensitive to researchers' choices. Hence, changes in students' answers are a better indicator of students' learning potential.

Second, differences between students' answers and the actual mean are much lower for Item20 compared to Item02. We calculated the difference between the actual mean and the estimated mean (= M_{diff}). M_{diff} is, as expected, lower for the *after* item. We explored if this difference was significant through a one-tailed paired-*t*-test, as we expected that dotplot items would support students in correctly estimating the mean from histograms in the *after* items. The assumptions for a paired *t*-test, such as the unimodality and rough symmetry of the paired differences, were checked and met. The results for *after* Item20 (M_{diff} = 1.1, SD = 1.8) compared to *before* Item02 (M_{diff} = 0.4, SD = 1.7) indicate that it is possible that dotplots improve students' performance on the *after* item, $t(49)$= -1.7, p = 0.0469 < 0.05. We consider this (and the next) p-value significant in the way the statistician Fisher intended: "in the old-fashioned sense: worthy of a second look" (Nuzzo, 2014, pp. 150–151). The 95%-confidence interval for the differences in M_{diff} is <-Inf, -0.014] and Cohen's d measure for effect size is 0.40. Altogether, this points toward an improvement in answers. Note that, on the one hand, effect sizes tend to be larger in researcher-made tests compared to general (standardized) tests as well as in studies with small sample sizes. On the other hand, (very) short interventions often have lower effect sizes (e.g., Bakker, Cai, et al., 2019). Furthermore, although students' answers' *correctness* improved for the double-histogram item after the dotplot items, this improvement is not significant, as p = 0.3428 > 0.05 (e.g., McNemar, 1947).

Table 5.5 Answers given by the students, $N = 50$

Item	Actual mean	Answer range correct answers*	Average of given answers	Difference between students' answers – actual mean	Number of students correct	Percentage of students correct
Item02	**2.7**	1.6–3.8**	3.8	1.1	19	38%
Item20	**6.3**	5.2–7.4***	6.7	0.4	17	34%

Item	Answer options (number of students with this answer)			Percentage of students correct
Item11	Ellen (2)	Titia (30)	**Same (18)**	36%
Item21	Elizabeth (26)	Monsif (1)	**Same (23)**	45%

Note. *Experts were also asked to give an answer to these items. Based on these results as well as students' preference for whole numbers, the answer range was set to +/-1.1 for all items. **If answers 1.5 and 4 had been included, 27 students (54%) would have answered correctly. ***If the answers 5 and 7.5 had been included, 31 students (62%) would have answered correctly. Correct answers are in **bold**.

Instead of attributing the smaller M_{diff} —for single histogram items—to the solving of dotplot items, one alternative explanation is that the mean of Item20 (M = 6.3) compared to Item02 (M = 2.7) is closer to the mean of the frequencies (M = 4.9 for both). Nevertheless, we do not expect that this is the case, as for another left skewed item of this sequence of items (Item06, not further reported here, see Chapter 4) the mean of the frequencies (M = 7.1) was also close to the mean of this item (M = 5.7), but the difference between actual and students' mean (M_{diff} = 1.2) was similar to *before* Item02. To further exclude this alternative explanation, we suggest enlarging the difference between the actual mean and the mean of frequencies by adding more data (e.g., 50 packages) to the graphs. This number of added packages should not be too high to avoid students guessing from the size of the numbers what the weights are, as may have played a role in an item with SAT scores according to Kaplan et al. (2014).

5.5   Conclusions and discussion

In this study, we answer the main research question of in what way Grades 10–12 students' histogram interpretations change after solving dotplot items. More specifically, we look at students' estimations and comparisons of means from histograms. We expected that solving dotplot items would focus students' attention on the measured variable (weight) being depicted along the horizontal axis. In turn, that would invite students to estimate the mean of the weights (along the horizontal axis) instead of the mean of the frequencies (along the vertical axis) in the histograms. We examined three indications that taken together can suggest detailed-level changes in students' histograms interpretations: a change in students' gaze patterns, a shift in students' strategy for solving the histogram items, and an improvement in students' answers. If the changes are for the better, the relevance of knowing them is that they could underpin the learning potential of using dotplot items before solving histogram items—a hypothesis put forward by researchers in statistics education.

For the first indicator—a change in students' eye movements—we looked at differences in students' gaze or scanpath patterns on the graph area through a machine learning algorithm. Two main differences—on student level—between scanpath patterns on before and after items were found. First, there were proportionally less horizontal directions (ESE/WNW) in the gaze patterns on the after items than the before items. Second, proportionally more vertical directions (NNW/NNE) were found in the after items. A horizontal gaze pattern is associated with an incorrect strategy while a vertical gaze pattern is associated with a correct strategy. Our best implementations of random

forests were able to accurately classify (roughly 80% of the instances) whether it was the first (*before)* or second (*after)* time a participant had seen an item in one of two pairings. What we can attest to, is a significant, discernable difference in the way participants looked at the *before* items and when viewing the mirrored *after* versions, even after accounting for mirroring in the graphs. As we could not identify other confounding factors[53], it seems reasonable to conclude that our findings exhibit evidence that students changed the way they approached an item when seeing its mirrored version later in this sequence of items. The results of our MLA are in line with the results of previous studies (Chapters 3 and 4).

We cannot be certain whether the observed differences in gazes indicate a change in strategies. Although scanpaths can disclose students' strategies at a detailed level, the relationship between eye movements and strategies is task-dependent (e.g., Orquin & Holmqvist, 2017; Russo, 2010). In addition, not every eye movement is part of a task-specific strategy (e.g., Schindler & Lilienthal, 2019). Therefore, other data—such as our second and third indicators—are often needed to support or refute conjectures about the association between scanpath patterns and strategies.

The second indicator of changes in students' histograms interpretations—a shift in students' strategy for solving the items—was evaluated by coding students' stimulated recall verbal reports. The excerpts provide evidence that at least some students changed their strategies, from an incorrect approach for estimating and comparing means from histograms to a correct approach, during or after solving the dotplot items.

The third indicator—improvement in students' answers—was explored through both answer correctness, the difference between students' estimation of the mean and the actual mean for single-histogram items, and the changes in students' answers on the double-histogram multiple choice items. Answer correctness did not change significantly on either item type. Nevertheless, the difference in students' estimation of the mean compared to the actual mean was significantly smaller for the *after* item compared to the *before* item. We use 'significantly' here in the sense Fisher intended: worthy of further investigation. Data collection with new and more participants—from the same population (Dutch Grades 10–12 pre-university track students)—is needed to investigate the hypothesis that this difference becomes smaller, that there is a change in multiple choice answers, and that both are due to solving the dotplot items.

---

[53] Students solved similar items in the sequence of items preceding the *before* items. Therefore, we consider it less likely that solving a similar but mirrored item contributed to the change in gaze behavior, although we cannot rule this out.

The three indicators taken together suggest that at least some students changed their strategy during or after solving the sequence of dotplot items. A change in gaze behavior was observable through our machine learning analysis with random forests. Depending on how learning is defined, this change could point to a learning effect of solving dotplot items.

Interpreting the results, we abductively arrived at the following explanations for our results. First, the change toward proportionally more vertical gazes on the after items, is in line with the conjecture that the absence of a vertical scale in dotplots can turn students' attention toward the horizontal scale which is where the variable is presented in both histograms and dotplots. These students possibly figured out that the mean can be estimated from the measured values along the horizontal axis. However, we cannot rule out that factors other than solving dotplot items could have contributed to this change.

Second, we consider the most likely explanation for the mixed results that solving dotplot items promoted *readiness for learning* (Church & Goldin-Meadow, 1986) about histograms. Having students reflect on their previous strategy while they were cued with their own gazes during retrospective verbal reporting then might have given them new insights. After solving the dotplot items, histogram items seem to lie within the region of sensitivity for learning, hence within students' zone of proximal development (Vygotsky, 1978). It is possible that the questions asked by the researcher (an adult), which were intended to figure out how students solved the items, unintentionally stimulated students' thinking by asking them to explain—hence, reflect on—their strategies. Further research is needed to check this explanation. An alternative explanation for the results would be that other items after the second series of histogram items induced students' thinking. Although we cannot exclude this alternative, we regard this to be less likely.

Further discussing the results, we note that this study is novel in the following ways. First, to the best of our knowledge our study is the first in education that combined a quantitative analysis of the scanpath patterns found in spatial gaze data with insights from a previous qualitative study about what part of the scanpath pattern is relevant for students' strategies (namely, the scanpath on the graph area only). The use of qualitative insights contributes to the validity of the study while the quantitative approach through machine learning analysis contributes to the reliability of it. Most eye-tracking studies that use spatial measures investigate the sequence of AOIs (Garcia Moreno-Esteva et al., 2020) and the same holds true for those combining it with MLAs (e.g., Garcia Moreno-Esteva et al., 2018). Instead, we used vectors (i.e., direction and magnitude) of saccades. Studies in education that utilize vectors are rare (e.g., Dewhurst et al., 2018). Second, novel is the

use of an MLA for finding differences in gazes that are relevant for changes in students' task-specific strategies between tasks.

Our study has several limiting factors. First, many of the participants' gaze data contained data loss. Although data loss is normal due to blinking or looking away from the screen, some data loss could be avoided by pre-excluding participants who wear glasses, contact lenses, or mascara. In addition, an eye-tracker could be used that is better in catching gazes from people with epicanthic folds (almond eyes). As we aimed for a naturalistic setting, we did not exclude any of such participants. In addition, for some participants we had sparse data. Some of these participants spent only a few seconds looking at a given item. This made predictions and training more challenging. Most of these participants appeared to parse the graph and answer the corresponding question(s) in a rapid but reasonable manner, although one participant appeared to scan the graph and answer the question in such a rapid way that it is unlikely that they had time to fully understand what the graph was depicting. Since no participants' data were removed, it is likely that some amount of data cleaning and removal of outlier participants would increase the accuracy of our random forests, although our data collection scheme does not allow us to know with certainty why a certain participant's gaze data were sparse for a particular item.

A second limiting factor was that we restricted our final analysis to the graph area of each item, excluding AOIs such as the axes labels and the graph title. The inclusion of these AOIs yielded more noise and worse results, but further work might investigate the possibility of productively including them. Third, answer correctness and students' strategies correspond only to a limited extent. Finally, and most importantly, our sample size—50 participants and 2 items yielding 100 participant-item pairings—is relatively small for machine learning and statistical analysis. Our results indicate strong evidence of a change in gaze patterns between the *before* and *after* items, but more data are needed to generalize these findings appropriately.

A theoretical contribution of this study is that having students solve 'messy' dotplot items can create readiness for learning histograms. A reflection phase seems to be needed to make use of the knowledge obtained. We speculate that this partly explains the results from the literature on dotplots (e.g., Garfield & Ben-Zvi, 2008b; Lyford, 2017). Another reason for these results, we believe, is that only 'messy' dotplots contribute to students' understanding of where the measured value is. Stacked dotplots already contain an information reduction step (the binning) that could lead to similar misinterpretations as for histograms (see Chapter 2). We, therefore, advise investigating whether stacked dotplots need to be avoided in secondary education.

As a first methodological implication, our study shows how an MLA in combination with eye-tracking data can be used to reveal phenomena that are of interest to researchers of education. Future use could include interpreting graphical representations in biology, physics, economics, and geography. By choosing features (variables) that are relevant to the phenomena of interest (here: students' strategies for solving a histogram item) and meaningful to the researchers, an MLA can give insights into subtle, detailed-level differences in students' strategies that are hard to detect through other research methods, such as time measures in eye-tracking research or qualitative analysis of gaze data by researchers.

A second methodological implication is that there seems to be a practice effect *within* a sequence of items for at least some students, in line with suggestions from Lumsden (1976). A practice effect refers to improved performance after 'practicing' (i.e., repeatedly solving of similar or the same items). This is also important for judging the validity of summative assessment. More research is needed to confirm this within-a-test practice effect. Catron (1978) found an effect of item types on an IQ test. For example, he showed that development of a strategy in strategic items improved performance on retesting. The present study does not consider an effect of item type. Further research is needed to find out whether and for what items the order and type of items influence the within-a-test practice effect.

What are possible implications of our findings? The ML approach is generalizable to other sequences of items or any instance when a user may wish to classify eye-tracking data into one of many discrete categories. Further analysis is needed to correlate the number of saccades of specific directions and magnitudes with particular viewing strategies. In other words, does the presence of certain features (variables, such as horizontal or vertical saccades) indicate students taking a particular strategy, and if so, is this strategy more common when viewing a before item as opposed to an after item? Moreover, we think our ML approach can also be used when researchers want to know whether solving X (a question about a graph or image) changes the way students solve Y (a question about a different type of graph or image).

For testing a future hypothesis that students' estimations of the mean from histograms become closer to the actual mean after solving dotplot items, we suggest making a sequence of 24 items: eight histogram items (improved versions of the existing items from the original sequence with more data in them as well as one extra left skewed single histogram and one extra double histogram), eight dotplot items (all items containing the same data as the first eight histograms) and then again eight histogram items (all mirrored versions of the first eight histogram items). To check a future hypothesis that giving students stacked dotplots is a less effective way to scaffold them, a variant of

this design could be made with stacked dotplots only, instead of 'messy' dotplots (with the stacks in between two values on the horizontal scale; all stacked dotplots contain the same data as the first eight histograms). To check a future hypothesis that the reflection phase is important, a variant with and without stimulated recall verbal reports could be conducted followed by another series of histogram items. In all these variants, machine learning analysis can support this hypothesis testing.

For practitioners, insight into what students learn from doing a sequence of items is also relevant for homework and formative assessment, in particular, if no feedback is given—which is quite a common situation (e.g., when there is less student-teacher interaction). The observed differences in gaze patterns together with the other evidence in this study, suggest that a sequence of items can create readiness for learning, but a teacher may still be needed to ensure that students reach their full potential.
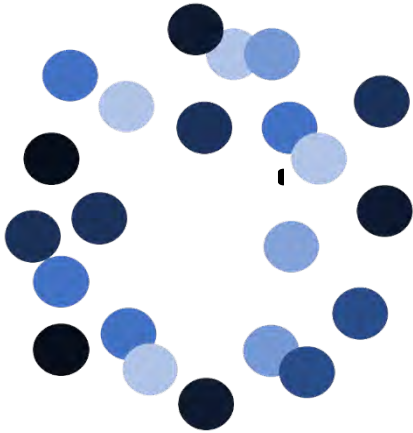
Understanding histograms in upper-secondary school: Embodied design of a learning trajectory

Understanding histograms in upper-secondary school: Embodied design of a learning trajectory

Abstract Many students persistently misinterpret histograms. We think this is partly due to a lack of embodied experiences. To answer the question of what sequence of tasks—designed from an embodied instrumentation perspective—can support students' understanding of histograms and the underlying key concepts, we created a hypothetical learning trajectory (HLT). We used five design guidelines based on our theoretical framework: (1) identify the actions that could have constituted the target artifact (histogram), (2) design motor-control or perception tasks to which these actions are the answer, (3) have students (digitally) perform these actions with feedback, (4) stimulate reflection on actions, (5) create possibilities for transfer of actions by varying contexts and environments. The main steps in the HLT are experiencing a lack of understanding, reinventing the role of both axes and arithmetic means in histograms and transfer. Our multiple-case study with five 10–12th graders suggests that most conjectures of the HLT were met, but transfer could be improved. Students' gestures indicated using actions from previous tasks to solve current tasks. The results suggest that embodied experiences with reflection contributed to overcoming some well-known misinterpretations. Overall, we show how students can be guided to reinvent more complicated mathematical artifacts from actions with simpler ones.

6.1   Introduction

What sequence of tasks designed from an embodied instrumentation perspective can support students' understanding of histograms and the underlying key concepts?

6.2 Review of the literature on histograms

6.2.1 Misinterpreting histograms

An extensive literature review revealed that many students persistently misinterpret histograms (Chapter 2). Misinterpretations included that Grades 7–12 students incorrectly interpreted the height of bars as people's heights (Bakker, 2004b) and calculated the mean from histograms by dividing the sum of frequencies by the numbers of bars (Ismail & Chan, 2015). The review study revealed two of the statistical key concepts that underlie students' difficulties when interpreting histograms: data and distribution. Much research focuses on specific misinterpretations related to distribution (e.g., center, variability, shape). Nevertheless, several of these misinterpretations may originate in the much less studied concept of data.

The concept of data encompasses "the need for data; how data represent characteristics or values in the real world; how data are obtained; different types of data, such as numbers, words, and so forth" (Garfield & Ben-Zvi, 2004, p. 401). Gould phrases this as "understanding who collects data about us, why they collect it, how they collect it", "understanding how representations [of data] in computers can vary and why data must sometimes be altered before analysis" (Gould, 2017, p. 22). It includes 'data moves' which is merging data, constructing new data based on existing data, and so on (Erickson et al., 2019), and the difference between variable (e.g., weight) and data (e.g., numbers representing the measured weights). For graphical representations, this concept of data encompasses how data are represented in, for example, histograms, boxplots, case-value plots, and along what axis the measured variable is represented (Chapter 2). This is a broader concept of data than that found in the GAISE II guidelines (Bargagliotti et al., 2020), in which data is used as 'raw data' and didactical choices needed to be made.

We focus on three aspects of the key concept of data in histograms that many students tend to misinterpret:

- What the data are. The number of bars is incorrectly seen as the number of cases N (e.g., Ismail & Chan, 2015; Sorto, 2004).
- How many variables a histogram depicts. Some people incorrectly think histograms display two statistical variables (e.g., Cohen, 1996; Meletiou, 2000; Zaidan et al., 2012).
- What the measured values are. Frequency (depicted along the vertical axis) is incorrectly seen as the measured value (e.g., Bakker, 2004a). In addition, graphs without context (only bars) can be histograms or case-value plots and should be avoided (e.g., Cooper & Shore, 2010). Some contexts are associated with specific axes in graphs (e.g., body height

6.2.2 Revealing students' understanding of data in histograms through estimating the arithmetic mean from a graph

6.2.3 Lessons learned from our previous studies

hand border value). We also learned that messy dotplots could create readiness for learning (Chapter 5; Lyford, 2017).

6.2.4 Advice for the content of tasks

The following advice can be extracted for the content of tasks from the statistics education literature. Use realistic problems (e.g., Biehler, 1997). Start with graphs in which all measured values are visible, such as dotplots (delMas & Liu, 2005), and then coordinate them with histograms, for example, via an overlay (e.g., Bakker, 2004a). Have students sort histograms (Garfield & Ben-Zvi, 2008a). Develop students' conceptions of mean, spread, and variability informally and in context first (Garfield & Ben-Zvi, 2008a). Have students explore histograms containing small and large variation and remove or add outliers. Students first need to predict or estimate the mean before determining or calculating it (themselves or by technology). Other recommendations for the content of tasks are to work with small and large data sets (Garfield & Ben-Zvi, 2008a), give feedback (e.g., by confronting students with results), have students construct graphs themselves (e.g., from a table, Eshach & Schwartz, 2002), let students flexibly use multiple representations (e.g., Lem et al., 2013c) and have them estimate or predict (e.g., what the mean is) before feedback is given. In doing so, students must be "forced to record and then compare," as otherwise, they tend to see only confirmatory evidence in the results (Garfield & Ben-Zvi, 2008a, p. 41). Many studies advise using measures of central tendency (e.g., Gal, 1995). We address this separately in the next section.

6.3 Theoretical framework for the design

We think the persistence of students' difficulties with histograms is partly due to a lack of sensorimotor experiences. Vygotsky stated that mathematical thinking is grounded in such experiences (1926/1997):

> When thinking of something round […] we realize through the movements of our eye muscles the very same adaptive movements, the very same focusing on objects which we had once perceived in actuality. Even the most abstract thoughts of relations that are difficult to convey in the language of movements, like various mathematical formulas, […] even they are related ultimately to particular residues of former movements now reproduced anew. (p. 162)

Figure 6.1 Example of finding—through a logical-historical reconstruction— the actions that could have constituted the target artifact (e.g., histogram). Actions reify in artifacts; in this process, artifacts get crystallized. Through an action with an artifact, students acquire or "get" it
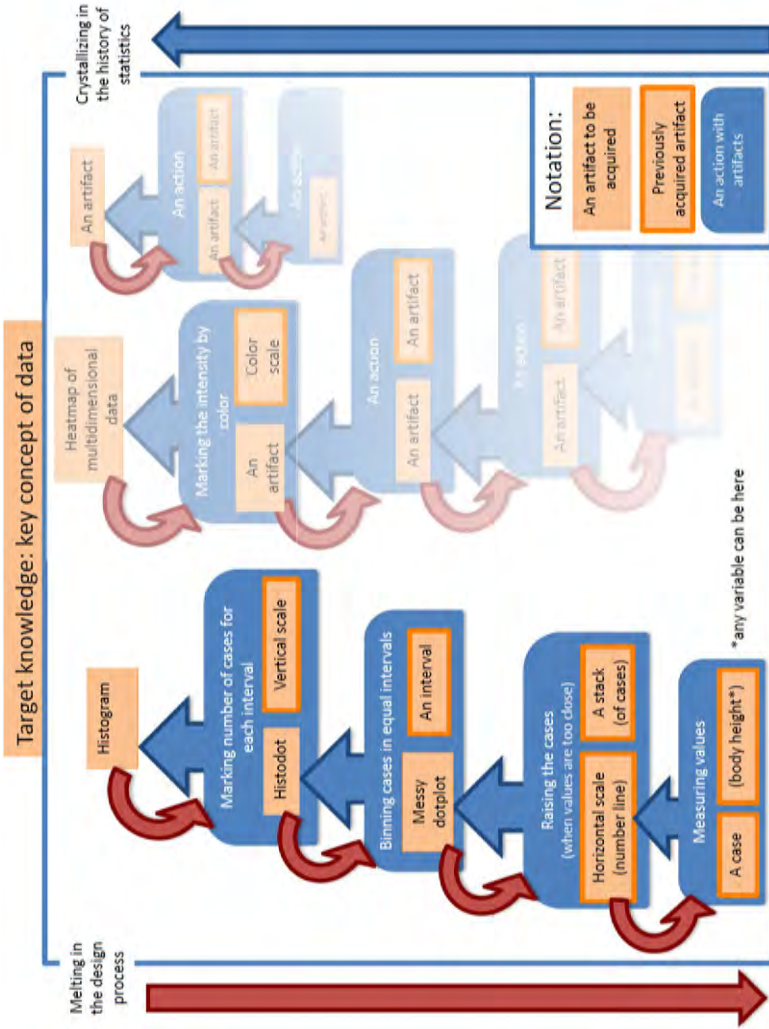
design

how
Identify the actions that could have constituted the target artifact

Design motor control or perception tasks to which these actions are the answer

applied this in one task by having students drag measured values to their correct position on a horizontal scale. (b) Create productive struggle (Abrahamson & Bakker, 2016; Kapur, 2014; Roth, 2019). For this, some ambiguity in task (formulation) might be needed (Foster, 2011). For example, we used a body height context which is associated with the vertical axis while these body heights are depicted along the horizontal axis. (c) Support students by reducing the complexity of the environment through limiting degrees of freedom (Bakker, Shvarts, & Abrahamson, 2019; Bernstein, 1940/1967), which makes tasks manageable and focuses the students' attention on the target knowledge. For example, in several tasks, students can change the height of a bar in a histogram—to see how this influences the mean—but not its position or bin width. Also, in the first tasks on the mean, we used histograms with only two bars. (d) Guide students to gradually reinvent mathematics through sensorimotor actions and/or perceptions. To accomplish this, in consecutive tasks, (mathematical and statistical) artifacts such as axis, numbers, bars or line segments are added that can guide students toward the mathematical and statistical discourse. Abrahamson et al. (2021, p. 168) call this "mathematical appropriation." For example, in several tasks, students drag a vertical line to their estimated mean. (e) Start with a macro learning problem that triggers intentionality within body-artifacts functional systems.

(3) *Have students (digitally) perform these actions with feedback.* This guideline contains two aspects: (a) have students perform the actions and (b) provide students with feedback. For persistent problems like students' conceptual difficulties with histograms, it is worth looking at who is doing the mathematical actions—the student or the software? In most statistical software, actions are hidden—for example, a histogram appearing directly when an option is clicked (e.g., InZight, Minitab[56], minitools) or displayed, such as moving dots in a dotplot to the correct bin (e.g., TinkerPlots, Fathom, CODAP[57]). Sometimes, both are available (e.g., VUstat, GeoGebra[58]). However, in embodied design, students initially perform the actions themselves (e.g., build a dotplot or histogram from data). Hence, the construction of target artifacts is not outsourced to the (digital) environment until students have reinvented and established each artifact themselves (Chase & Abrahamson, 2015). For instance, height of bars in a histogram is not outsourced until reinvented. Therefore, the *unit height* in our histogram overlay onto a dotplot in our first tasks is not equal to the *height of a dot*, and a *stacked* dotplot is avoided. We chose to give feedback after the students answered: a check box

---

[56] https://www.stat.auckland.ac.nz/~wild/iNZight/; https://www.minitab.com/en-us/

[57] http://tinkerplots.com/; https://Fathom.concord.org/; https://CODAP.concord.org/

[58] https://www.vustat.eu/; https://www.geogebra.org/

y

V

7

Stimulate reflection on actions. =

\

h

Create possibilities for transfer of actions by varying contexts and environments #   u
=

u   7
"   \
o   u
u
u   u

6.4   Method

@
@
U M   k   u

#

\

‡   #

)   )

‡

analyzed the HLT step by step. For each step of the HLT (Simon & Tzur, 2004), we analyzed each student's reasoning in relation to the conjectured learning and aligned this with extra information about each case. Furthermore, we merged the gained information about each HLT step by comparing and contrasting the cases—in line with the method of cross-cases analysis. Typical tasks for each step in the HLT are described in a section including results; an overview of the HLT is found in the data analysis section and some tasks are described in Appendix A of this chapter.

6.4.1 Participants

Participants were five, pre-university track students in Grades 10 and 12, see Table 6.1. They all took the Mathematics A course on applied analysis in economics and health contexts, and statistics (Daemen et al., 2020). Their mean self-reported mark for mathematics was 6.7 on a ten-point scale (10 is highest, 1 is lowest; 6.9 for females, 6.5 for males), indicating normal mathematical abilities. Participants were given a 30-euro fee for their participation. Approval from the Science-Geosciences Ethics Review Board was obtained under number Bèta S-21578, and written consent of participants and their legal representatives (if necessary) was obtained.

The participants' primary experiences were based on the most common textbooks that introduce histograms in Grade 9 after introducing stem plots and frequency tables. In Grades 10–12, students with Mathematics A re-encounter histograms. Textbooks sometimes confuse histograms and case-value plots and pay no attention to relevant differences. Students use a calculator for standard deviation, mean, median, first and third quartiles, and interquartile range and learn to read off values from histograms and boxplots and to draw these graphs. Comparing graphs (samples) is done through calculations in hypothesis testing only.

Table 6.1 Participant characteristics

Student	Age	Grade	Sex
S1	15	10	Female
S2	16	10	Female
S3	16	10	Male
S4	18	12	Male
S5	18	12	Female

6.4.2 Data collection

Students filled in a questionnaire on their characteristics and pre-knowledge. Students' discussions were audio and videotaped. The students' worksheets and grid papers were also collected.

6.4.3 Setting of the intervention

o   u   O   O uOO y
y   u   7

                u

                    u

                                    uOO

        ‡

                            u

            V

                                                    o
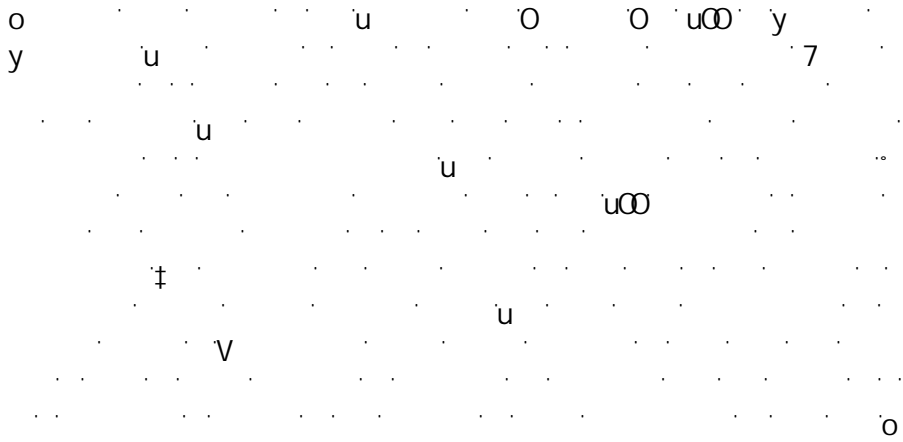
Figure 6.2 o   u
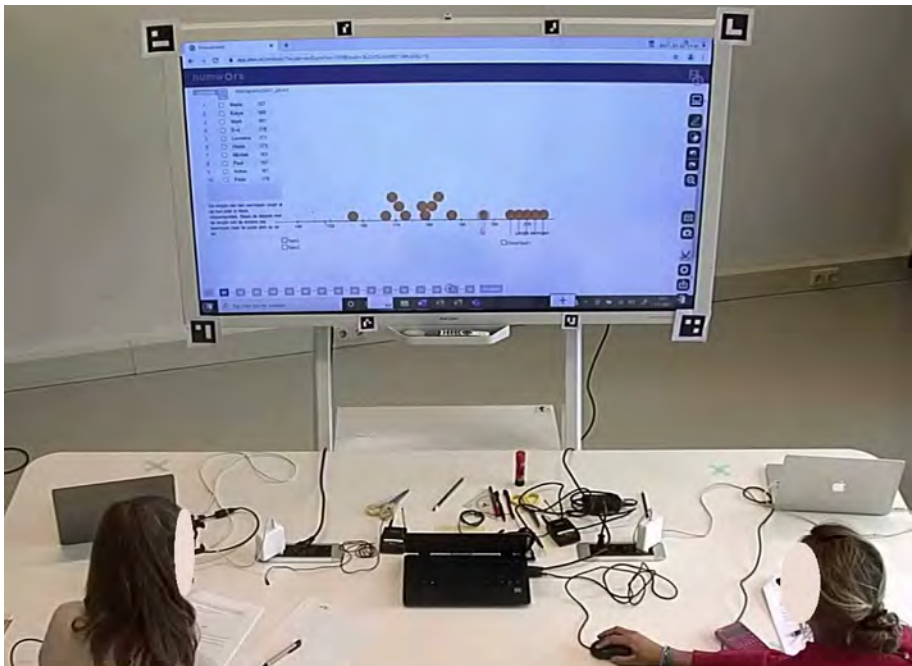
6.4.4 Data analysis—hypothetical learning trajectory

@   =Ou u   @

guideline for teachers and software designers both for the teaching experiment and for future use and 3) to analyze what actually happened compared to what we conjectured would happen. When conjectures are supported, tasks are kept; if not, we try to explain why and how tasks could be improved in a future redesign. In addition, the first author held mini-interviews with some (pairs of) students, with questions—Can you describe what the similarities and differences are between those graphs? (case-value plot and histogram, Task 1)—addressing H1a: *By comparing means and variation of data in two graphs, students experience that they focus on most apparent features that are similar for both graphs (such as shape, number, and position of bars), but irrelevant for this comparison.* The mini-interviews, together with students' written materials, and videotaped discussions, were to verify whether the conjectures of the HLTs were met. Some relevant conjectures are discussed in the next section.

Table 6.2 Summary of the HLT. Description of tasks and full HLT, see online materials

Step	Example task	Activities	Example conjecture
1 Learning initiation: Experiencing lack of understanding		Compare means and variation in a histogram (left) and case-value plot (right)	H1b: By experiencing initial confusion or misunderstanding, students' intentionality and motivation for upcoming tasks is established.
2 Reinventing the role of the horizontal scale in univariate graphs		Slide dots to their position on a scale.	H2b: By horizontally moving dots to their correct position on a horizontal scale, students notice the *position* of a dot depicts the measured value.
3 Reinventing the role of the vertical scale in histograms		Build a histogram overlay from a dotplot.	H3b: By moving the (orange) sliders up, students notice the height of the bars is related to the number of cases in a bar when class intervals are equal.

Phase	Image	Description	Hypothesis
4 Reinventing arithmetic means in histograms		Establish relation between data and mean; discover influence of outliers, gaps, distribution, on mean.	H5a: By finding the balancing point of the graph, students perceive the mean can be seen as the point where the graph is "in balance."
5 Confirming learning: transfer to other situations		Construct and interpret histograms on paper. Sort histograms and look-alikes.	H18b: By drawing a histogram on paper from a frequency table, transfer to another environment (paper) is established.

6.5 Materials, conjectures, results, and ideas for redesign

6.5.1 HLT step 1: Learning initiation–experiencing not understanding

Materials and conjectures HLT step 1: Task 1

Figure 6.3

\#

u

u
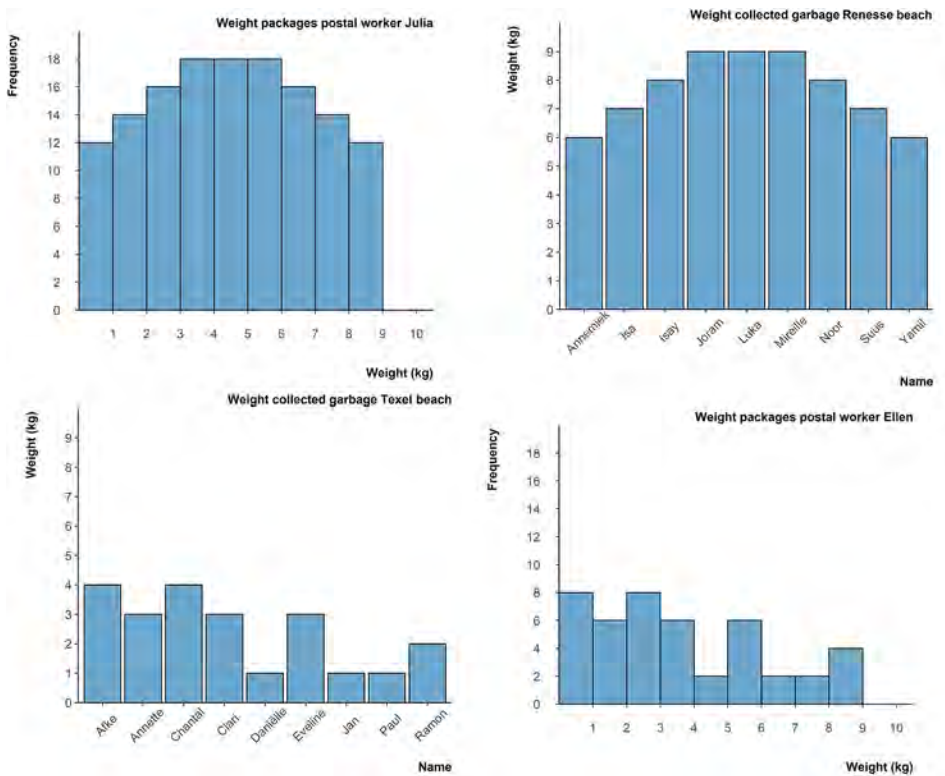
u = u *transfer* =Qu
u

u

= "

= "

Results HLT step 1

Table 6.3 7 u *N*

	u		u	
	U	†	U	†
O	**2**			
k		**2**	**3**	**5**
o				

Note # **bold**

u u

o u ‡
=

o \ \ K ‡

o y @

o \ \
x

o ' O @ @
"

o o

V o

O

u

o

Ideas for redesigning HLT step 1

describe

6.5.2   HLT step 2: Reinventing the role of the horizontal scale in univariate graphs

Materials and conjectures HLT step 2: Task 2

positioned

histogram is related to the number of cases, not to the measured value 'body height.' When all dots are placed, a check button is available. The conjectures are:

> H2a: By moving dots, students perceive that every dot in a bar stands for one measured value.
>
> H2b: By horizontally moving dots to their correct position on a horizontal scale, students notice that the position of a dot depicts the measured value.

Figure 6.4 In Task 2, the body heights of ten students are already depicted in a dotplot. Students are asked to position the heights of the six students at the far right by sliding these dots to their correct place on the number line. The labels on these six dots are (from left to right): 191, 176, 161, 185, 174, and 164
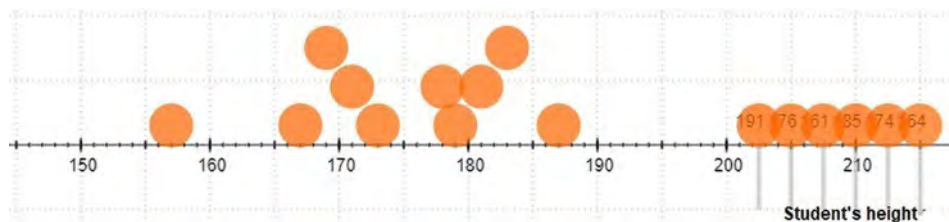
Results HLT step 2

Despite our design guidelines, we did not see much productive struggle. All students immediately tried to move the dots to their correct position. Students understood that the values along the horizontal axis were depicting body height. One student noticed he cannot choose the vertical position of balls:

S4: Well, you can't choose in height.
S5: No, he adjusts it himself.

One of the mini-interviews illustrates that the conjectures were met for these students:

S1: That determines the [body] height. Of the student.
T: And how, um, how do you see that it has a certain height?
S2: What do you mean? Where it is located? Like, on the *x*-axis [gestures horizontally with index finger] or something?

Ideas for redesigning HLT step 2

The action of connecting the data points with the table is missing in the current design of this task. In the next cycle, student names rather than numbers will be put on the balls so that students need to use the table.

6.5.3 HLT step 3: Reinventing the role of the vertical scale in histograms

Materials and conjectures HLT step 3: example Task 3

Tasks 3 and 4 both aim at reinventing histograms. The target knowledge is some aspects of the key concept of data, such as the number of cases and the measured values. The context is creating a histogram that supports a school principal in deciding which two chair sizes to buy. Each chair size is suitable for students with certain body heights. The histogram depicts students' heights. As chair sizes are related to students' heights, the histogram could be used for taking an informed decision about what chair sizes to buy. Both tasks concentrate on the following content: the height of a bar in a histogram depicting the number of cases in it (only when bin widths are equal). In Task 3 (Table 6.2 and Figure 6.5), students first drag down sliders to make the bins. Next, a pulled-up strip creates the height of the bar. In this task, a goal-oriented sensorimotor action aims to facilitate the perception of the height of the bar depicting the number of cases. When all bars are done, there is a check button with a green v or red x as feedback. The size (height) of the dots is purposefully not chosen as a unit height in the histogram, as we want students to think about what the height of each bar represents.

Tasks 3 and 4 have a similar layout. In Task 4, we created a context in which students need to move some dots to a new position based on changes in the data. As this influences the histogram, students also need to adjust the bars' heights. Task 4 aims to ensure that students understand the relation between the height of a bar and the number of dots (cases) in it. After Tasks 2–4, we expect students to understand these three most important aspects of the key concept of data: perceive that there is only one statistical variable (students' height), that the values of this variable are along the horizontal axis, and that the height of the bar is related to the number of cases in each bar (equal bin widths only). The conjectures are:

> H3a: By dragging the (grey) separating lines down, students notice that there are different measured values (students' height) depicted in one bar.
> H3b: By moving the (orange) sliders up, students notice that the height of the bars is related to the number of cases in a bar when class intervals are equal.

Results HLT step 3

All students mention that they do not know what a histogram is. S5 slightly hesitates when discussing the 10 cm steps. She first gestures horizontally to indicate which axis this 10 cm refers to and then states: "Yes, that's right".

Note that this gesture matches the enactments in the previous task and helped this student to rethink what runs along the horizontal axis. In addition, S4 dragging down sliders—creating bins—was enough for S5 to see it boxed as shown by her bimanual gesture, in which she uses her thumb and index finger as if grabbing two boxes, and moves these down, repeating S4's action. Her gestures grounded her understanding. The teacher intervenes a bit later to explain they can pull the orange slider up. Guided by S5, S4 immediately pulls it up to the correct height for each bar. S1 and S2 understood the horizontal binning ("from 170 to 180 is one size [of students' length for which chairs are to be bought]"), but did not know what the height of each bar represents. Their first attempt was to make all bars as high as the dotplot (Figure 6.5, left). Next, they used the context: two bars for the chosen chair sizes (Figure 6.5, middle). S2 plays with the middle bar and raises it all the way up to the maximum height possible (seven). This movement allowed for a new perception, as now S1 says: "Oh, maybe you should put as far up … as there are people in the box". Once the histogram is constructed, they discuss the graph:

Figure 6.5 Several attempts of S1 and S2 to construct a transparent histogram over a messy dotplot; the right one being correct
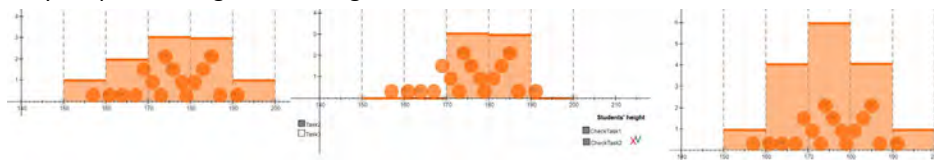
S2:   What is this scan? [Figure 6.5, right]
S1:   This is a histogram.
S2:   Huh, but he [the school principal] could only order two, right?
S1:   Yes, but this may help him, because now he knows…. now he can see that he has to be in those first three [she means the middle three] then he needs anyway that one [points to the middle bar] and then he can choose between those other two [left and right of the middle bar] because he knows how many people are in there.
S2:   Oh, that's clever.

In the mini-interview, the teacher-researcher asked: "And is everyone in such a bar the same height?" Both S1 and S2 pertinently answered no, in line with H3a. S4 and S5 said the height of each bar represents the number of people in that bar. S3 worked alone. He first used the context and created a histogram with two bars: each bar representing the number of chairs to be ordered (Figure 6.6), but he could not get the bar at the height he wanted (eleven). S3: "Wait. There are sixteen students [reads the text again] … and I can specify a

Figure 6.6
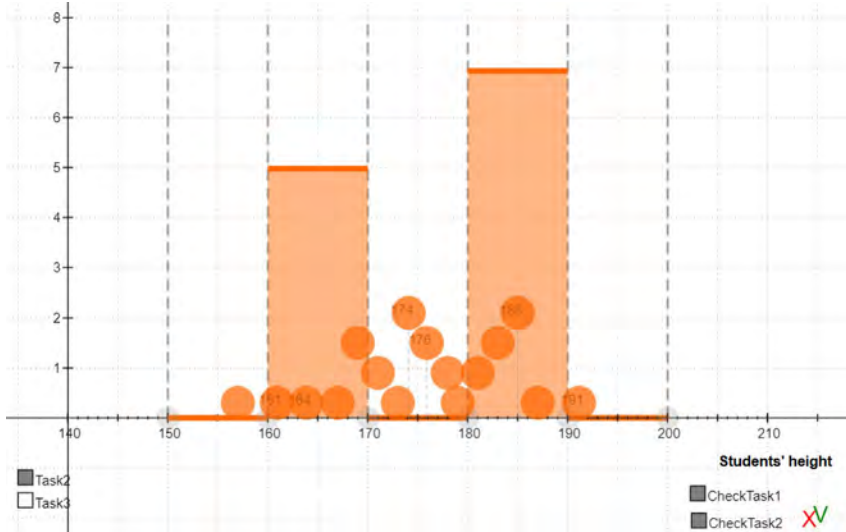
Ideas for redesigning HLT step 3

6.5.4   HLT step 4: Reinventing arithmetic means in histograms

Materials and conjectures HLT step 4: example Task 8

button. Then, a line with a blue triangle underneath it appears (Figure 6.7, left). The triangle can slide horizontally. To reach the task's aim, students untick the balance button and alternately add high (10) and low (1) marks (Figure 6.7, middle and right) and estimate and check means in between. During estimation, the tool was not visible. The idea of a balance tool is based on what students *spontaneously* described when solving similar tasks in our previous eye-tracking research (Chapter 3). Based on our design guidelines, we, therefore, introduced this artifact. The conjectures are:

> H8a: By adding low and high numbers alternately, students experience that higher numbers do not contribute more to the mean than lower numbers.
> H8b: By adding lower numbers further away from the mean and higher numbers closer to the mean, students perceive that numbers further away from the mean influence the mean more than numbers closer to the mean.

Figure 6.7 Example stages Task 8. Dotted line: students' estimation. Blue line: balance. Checking the estimation (left), adding a new mark (middle, balance tool not ticked), comparing estimated and actual mean (right). The solid vertical line and the dotted line are bolded here for clarity
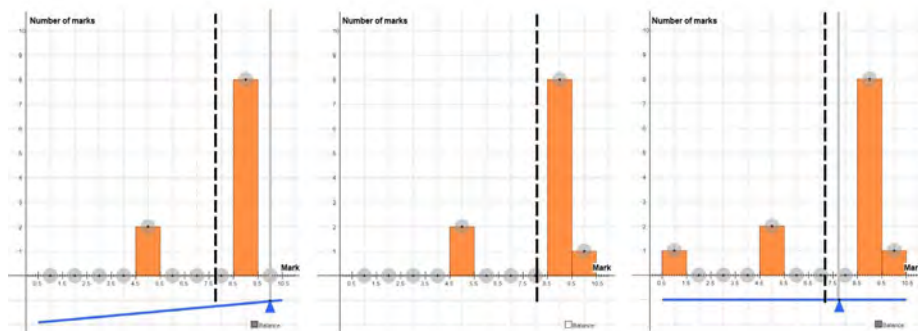
Results HLT step 4

While performing Task 8, S1 and S2 spontaneously formulated three ideas. We speculate that these were provoked by having students repeatedly estimate the balance point, write their estimation down, and then check it. Furthermore, the question "What do you notice about the balance points?" may have helped to form their ideas:

- If you add a ten and a one the mean stays the same (S1)
- Alternately adding tens and ones moves the mean toward 5.5 (S2)
- Adding a ten increases the mean (which is then around 7) by 0.2 and adding a one decreases the mean by 0.4. (S1)

Ideas for redesigning HLT step 4

Figure 6.8 Possible redesign of Task 5 from *action*-based embodied design perspective (left: incorrect, as mean is not in the correct position, right: correct, as mean and heights of bars are at a correct position). The left bar can be moved up and down only; the right bar is fixed. The triangle (mean) can only be moved horizontally
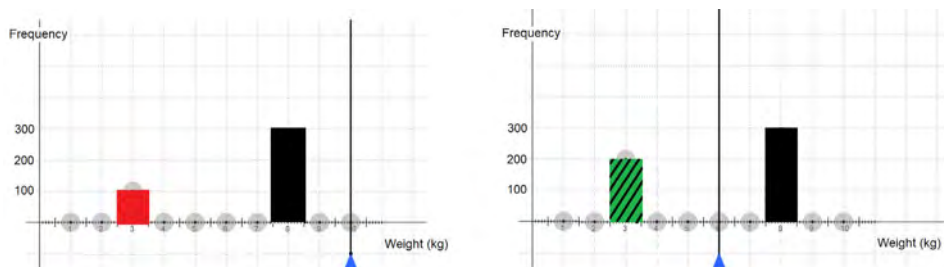
As an example, we could redesign Task 5—the first task that uses the mean (by asking for the balancing point in a bimodal histogram). One option is to introduce the mean through an *action-based* embodied design (Abrahamson et al., 2020) by using a motor-control problem. For example, one bar could have been drawn already (right, black bar, Figure 6.8) and one could have the students manipulate the height of a second (left, red) bar as well as the horizontal position of the (blue) triangle (mean) until they hit a correct combination of both. Students can then be asked to keep the task green while adjusting the bar and triangle continuously. Furthermore, a green/red frame (or screen) might be preferable over colored bars, as colored bars in the foveal view area might hinder the creation of perceptual structures (Bakker, Shvarts, et al., 2019). Having students fluently solve one or more motor-control problems regarding the mean, allows them to *reinvent* the balance artifact themselves before we introduce it. Therefore, in our example, the horizontal strip of the balance is removed and only the triangle is kept as a preparation for future introduction of this balance. Some applications use the horizontal axis as a balance (e.g., TI, 2015). As the digital environment we used would require too much reprogramming, we used a horizontal strip below the axis instead. It is left for future research to investigate how the placement of the balance (and triangle) influences students' conceptualization of the mean.

In retrospect, we note that Task 8 provided an opportunity to further mathematize by working toward the algorithm for calculating the mean (the process, Skemp, 1976) through the equilibrium of moments. In a future design cycle, we could, therefore, have students first estimate and then calculate the mean, and have them reflect on the algorithm. The benefits of *estimating* the mean from a histogram for large datasets might then also become more evident.

6.5.5 HLT step 5: Confirming learning—transfer to other contexts and environments

Materials and conjectures HLT step 5: example Task 18
The aim of step 5 is to create transfer to other contexts and environments. In Tasks 5, 14, and further, we vary contexts. As histograms are specifically designed for large datasets, we continued from Task 15 with increasingly large sets in realistic contexts. Since several studies show each environment creates other challenges (e.g., Alberto et al., 2022), some tasks were delivered on paper (Tasks 18–20 and 22). In Task 18, for example, students were asked to construct a histogram from a frequency table with given bin widths (*annual* income classes, see Appendix A of this chapter). Students first needed to draw a histogram for a given frequency table and then draw lines in it. Context and conjectures for this task were:

> If you have less than 1694 euros per month to spend as a family with one child, you are officially poor. At 1850 euros you don't have much but just enough […]. Indicate with a blue line in the histogram where the poverty line lies. For single people, the poverty line is 1039 euros per month. Indicate this with a red line.

H18a: By drawing a histogram from a frequency table, students perceive how an interval (e.g., < -10, 0]) is represented on the continuous horizontal scale in histograms.

H18b: By drawing a histogram on paper from a frequency table, transfer to another environment (paper) is established.

H18c: By drawing vertical lines for other values than the mean, students notice what part of the population is to the left of this line.

Results HLT step 5
Students performed as expected on most of the transfer tasks. The results of Task 20 (Table 6.3) indicate an improvement compared to Task 1. An exception is paper Task 18, which has mixed results. In it, all students but S2 (Figure 6.9, bottom left) correctly constructed a histogram with a continuous scale (ratio measurement level) along the horizontal axis, suggesting that H18ab is met. S2 asked S1 if annual income must go along the *x*-axis, and she horizontally gestured out this axis, in line with the newly built body artifact (here: *x*-axis) functional system in Task 2. S2 then gestures a horizontal line while confirming "annual income at the *x*-axis" and gestures a vertical line for "number of households on the *y*-axis." This gesturing suggests that they consider variables as enacted on a coordinate plane, which was what we aimed for in HLT steps 2 and 3.

To our surprise, three students drew horizontal lines for income: S1, S4, and S5 (Figure 6.9). The other two students corrected their initial horizontal

lines. S1 looked at the number [1039], then at S2, put her triangle ruler horizontally, looked at the number again, and drew a horizontal red line [at 1039, the *monthly* income e.g., Goderis et al., 2019]. She was attracted to the more similar numbers along the vertical axis (number of households in thousands) rather than the *annual* income numbers (in thousands, CBS, 2021) on the horizontal axis. When finished, S1 and S2 compared graphs, and S2 gestured a vertical line on the paper, first with her index finger and then with her whole hand, saying "you should indicate it like this." S1 responded: "but it is just…" and then paused, with her index finger on the words "X annual income thousands," marking the variable along the horizontal axis. She thought, then said: "Oh uh ha-ha," and then laughed "Oh, but it indeed needs to be like this," and gestured a vertical line. S1 noticed her mistake but did not change the drawing. Students S4 and S5 also were attracted to the vertical axis and drew these horizontal lines, most likely also using similarity between numbers. Again, we see no thinking. One explanation (cf. Kaplan et al., 2014) is that the magnitude of the numbers (number of households in thousands) along the vertical axis seemed to better match the magnitude of monthly income. To avoid this, students should first convert monthly income to annual income (e.g., by multiplying by 12 or 13). From an embodied perspective, we interpret this as students' body-artifacts (numbers) functional systems for finding similar numbers being so strong that they do not even think about what axis the number should go on.

Ideas for redesigning HLT step 5
Note that the intervals in the histogram of S2 could indicate that this student sees the bins as categories instead of a numerical scale. This conjecture could be tested in a future design cycle by presenting students with an unordered frequency table or with zero frequencies. Furthermore, to establish a better transfer of the measured values positioned along the horizontal axis, we suggest adding reflection questions. Next, in step 4 of the HLT, we could design some tasks that ask for the position of other values than the mean, with feedback. Another option is to extend HLT step 2 to paper before introducing step 3. Furthermore, the problem of monthly and annual income could have been avoided by providing annual income in all cases. However, for classroom use, we would prefer to keep the numbers from the original data sources as students could also encounter these in their daily life. Next, we would have a classroom discussion on the need to convert some numbers, rather than avoiding this problem. In addition, when working toward density histograms, using frequency for the height of bars in histograms might hinder further conceptualization. Using relative frequencies might solve this problem.

Figure 6.9
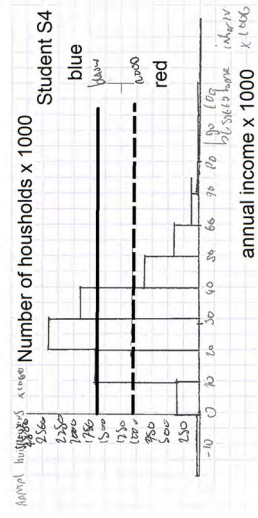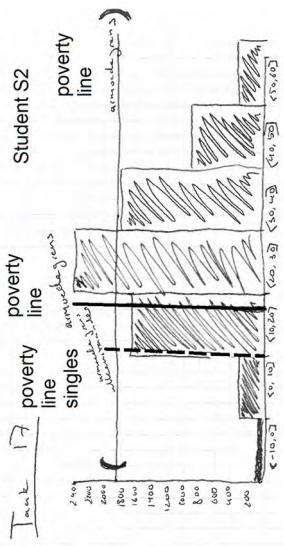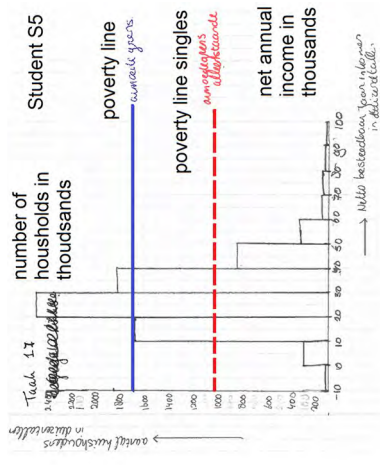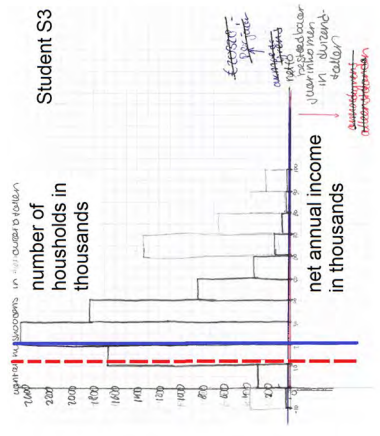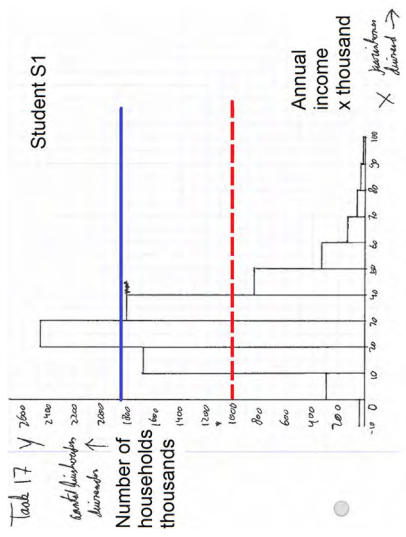
6.6   Conclusions and discussion

In this study, the central question was: What sequence of tasks designed from an embodied instrumentation perspective can support students' understanding of histograms and the underlying key concepts? Our answer is a 5-step hypothetical learning trajectory (HLT). Those steps were designed according to an embodied instrumentation approach in which learning is seen as enactment: Step 1 sets up a general goal for learning by having students experience a lack of understanding; Steps 2–4 aim to fulfill this goal by having students become aware of the role of the horizontal scales and reinvent the role of the vertical scales and arithmetic means in histograms. Transfer to other context and environments in Step 5 was a way for the students—and researchers—to confirm whether learning occurred. A sequence of 22 tasks was created to foster these steps. The tasks explicitly asked for goal-oriented sensorimotor actions that prompt students to perceive the most difficult aspects of how data, and their distribution, are depicted in histograms. Affordances of the digital environment allowed for direct exploration of histograms' qualities: axes, dots representing cases, histogram bars, and mean were directly extending students' hands in their body-artifacts functional systems, thus allowing for instrumented actions of representing data with histograms (Shvarts et al., 2021). Furthermore, the students were asked to search for the task's solutions, rediscovering several aspects of a histogram and including them in their emerging functional systems. The process of searching for the task solution is conceptualized as a productive struggle (e.g., Kapur, 2014; Roth, 2019). Productive struggle can be understood as "students attempting to make sense of something that is not immediately apparent, working toward reconfiguring their understanding of facts, ideas, or procedures" (Reitemeyer, 2017, p. ix). An example of this is the unit height for bars in histograms not being equal to the size of dots (e.g., Task 3), which invited students to explore the situation and notice this height is equal to the number of measurements (in the equal bin widths case).

Comparing students' performances with the conjectures from the anticipated HLT, our case study shows most conjectures were met. Students experienced misunderstanding in the first step, had no trouble imagining the role of the horizontal scale, struggled but reinvented the role of the vertical scale in histograms, seemed to have an easy task estimating the balance point of a histogram and stated that its practical relevance is that it is the arithmetic mean. The final task showed students mostly could transfer the acquired knowledge to paper, so the functional system formed within a digital environment could easily re-emerge in a different environment. Students' gestures indicated using actions from previous tasks to solve current tasks.

Taken together, the results suggest that embodied experiences with reflection contributed to overcoming some well-known misinterpretations. Yet, some difficulties occurred when numbers were given that seemed to better match the numbers along one of the axes. So, we suggest adding transfer tasks also after the steps of the HLT that are dedicated to horizontal and vertical actions. In addition, including other mathematical notions into the emerging functional systems can be improved, for example, by fostering students' reinvention of the algorithm of calculating the mean based on their sensorimotor estimations. To further develop students' notions of distribution and variability, for example, in density histograms, the artifact "area" may need to be included in the design, and the artifact "interval" may need to be reinvented by students (Boels & Shvarts, 2023).

A limitation of the study is the small number of students in it and their varying previous experiences with histograms. Moreover, in most classrooms, there will not be time to spend 4–5 lessons (3.5 hours) on 'one' topic. In further discussing limitations, "It is important to acknowledge that the complexity of students' […] learning, and of the designed learning environments, makes it impossible to specify completely everything that transpires in the course of a design study." (Cobb et al., 2016, p. 40).

A methodological contribution of our work is our design guidelines. As the generalization and value of design guidelines come from the iterative process of letting the guidelines do the actual work (Bakker, 2018), we now revisit our theory-driven design guidelines based on the empirical tryout. Using our first design guideline (identify the actions that could have constituted the target artifact), again during the evaluation of the results, we reconstructed how we melted artifacts back to the actions that are crystallized in them (Figure 6.1). It made us aware that we did not pay enough attention to the binning action that histograms reify. For future research, we call researchers to question all pre-given aspects of the artifacts they use and to reveal artifacts' origins.

The second design guideline—to design motor-control or perception tasks to which these actions are the answer—helped us think about redesigning HLT step 4 (estimating arithmetic mean). This guideline includes productive struggle and as we can see from the results, productive struggle can create aha moments. Moreover, the absence of such moments for crucial steps in the HLT (e.g., step 2) might underlie difficulties during transfer to another environment (e.g., positioning lines in Task 18). We, therefore, suggest using "create productive struggle for crucial steps in the HLT" as a separate, sixth design guideline. Theoretically, it means that tasks that students' emerging functional system is solving should be new enough for the learning process—rather than simple recollection—to happen.

The third and fourth guidelines—on performing (digital) actions with feedback and reflecting on them—need no further elaboration. They are solid guidelines matching previous work on embodied design (Abrahamson et al., 2020; Alberto et al., 2022). In Task 18, we saw that two students did not discover their misplacement of poverty lines. This observation makes us wonder if further learning would have been induced if we had included feedback here.

The fifth guideline is to create possibilities for transfer of actions by varying contexts and environments as we did in steps 4 (different context) and 5 (paper). The difficulties students encountered on paper Task 18 stress its importance. A functional system needs to be flexible and adaptable to various environments. Therefore, such a transfer is also desirable within each crucial step and not only at the end of an HLT.

A scientific contribution of our work is that it substantiates the general ideas of embodied instrumentation by showing how more complicated artifacts (e.g., histograms) can be reinvented from actions with simpler ones (e.g., positioning dots on a scale, Figure 6.1). Our HLT can be seen as a further step from a general theory of embodied instrumentation toward a *domain-specific instructional framework* on teaching how data and their distribution are depicted in univariate graphs such as histograms, dot, box, stem-and-leaf, and hatplots (Konold, 2007) and histodots (Chapter 2). Unlike other researchers who work on general principles of enactive pedagogy (e.g., Abrahamson et al., 2021), we try to work out a design framework that helps to design for a specific mathematical domain (e.g., statistics) and topic (e.g., histograms).

We now discuss recommendations for future research and design. In the previous section, possible improvements of the tasks are suggested based on the results. Further solidification of students' understandings requires further enactments with histograms. In line with statistics education literature, we suggest adding more comparison tasks for dotplots (messy and stacked) and histograms as comparing data of two groups is core to statistics and important for developing statistical literacy (e.g., Garfield & Ben-Zvi, 2008a). Students could first be asked which group is better (cf. Watson & Shaughnessy, 2004) and then which group has a higher mean and which has higher variability. This task is similar to Tasks 1 and 20 in our HLT, but now with the *same* type of representation. Furthermore, we suggest adding graphs with the same ranges to support students proceeding from informal measures for variability (range) to more formal ones (e.g., deviation from the mean). Students can also be asked to produce at least two different datasets for a given histogram or to collect their own data and depict these in univariate graphs (e.g., Garfield & Ben-Zvi, 2008a). In future cycles, it is advisable to guide

students to reinvent other univariate graphs as "in statistics different […] representations are used to identify different aspects of the same data (transnumeration)" (Burrill & Biehler, 2011, p. 64). A future design cycle could profit from other approaches to embodied designs, such as probability (Abrahamson et al., 2020), for example, by having students stack paper cards with measured values into bins.

This leads to several questions for future research and redesign. For instance, how should our HLT be placed along the curriculum? Little is yet known about how students in Grades 6–8 interpret histograms (e.g., Bakker, 2004a; Whitaker & Jacobbe, 2017). Are more concrete models, such as the balance model with blocks and a ruler, suitable for Grades 4–6 (e.g., O'Dell, 2012)? What are the benefits or downsides of starting at an earlier age? Our HLT guides students in a specific direction. Given our design guidelines to carefully add reinvented artifacts in successive steps, this seems appropriate for *initial* learning. How to proceed? Are all suggested tasks needed in Grades 10–12 or can some tasks be combined? How could—or should—our HLT deal with students who have already had experience with the target knowledge as well as with several mathematical and statistical artifacts? Students' mathematical backgrounds can hinder the development of new forms of perceptuomotor structures (Shvarts & Van Helden, 2021). Could our HLT benefit from some flexible adaptation for Grades 10–12 students as well as tertiary students, as their mathematical backgrounds can be extremely diverse (e.g., Bor-de Vries & Hoogland, 2020)? In addition, design-based implementation research (Fishman et al., 2013) recommends that future design cycles involve more stakeholders and occur in classrooms. Designing from an embodied instrumentation perspective highlights that software designers need to think carefully about what kind of actions (crystallized in artifacts) they outsource to the software and what actions they transform into tasks for the students. For example, most software can automatically create histograms but lacks possibilities for students to reinvent them. Similarly, software designers might include an option for students to freely drag two graphs to a position suitable for comparing the graphs instead of presenting graphs in an already comfortable position for comparison. There is a risk of outsourcing actions to software too early, which hinders students to notice critical aspects of mathematical practice and artifacts. Our study exemplifies how designing from an embodied instrumentation perspective can help detect such actions. We call on software designers to create opportunities for students to perform these actions themselves during initial learning of mathematical and statistical concepts.

Appendix A HLT and example tasks

A.1 Hypothetical Learning Trajectory (HLT) (Simon & Tzur, 2004)

Task 1	Example

Description of task

Paper and pencil

Compare the mean and variation in the weight in both graphs. Reflect on the answer and what a bar in the left-hand graph is depicting

Design considerations

By comparing histograms and case-value plots, students experience the (often unnoticed) confusion they have with these two representations.
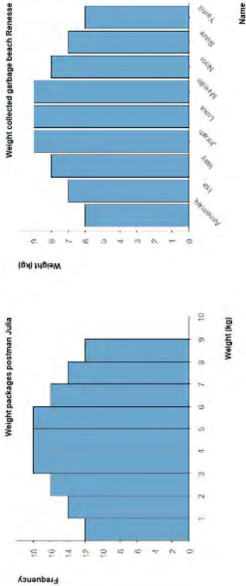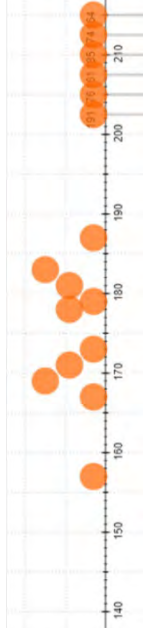
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Example of possible test item from international statistics tests. In which graph is the mean and variation in weight bigger? Explain the answer (how and why). What is the weight of the packages in the leftmost bar of the left-hand graph?	Posing a learning problem for students in which they experience not understanding leading to productive struggle. This is triggering students' intentionality and motivation (through the body-artifact functional system). Also: measuring pre-knowledge.	Comparing graphs, using (informal) measures of center and spread, reflecting on how the answer was found and why this answer is correct, thinking about the spread of weights in one bar.	H1a: By comparing means and variation of data in two graphs, students experience that they focus on most apparent features (such as number and height of bars) that are similar for both graphs (shape, number and position of bars) but irrelevant for this comparison. H1b: By experiencing initial confusion or misunderstanding, students' intentionality and motivation for upcoming tasks is established.

Task 2	Example	Description of task
	Note that this is the first digital task.	**Description of task** Slide six dots depicting the height of six students to the correct position in the dotplot. The height of ten other students is already depicted in the dotplot. Continuous movement possible. [Note that the dotplot lay-out could be improved. Due to constraints of the digital environment, this was not done.] **Design considerations** Messy dotplots have in common with histograms that they reify the action of positioning cases. As dotplots have only one axis, we focus students' attention on the horizontal axis depicting the measured values as a preparation for histograms. That dots are raised when too close to each other can trigger a naïve perception of higher is more. We avoided a stacked dotplot as it might induce the same difficulties related to the height as for histograms. Implicitly this task addresses that there is only one measured variable.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
The height of students is measured and some students heights' still need to be placed on the correct position of the scale. Feedback: tick check task.	Have students understand data. Create an opportunity for students to perceive that every dot is a case, and the horizontal orientation of the statistical variable.	Goal-oriented horizontal hand movement. Students are asked to move a dot to the target position, thus actively associate the value of the variable with the position on the horizontal axis.	H2a: By moving dots, students perceive that every dot in a bar stands for one measured value. H2b: By horizontally moving the dots to their correct position on a horizontal scale, students notice the position of a dot depicts the measured value.

Task 3 **Example**

Description of task

Slide down the borders for class intervals (fixed, given intervals). Move a slider up in discrete intervals of 'one' in accordance with the number of dots in a bar. Dots are visible roughly on their correct horizontal position (vertical position is irrelevant). Size of the dots is unequal to size 'one'.

Design considerations

By sliding down the borders for classes, students actively create these classes.

By using an histogram overlay onto the dotplot, all data values are still visible.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
School principal wants to buy chairs of different sizes for a class. Chair sizes depend on students' height. Histogram is used for advice. Feedback: tick check task.	Have students understand data. Create an opportunity to perceive that the cases in a bar of a histogram contain different values and that the height of the bar is the number of cases in the bar.	Goal-oriented vertical hand movement. The students are required to count the number of cases in a bar and depict it by the height of the bar (the vertical position of the strips).	H3a: By dragging the (grey) separating lines down, students notice that there are different measured values (students' height) depicted in one bar. H3b: By moving the (orange) sliders up, students notice the height of the bars is related to the number of cases in a bar when class intervals are equal.

247

Task 4	Example

Description of task

Move some dots from their position in task two to a new position. Move the sliders (top of the bars) to adjust the bars height to the new situation.

Design considerations

Both actions of positioning cases along the horizontal axis and raising bars up according to the number of cases in each bar are now combined. The target knowledge (key concept of data) is addressed in task 2–4.
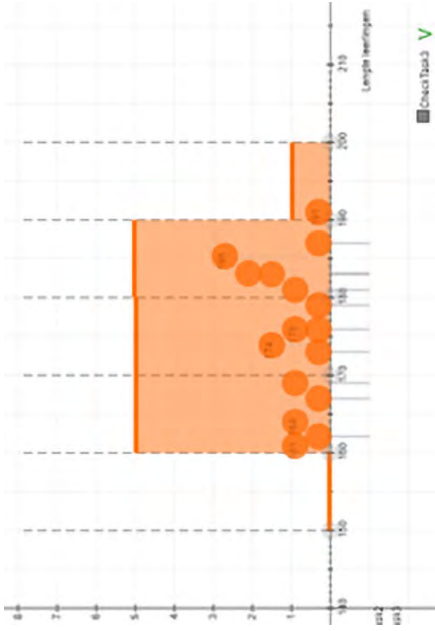
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Three girls wear high heels and, therefore, need higher chairs. Will the adjusted histogram influence the advice to the school principal? Feedback for histogram: tick check task.	Have students understand data. Create an opportunity to perceive to ensure that students understand the relation between the height of a bar and the number of dots (cases) in this bar	Combination of task 2 and 3. Goal-oriented horizontal and vertical hand movement. The students are required to move some cases and adjust each bars' height accordingly.	H4a: By moving dots, students perceive that the new position of cases can influence how many cases there are in a bar and thus its height. H4b: By moving only a few dots, students notice that the shape of a histogram is sensitive to small changes in the data.
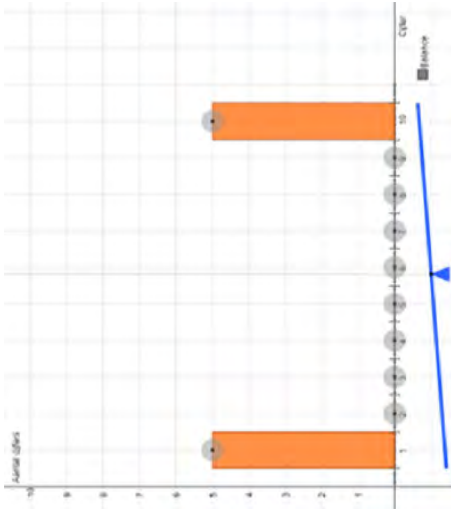
Task 5	Example
	Description of task Create a histogram from marks in a table. Find the balancing point and practical meaning of this point. **Design considerations** The new target knowledge is the key concept of distribution. The mean—as the balancing point of a histogram—is introduced as an object that can be estimated from the graph. By using a bimodal histogram with two bars only, we expect students to perceive the balancing point as being the arithmetic mean. Note that the numbers in the given table (only visible in the online tasks) are the middle values of each bar.
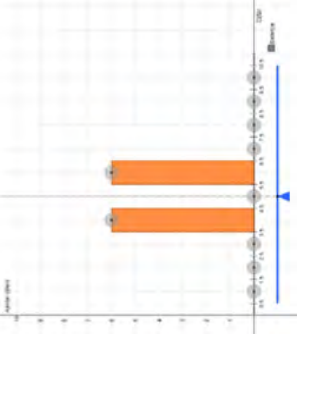
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
The next series of tasks is designed to discover some specific properties of a histogram. Anna has these marks. Create a histogram for these marks. Drag the triangle (of a slider) to the point where the balance (graph) is in balance. What is the practical meaning of the number corresponding to this point? Feedback: blue line is horizontal.	Have students understand distribution. Create an opportunity to perceive that the mean depends on both the spread of the bars as well as the height of the bars.	Coordinating horizontal and vertical hand movements with horizontal and vertical eye movements. The students are required to find the point where the graph is 'in balance'. The blue line will be horizontal in this balancing point.	H5a: By finding the balancing point of the graph, students perceive the mean can be seen as the point where the graph is 'in balance.' H5b: By seeing the big difference between the bars, students notice that the mean may not always be the best measure of center for a bimodal distribution or a distribution with a large spread.
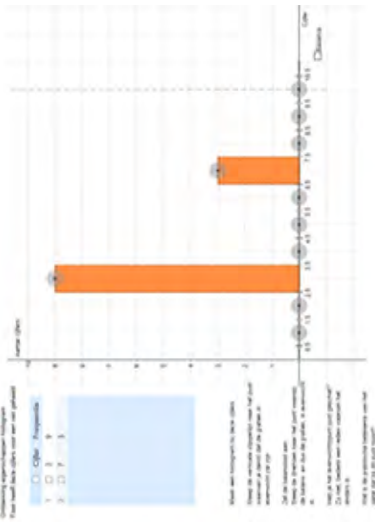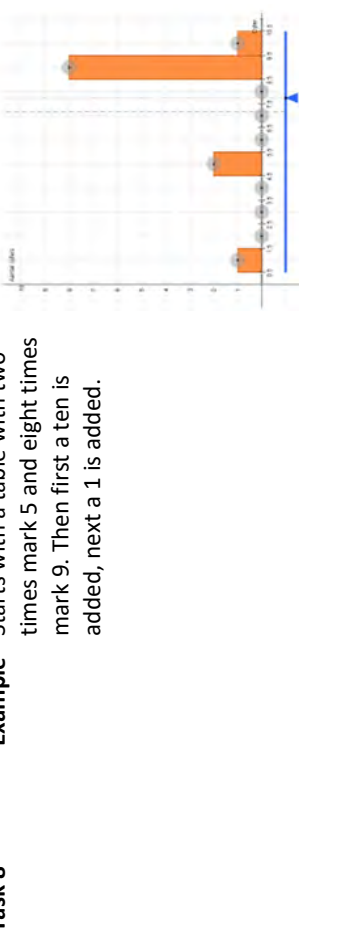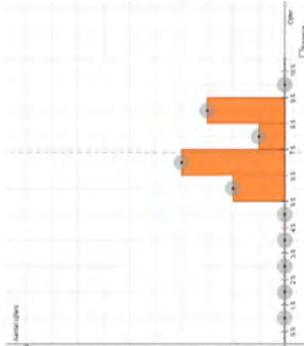
Task 6 **Example** **Description of task**

Estimate the balancing point before you drag the triangle to this point.

Design considerations

As we are not sure students will perceive the balancing point as being the mean, this task is similar to the previous, but now with bars closer to each other, expecting this will make it easier for students.

Note that this histogram has a different scale: instead of showing the middle values of each bar, the borders of each bar are shown. Furthermore, we introduce a vertical line that can be moved first (to indicate an estimation of the balance point, hence of the mean) before the balance point is checked.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Lonneke got these marks for one topic.	Have students understand distribution.	See task 4. In addition, focusing on the vertical line being an estimation of the mean.	H6a: By estimating the balancing point and then drag a vertical slider to this point, students notice that the mean can be perceived as a point where the graph is 'in balance.'
Create a histogram with these marks.	Create an opportunity to ensure students perceive that the mean		
Drag the vertical dotted line to the point where you think the graph will be in balance.	depends on both the spread of the bars as well as the height of the bars.		H6b: By moving the vertical slider to a position without measured values, students notice that the mean is not
Drag the triangle to the point at which the balance, and hence the graph, is in equilibrium.	Stimulate this perception by have students first estimate before dragging.		necessarily equal to a measured value and that it is possible that none of the bars represents the mean.
What is the practical meaning of the number associated with this point? Feedback: blue line is horizontal.	Have students experience the influence of the 'zero' bars and that the mean can be a value that is not measured.		H6c: By noticing that the horizontal scale is different compared to the previous task, students perceive that they need to pay attention to the scale.

Task 7	Example	Description of task
	Table given with nine times mark 3 and three times mark 7	Estimate the balancing point before you drag the triangle to this point

Design considerations.
One issue students might have is that they might think the arithmetic mean is the middle value of bars. Another is that higher values 'count more.' This task addresses both issues. Furthermore, in this task we connect a frequency table (hence, containing aggregated data) with a histogram (also containing aggregated data). |

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Floor got these marks for one topic.
Create a histogram with these marks.
Drag the vertical dotted line to the point where you think the graph will be in balance.
Drag the triangle to the point at which the balance, and hence the graph, is in equilibrium.
What is the practical meaning of the number associated with this point? Feedback: blue line is horizontal.	See task 5. In addition: being able to make a histogram from a frequency table.	See task 5.	H7a: Through the feedback of the balance tool, students notice that the mean is not always equal to the middle value.
H7b: By experiencing that the mean is closer to higher bars in a histogram, students perceive that 'heavy' bars count more.
H7c: By experiencing that the mean is closer to lower numbers, student perceive that lower numbers can contribute more to the mean than higher numbers.

| Task 8 | Example | Starts with a table with two times mark 5 and eight times mark 9. Then first a ten is added, next a 1 is added. |

Description of task

Estimate the balancing point before you drag the triangle to this point.

Then add a mark of ten. Estimate the balancing point before you drag the triangle to this point.

Repeat this procedure with adding a one.

Design considerations

This task elaborates on the insights we expect students to perceive in the previous task. It also works toward mathematization of the arithmetic mean.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Orlando got these marks for one topic. Create a histogram with these marks. Drag the vertical dotted line to the point where you think the graph will be in balance. Drag the triangle to the point at which the balance, and hence the graph, is in equilibrium. Feedback: blue line is horizontal. Then this is repeated by unticking the balance tool, adding a 10, making a new estimation, ticking the balance tool, dragging the triangle to the point at which the balance is in equilibrium. This is once more repeated by adding a 1.	Understanding that high numbers do not count more than lower numbers. The mean is influenced by the distance of the extra mark to the mean: the bigger the difference, the higher the influence on the mean.	See task 5.	H8a: By adding low and high numbers alternately, students experience that higher numbers do not contribute more to the mean than lower numbers. H8b: By adding lower numbers further away from the mean and higher numbers closer to the mean, students perceive that numbers further away from the mean, influence the mean more than numbers closer to the mean.

Task 9	Example	Description of task
	Optional task. Example of possible task.	Play around. Give your peer a histogram for which s/he needs to find the balancing point. **Design considerations** By giving students the freedom to explore, we expect they discover some aspects of symmetrical and skewed distributions in histograms. In this way we expect to establish their perception of the arithmetic mean as a measure for a distribution.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Now make up your own histogram for which you think the balance point is difficult to find. Give this histogram to your classmate. For your classmate: Estimate the balance point by dragging the dotted line there and then drag the triangle to the point where the histogram is in balance. Then swap roles. You may do this a few times.	Discovering the influence of different distributions in histograms. Perceive the relation between the data and the mean.	Creating histograms. Challenge your peer and yourself. Think about what 'difficult' histograms may look like.	H9a: By constructing several histograms, students reinvent symmetrical and very skewed distributions. H9b: By constructing several histograms, students develop a sense of the relation between the distribution of the data (here: the shape of the histogram) and the position thus value of the mean.

Task 10	Example	Similar to task 8, but students are now asked to create a histogram for which they can estimate the balance point themselves.	**Description of task** Play around. Give your peer a histogram for which s/he needs to find the balancing point for which you can give a good estimation for the balance point. **Design considerations** This tasks elaborates on the previous task by asking students to focus on means they can estimate. We expect that in this way they think more carefully about the relation between mean and distribution, and in this way further establishing their perception of the arithmetic mean as a measure for a distribution.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Now make up your own histogram for which you think you can give a good estimation of the balance point. Write this secret estimation down. Give this histogram to your classmate. For your classmate: Estimate the balance point by dragging the dotted line there and then drag the triangle to the point where the histogram is in balance. Then swap roles. You may do this a few times.	See task 8.	See task 8.	H9a: By constructing several histograms, students reinvent symmetrical and very skewed distributions. H10: By constructing histograms for which they can estimate the mean, students perceive that both the spread and height of the bars in a histogram influence the mean.

Task 11	Example	Description of task
		Description of task Creating a histogram from a table with decimal marks, estimating and checking the mean. **Design considerations** In previous tasks we used integer numbers. The idea of a bar containing different measured values is here re-established by using decimal numbers in the table.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
LeBron had these marks for a test. Create a histogram from these marks. Drag the vertical line to the point where you think the graph will be in balance. Drag the triangle (of a slider) to the point where the graph is in balance. Feedback: blue line is horizontal.	Establish students' perception of the relation between the data and the mean. Have students perceive that a histogram is specifically meant for continuous data. Draw students' attention to the variation of data within a bin (class interval).	See task 4. In addition: creating bars with different numbers in one bar.	H10: By constructing histograms for which they can estimate the mean, students perceive that both the spread and height of the bars in a histogram influence the mean. H11a: By constructing a histogram containing measured values with decimal numbers, students again notice that each bar in a histogram contains different numbers (cf. H3a). H11b: By receiving feedback on their estimation of the mean, students using an equal area strategy notice that this strategy does not lead to the mean (but the median) if the histogram is not symmetrical.

Task 12	Example	Description of task

Description of task

Investigate the influence of one extra (extreme) value on the mean.

Histogram is given; students have the possibility to move the estimation line and drag the slider. Also, the height of each bar can be adjusted.

Design considerations

Addressing another aspect of the mean: the influence of new data points being added. At first, it seems impossible to solve this task (creating productive struggle) until students realize that an exact answer is not possible and not necessary.
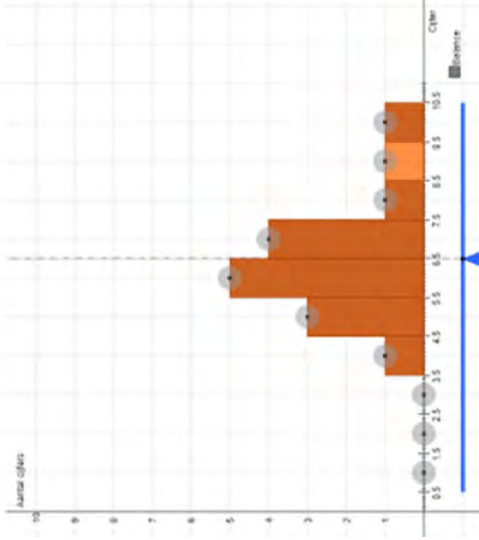
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Divya has gotten these marks for a test. Which score does she have to get for the last test to get a 7 on her final list?

Feedback (optional): blue line is horizontal. | Establish that an extra value is influencing the height of a bar. Influence of one extra (extreme) value on the mean.

Discussion on the exactness of mean and values in a histogram.

Note that a mean of exactly 7 seems not possible in this histogram; only 6.6 (or lower) seems possible. | Adjusting the slider to the intended mean of 7, horizontal hand movement. Then adjusting the height of one bar to get the balance in balance (vertical hand movement). | H12a: By adding a data point to get a specific mean, students notice how an extra data point influences the mean.

H12b: By observing that the exact values in a bin are not given, students notice that in a histogram an exact determination of the mean is most often not possible due to aggregation of the data in a bin. |

Task 13

Example

Description of task

Investigate the influence of two different values in the same bar on the mean.

Histogram is given; students have the possibility to move the estimation line and drag the slider. Also, the height of each bar can be adjusted.

Design considerations

Elaborating on the previous design and the mean as well as the aggregation of data in a histogram.
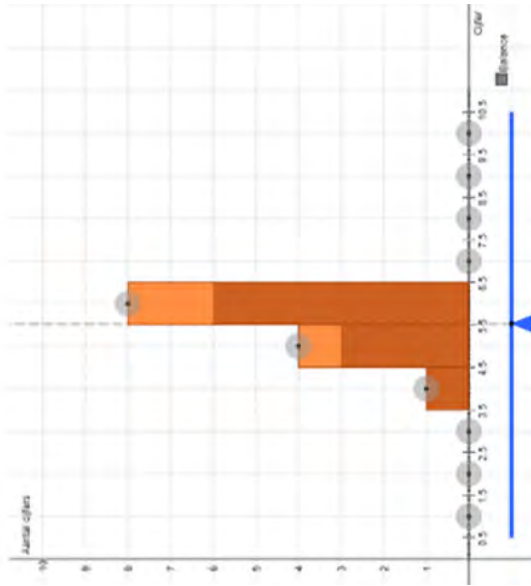
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Nine of Divya's test marks are already in the histogram. Next, she gets a 6.1 and a 5.6. Adjust the histogram by adding the two new marks. What mark does she need to get for the next test to get a 6 on her final list?	Perceive that a histogram is aggregating information and throwing away exact information.	Making one bar 'two' higher (vertical hand movement). Adjusting the slider to the intended mean of 6, horizontal hand movement. Then adjusting the height of one bar to get the balance in balance (vertical hand movement).	H13a: By adding the marks 6.1 and 5.6 students notice that two different data points can be in the same bar. H13b: By adding two marks, students experience that the influence on the mean of adding data points close to the mean is limited.

Task 14

Example

Description of task
Exploring the mean in a different but known context.

Design considerations
Vary context to create transfer.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Create a histogram with this table. Drag the vertical dotted line to the point where you think the graph will be in equilibrium. Drag the triangle to the point where the balance is in balance. What is the practical meaning of the number corresponding to this point? Feedback: blue line is horizontal.	Perceive the influence of a different context and distribution	See task 4. Reflecting on the first tasks in the DME.	H14: By estimating the mean in a different context, students transfer their activities to other situations.

Task 15	Example	Description of task
		Exploring the mean in a different context.
		Design considerations
		Vary context and using larger, realistic data sets to create transfer.
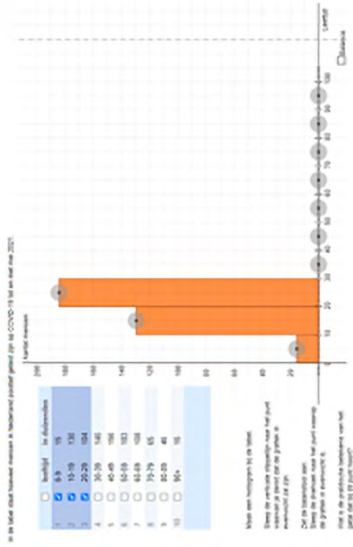
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
The table shows how many people in the Netherlands have tested positive for COVID-19 up to May 2021. Create a histogram with this table. Drag the vertical dotted line to the point where you think the graph will be in equilibrium. Drag the triangle to the point where the balance is in balance. What is the practical meaning of the number corresponding to this point? Feedback: blue line is horizontal.	See task 14, now for a larger data set.	See task 4.	H15: By estimating the mean in a different context, students transfer their activities to other situations.

Task 16	Example	Description of task
		Rebuilding the histogram of the previous task and choosing bin widths **Design considerations** Vary context and using larger, realistic data sets to create transfer.
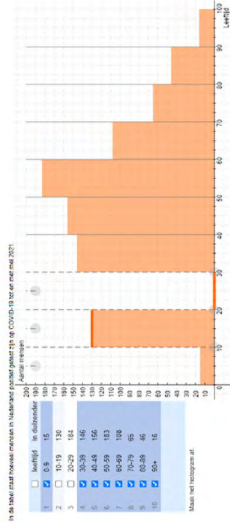
Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Finish the histogram.	Perceive that bin widths are a choice.	See task 14. In addition, students drag down the sliders for some bins. They have the possibility to create smaller (or larger) bins.	H16: By dragging down (grey) lines which create bins, students perceive that bin width is a choice.

Task 17	Example	Description of task
	Note that this is the last digital task.	Exploring influence of unequal class size (unequal bins). Splitting one bin into two bins. **Design considerations** Transfer to density histograms in a context that is now familiar (from previous tasks).

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
The table shows how many people in the Netherlands have tested positive for COVID-19 up to May 2021. The school principal wants to know whether different COVID-19 protection measures need to be taken for students within the age group 10-19 years old. Therefore, we split the 10-19 age group into 10-14 and 15 – 19. The new numbers are: 10-14 à 51903 and 15-19 à 77820. Split the bar for those two age groups into two bars according to the proportion in each age group and in such a way that the total area of the histogram does not change. What is the vertical axis now depicting? What is your advice to the school director regarding different COVID-19 protection measures for younger and older students?	Unequal class intervals require density on the vertical axis. Understanding that the total area of a histogram does not change (100%).	Coordinating the area of a histogram with the height of the bars after the split.	H17a: By keeping the total area of two smaller (5-years) bins equal to the area of the former larger (10-years) bin, students discover that the vertical axis in a density histogram does not depict frequencies. H17b: By keeping the total area of two smaller bins equal to the area of the former larger bin, students question what the vertical axis depicts and in this way reinvent a notion of density (e.g., number of people per 5 years) in a histogram with unequal bin widths.

Task 18	Creating a histogram from a frequency table with given bin widths.	**Description of task** Constructing a histogram for income classes on paper. Using different artifacts (paper and pencil). **Design considerations** Transfer to paper based on a realistic context. Construct a histogram and come back to positioning cases (now in the form of a vertical line depicting a poverty line for income).
Context and/or question for students	**Learning goal teacher**	**Learning activities**	**Hypothetical learning process**
Draw a histogram for the table below and then answer the questions. If you have less than EUR 1,694 per month to spend as a family with one child, you are officially poor; at EUR 1,850 you do not have much but just enough (if there are no setbacks). Indicate with a blue line in the histogram where the poverty line approximately lies. For single persons, the poverty line is 1,039 euros per month. Indicate with a red line where this is. What do you think of the statement: The Netherlands are a rich country. A middle income is more than enough to get by on?	Establishing that students can draw a histogram on paper. Relating the learning activities so far to school book problems. Understanding that both the table and the histogram can only give estimates of proportions or percentages.	Students make decisions on size, placement of the numbers on the horizontal scale, transforming class intervals (thus a seeming ordinal scale) into a continuous scale.	H18a: By drawing a histogram from a frequency table, students perceive how an interval (e.g., $< -10, 0]$) is represented on the continuous horizontal scale in histograms. H18b: By drawing a histogram on paper from a frequency table, transfer to another environment (paper) is established. H18c: By drawing vertical lines for other values than the mean, students notice what part of the population is left of this line.

Task 19	Example	Description of task
		Estimation of percentages from the total area of the histogram for a specific proportion. **Design considerations** Transfer to area being 100% as a preparation for density histograms.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
In reality, the distribution is much more skewed than you might think with the histogram from the previous question. About how many percent of households have less than 27000 euros disposable income per year? What percentage have more than 50000 disposable income per year? What would it mean for incomes if we went to 14 euros gross an hour as a minimum wage (that is 2467 gross a month)? The minimum wage is now about 1684 gross per month.	Preparation for a density histogram. Relating the height of the bars to a proportion of 100%.	Students draw vertical lines to support their estimation of the area at the left hand side of this line or at the right hand side of this line.	H19a: By estimating the percentage of a population in a more fine-grained version of a previously drawn graph, students notice that the total area of all bars represents 100% of the data points and that this can be used to estimate the percentage for a proportion of the total area (i.e., the population). H19b: In addition, students perceive that it is impossible to give an exact estimation of the percentage or proportion of people if a measured value is cutting a bar into two portions.
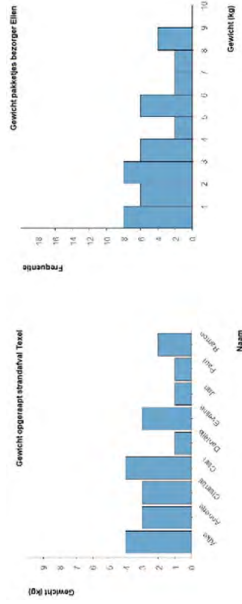
Task 20	Example	Description of task
		Comparing mean and variation in histogram and case-value plots.
		Design considerations
		Transfer to the starting task of this sequence.

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Both graphs show the weight. On the left is the weight of the parcels delivered by postal worker Julia. On the right, you can see the weight of the beach waste collected by different pupils. Which of the following statements about these graphs is true? The average weight is greater in the graph on the left, the graph on the right, or in both graphs approximately the same. The variation in weight is greater in the graph on the left, the graph on the right, or in both graphs approximately the same. Explain how you arrived at your answer and why you think this is so. What is the weight of the packets in the leftmost bar of the left-hand graph?	Distinguishing histograms from case-value plots. Ensuring that students understand the difference.		H20a: By comparing means and variation of data in two graphs, students experience differences in how and where measured values and variation are depicted in a histogram and case-value plot. H20b: By experiencing initial confusion or misunderstanding, students experience that they need to carefully read graphs before jumping to conclusions.

Task 21	Example	Description of task
	Two videos with gazes students: L08_it06.avi and L02_it06.avi. Both students in the video look at the histogram below	**Description of task** Watching other student's gazes and understanding what these students are doing **Design considerations** Confronting students with gaze patterns of other students.
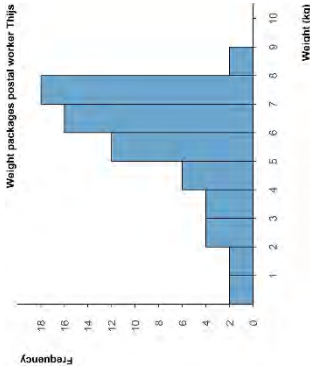
Weight packages postal worker Thijs (histogram: Frequency vs Weight (kg))

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
In this video, you can see how a student looks at a graph. The red dots are the places where a student looks longer, the thin lines are rapid eye movements between those places. The question this student answers is: What is approximately the average weight of the packages that Thijs delivers? The first student answered 6. The second student answered 8. Think about what you think the right answer should be and explain what the student may have been thinking here. You can watch the video as many times as you like.	Establishing a correct gaze pattern for histograms. Ensure that student note how and where the arithmetic mean is depicted in a histogram	Looking at other students' gazes and looking in a similar gaze pattern for correct gazes.	H21a: By watching other students' strategies, students experience that the estimation of mean and variation from a histogram require specific procedures. H21b: By watching other students' strategies, students perceive that there are different solution strategies for estimating the mean from graphs, some of which are effective for histograms and others not. H21c: By watching other students' strategies, students notice that most apparent features of graphs with bars are not always relevant.

Task 22	Example	Description of task
	Examples of histograms and other graphs with bars. Only graphs that are unambiguous. These came from school books, newspapers and previous research.	Sorting histograms and case-value plots

Design considerations
Transfer to graphs with bars found in newspapers, school books etc. |

Birth weight baby girls hospital 18 Dec 1997 — Frequency vs Birth weight in gram

Context and/or question for students	Learning goal teacher	Learning activities	Hypothetical learning process
Sort the graphs below. Put the histograms together and the graphs that are not histograms together. How would you explain to a class mate what a histogram is.

Explain how you sorted and why graphs are histograms or not. | Ensure that students can distinguish histograms from other graphs in realistic contexts. | Students categorize and discuss different graphs. | H22: By sorting histograms and look-alike graphs with bars, students notice the difference between relevant and most apparent features of graphs with bars. |

A.2   Description of selected tasks

Some pseudonymized videos as well as the paper version of the lesson materials can be found in a data repository (accessible for researchers on request). The link to all *digital* tasks used in this research is (in Dutch): https://app.dwo.nl/embod/?locale=en&profile=108&hash=%23s%3A698275#s :698275. Below, for each task discussed in the article, as well as some other tasks, relevant graphs or screen shots are shown to give the reader an impression of the tasks.

A.2.1 Task 1

Task 1 was a paper-and-pencil task. For this research, this task was projected on a digital whiteboard in a Word document.

International research shows that some graphs from newspapers and scientific articles are more difficult to understand than others. The question below might appear in such an international test. By the end of this series of tasks, you will find it easier to answer the questions below.

Figure A.1 Graphs used in paper task 1
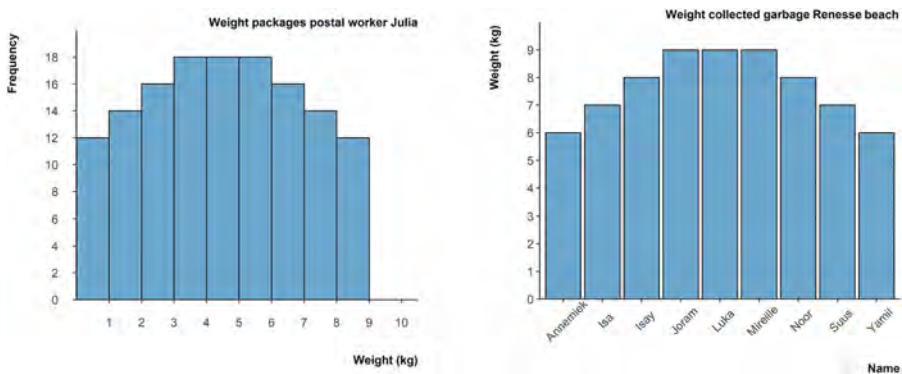
Both graphs show weights. On the left, you see the weight of packages delivered by postal worker Julia. On the right, you see the weight of beach waste collected by different students. Which of the following statements about these graphs is true?

The average weight is larger in the
- graph on the left,
- graph on the right,
- approximately the same in both graphs.

The variation in weight is larger in the
- graph on the left,
- graph on the right,
- approximately the same in both graphs.

Explain how you arrived at your answer and why you think this is so.

A.2.2 Task 2

Figure A.2 Start and end screen task 2
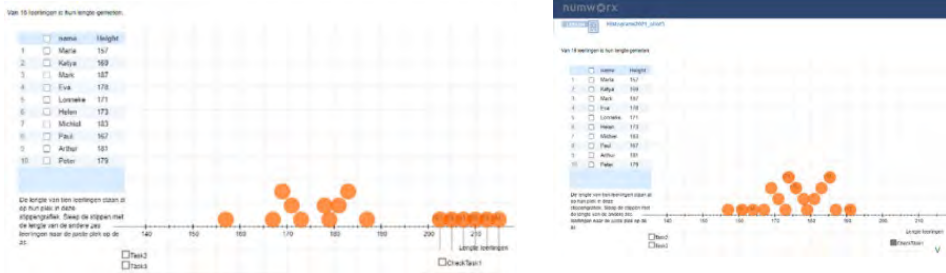
Note. After solving the task, students are asked to tick the CheckTask2 box on the bottom left. The feedback is either a green v (correctly solved) or a red x (incorrectly solved).

A.2.3 Task 3

Figure A.3 Screens task 3 where students advise a school principle (see, for example, Eshach & Schwartz, 2002)
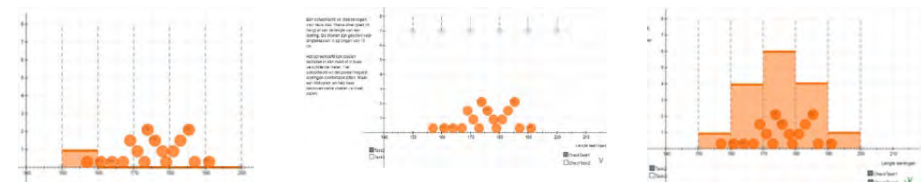
Note. In task 3 students make a histogram overlay on the dotplots. Start screen (left), all sliders pulled down and the first one pulled up (middle) and solved and checked task (right). The context for the task is given on the left side of the screen.

A.2.4 Task 4

Figure A.4 Task 4 requires moving data points and adjusting the histogram from task 3. Screen shot of the task after being correctly solved and checked by the students
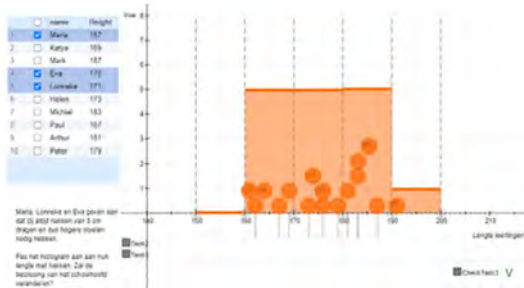
Note. Ticking boxes for keeping track of what has been done (left side of the screen) is optional.

A.2.5 Task 5

Figure A.5 Screen shots of task 5
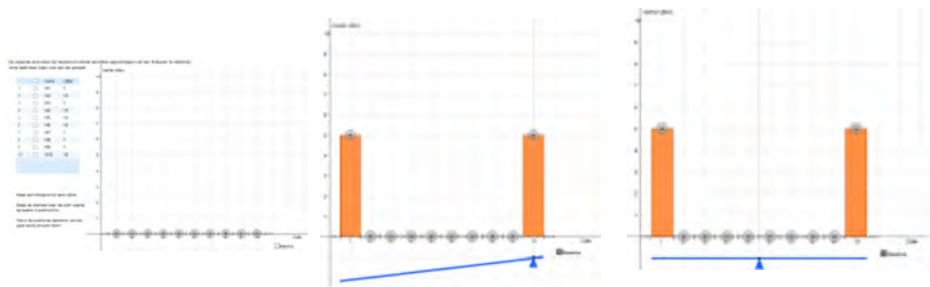
Note. Start screen with the context (left) and no bars, all bars pulled up and balance tool ticked (middle) and solution (right). Students are asked to first pull the bars up according to the table, then tick the balancing tool to find the balancing point.

A.2.6 Task 6

Figure A.6 Screen shots of task 6
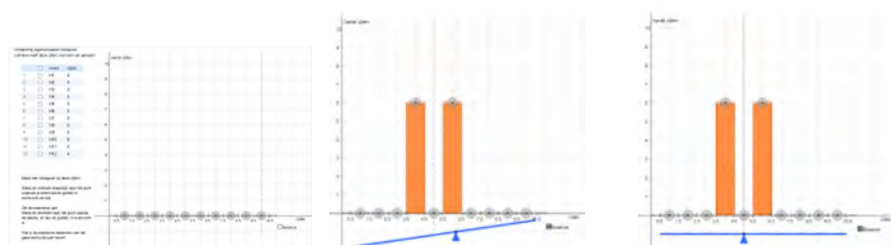
Note. Start screen (left) with no bars, bars pulled up and vertical line dragged to the estimated balance point (middle), finding the balancing point (right). In addition to the

previous task, students are now asked to drag the vertical line to their estimation of the balancing point.

A.2.7 Task 7

Figure A.7 Screen shots of task 7
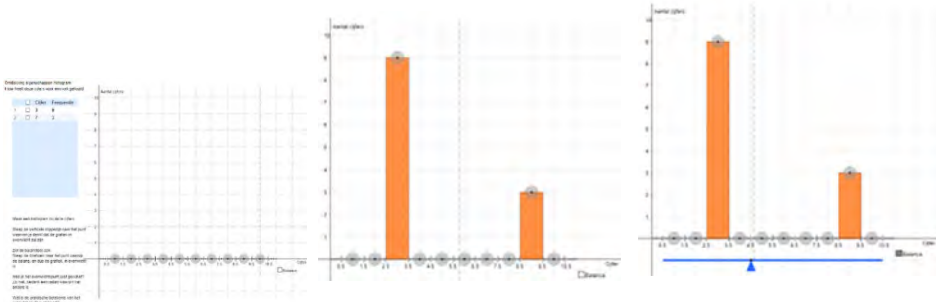
Note. the vertical dotted line at students drag to the left to indicate their estimation of the mean, before they check it with the balance tool.

A.2.8 Task 8

Figure A.8 Screen shots of task 8. Initial histogram (left) and histogram after alternately adding some 1s and 10s (right)
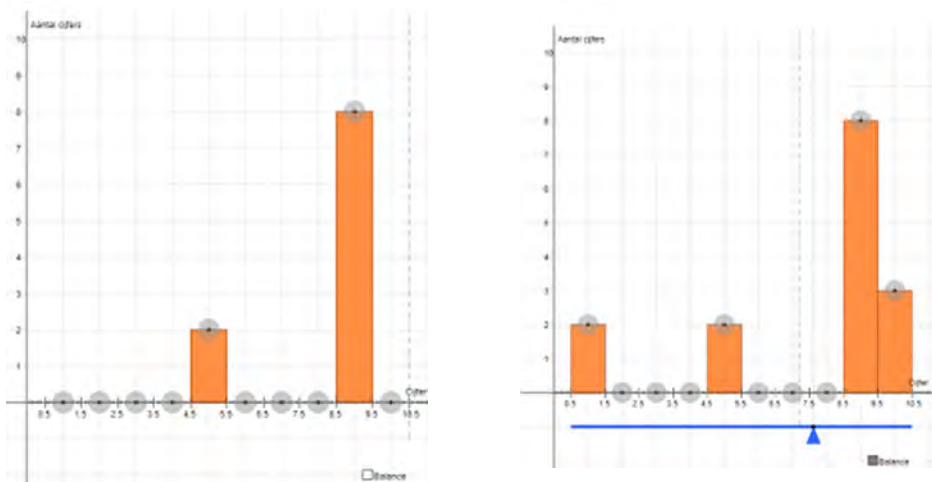
A.2.9 Task 9

Figure A.9 Screen shot of task 9. Possible histogram, estimation (vertical dotted line) and balance tool
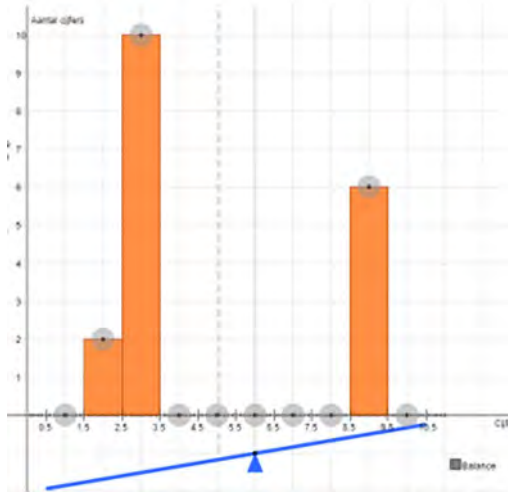
A.2.10 Task 10

Figure A.10 Screen shot of example histogram, with estimation (dotted line) and balance tool
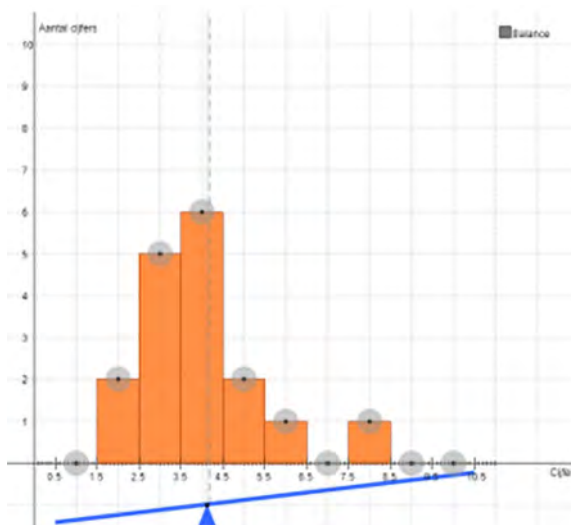
A.2.11 Task 18

The assignment was: Draw a histogram for the table below (use the attached grid paper) and then answer the questions. For some of the questions, readers are referred to the article.

Table A.1

Net annual income in thousands	Number of households in thousands
< -10, 0]	36
< 0, 10]	299
< 10, 20]	1690
< 20, 30]	2471
< 30, 40]	1822
< 40, 50]	845
< 50, 60]	327
< 60, 70]	134
< 70, 80]	66
< 80, 90]	36
< 90, 100]	22

A.2.12 Task 20

Enlarge this task to view it on your screen
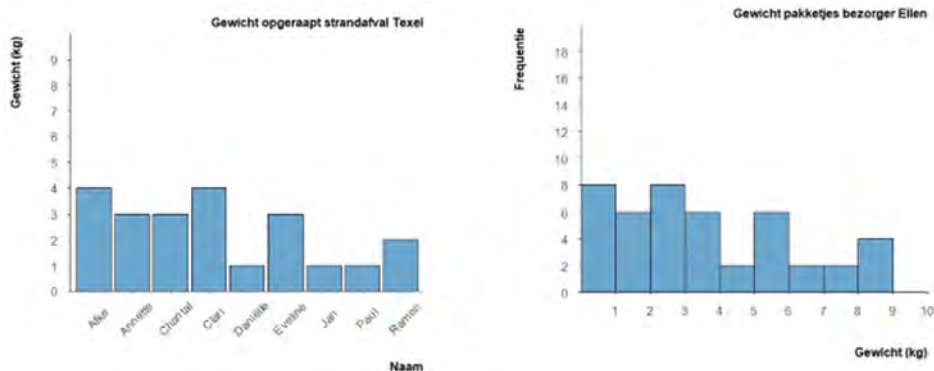
Both graphs show weights. On the left, you see the weight of beach waste collected by different students. On the right, you see the weight of packages delivered by postal worker Ellen. Which of the following statements about these graphs is true?

The average weight is larger in the
- graph on the left,
- the graph on the right,
- approximately the same in both graphs.

The variation in weight is larger in the
- graph on the left,
- graph on the right,
- approximately the same in both graphs.

Explain how you arrived at your answer and why you think this is so.
What is the weight of the packages in the most left bar in the graph on the right?
Why?

What is the weight of the garbage on the beach in the most left bar in the graph on the left?
Why?

Moving toward new tools for research and teaching statistics: General conclusions, discussion, and implications

"One never notices what has been done; one can only see what remains to be done." [61]
Marie Curie

This chapter is partly based on
Boels, L. (2023). Reflections on gaze data in statistics education. *Teaching Statistics,* 1–12*.* https://doi.org/10.1111/test.12340

#

7.1 Research aim and answer to main research question

The aim of this research was to contribute to an empirically grounded theory on how to teach statistical literacy through histograms. As explained in the introduction, we expected that a review of the literature and a small-scale eye-tracking study would both have been input for a larger design study (Bakker, 2018). However, the topic of this research turned out to be much tougher than initially expected, as histograms are used in numerous disciplines, for example, to present research outcomes. A search for 'histograms' in Google Scholar nowadays will lead to millions of hits. It is impossible to summarize all literature about how histograms are used in research and education and what is known about them. Moreover, from the review study (Chapter 2), it became clear that the few interventions in statistics education that had been reported were not very successful. Consequently, this provided few starting points for the design. Therefore, substantially more of what is known as "front-end" work (McKenney, as cited in Bakker, 2018, p. 142) proved necessary before a new approach to teaching histograms could be designed. This front-end work included better understanding students' conceptual difficulties with histograms through an eye-tracking study (Chapter 3), students' interpretations of dotplots (Chapter 5) as dotplots can draw students' attention to the variable being presented along the horizontal axis in both graphs, and formulating design criteria (Chapter 6).

In addition, what emerged was that students lacked experience with what and how data are represented in histograms. This suggested that students had insufficient embodied grounding. Given the successes of this approach in other mathematical topics, we applied an embodied instrumentation approach in the intervention for which we conducted the first cycle only (Chapter 6). In addition, the first eye-tracking study opened up future possibilities for the automatic identification of student strategies on histogram tasks (Chapter 4). We explored these opportunities by comparing an interpretable mathematical model (IMM) for which a machine learning algorithm (MLA) provided a baseline. Our revised overall research question was:

> RQ: How can pre-university track[62] students in Grades 10–12 be supported in understanding histograms?

We mostly concentrated on students with Mathematics A, as these students have statistics in their curriculum. One part of the answer to this question is that more attention to *the key concept of data* is needed, as many difficulties

---

[62] Pre-university track is 'vwo' in Dutch.

related to this key concept seem to underlie difficulties with the key concept of distribution (Chapter 2). This focus includes developing students' understanding of *data* in graphs of univariate data such as dotplots, stem-and-leaf plots, and histograms.

A second part of the answer to the main research question is a hypothetical learning trajectory (HLT) (Simon, 2020). We presented the first cycle of a future design study (Chapter 6). We formulated and tested design criteria that can be used in such a future study. Our HLT can be seen as a further step toward a domain-specific instructional theory on how to teach students to understand how data and their distribution are depicted in univariate graphs such as histograms, dot-, box-, stem-and-leaf, and hatplots (Konold, 2007) and histodots (Chapter 2).

As a third part of answering the main research question, we investigated whether it would be possible to identify students' task-specific strategies when estimating means from histograms. This could be a first step toward automatic feedback based on students' scanpath patterns on only the graph area of histograms in a future Intelligent Tutoring System. We showed that automatic identification is quite possible with a machine learning algorithm and an IMM (Chapter 4).

In the remainder of this chapter, we reflect on the study's scientific contributions (7.2) and methodological limitations and contributions of our work (7.3) and describe implications and recommendations for future research (7.4) and educational practice and design (7.5).

7.2 Scientific contributions

Our studies led to several scientific contributions. The most important ones concern an emphasis on the key concept of data and task-specific gaze patterns, a focus on attentional anchors, and an embodied instrumentation approach leading to a local instruction theory and theoretically and empirically underpinned task design guidelines. Below, we briefly discuss and elaborate on these contributions.

7.2.1 Emphasis on the key concept of data

In the review study, we speculated that the persistence of people's misinterpretations of histograms is partly due to overlooking the impact of data-related conceptual difficulties. We thought that this might also result in underreporting of misinterpretations regarding data-related conceptual difficulties, as well as misinterpretations regarding shape and center.

---

1. The key concept of data

According to literature (Erickson et al., 2019; Garfield & Ben-Zvi, 2004; Garfield & Ben-Zvi, 2008a; Gehrke et al., 2021; Gould, 2017; Ridgway et al., 2011), the key concept of data encompasses:

- The context of the data, including:
  - Need for data; why they were collected
  - Data as a representation of real-world phenomena
  - Who collected the data and how
- The different representations of data, including numbers, texts, pictures, how data representations in computers can vary, and summaries or aggregated forms of data such as graphs and tables.
- The characteristics of data, including the difference between a variable (e.g., weight) and data (e.g., numbers representing the measured weights), what the statistical variables are, and the measurement level.
- Why altering data is sometimes needed before analysis is possible, including data wrangling or moves such as:
  - Data cleaning, dealing with missing data or outliers
  - Merging data(sets)
  - Constructing new data based on existing data
  - Selecting or generating variables
  - Filtering, grouping, or ungrouping data
  - Aggregating or summarizing data

This is a broader concept of data than that in the GAISE II guidelines (Bargagliotti et al., 2020), in which didactical choices have been made. Data themselves are not an object but represent a phenomenon in the real world. We cannot think about data without thinking about their representation: numbers, tables, photos, graphs (cf. Bakker & Hoffmann, 2005; Gal, 1995). For graphical representations, the concept of data encompasses how data are represented in, for example, histograms, boxplots, and case-value plots, including (Chapters 2 and 3):

- How many statistical variables are depicted in the graph
- The measurement level of data (e.g., nominal, ordinal, interval, ratio)
- Along which axis the variable is presented

Contributing to statistics education research, we have, therefore, placed the statistical key concept of data (Box 1) more to the forefront in the rest of our studies. First, we did so by presenting students with several different graphical

representations of univariate data during our eye-tracking data collection: histograms, messy dotplots, stacked dotplots, and horizontal histograms (bars rotated 90 degrees clockwise). It would be interesting to pay more attention in future research to students' conceptual understanding of data, not only for data that fit the sample-population assumptions but also for contemporary data or data collected by others.

Second, we brought 'data' more to the forefront by contrasting histograms (univariate data) and case-value plots (bivariate data) during the eye-tracking data collection and some tasks in the intervention, confronting students with their confusion about these two types of graphs (e.g., Burrill, 2019; Cooper, 2018; Kaplan et al., 2014). In addition, in the intervention study, students were given different graphs of univariate data (cf. Bakker, 2004a), they were asked to interpret other students' gaze data using diverse strategies on a single histogram task, and they were asked to sort graphs with bars into two categories: histograms and 'other'.

Data and distribution are related. As we suspected that misinterpretations regarding shape[63] and center are underreported in literature, we addressed both in the eye-tracking data collection by providing histograms and case-value plots with different distributions—hence, also different shapes—and by asking students to *estimate* and *compare* arithmetic means. According to Gal (1995), asking students to estimate the arithmetic mean from graphs can reveal their conceptual understanding of the data. The mean can be regarded as a precursor for variability, as variation is often assessed compared to a measure of center (e.g., standard deviation is always calculated from the mean). Moreover, variation is barely taught in Grades 10–12, while the arithmetical mean is familiar to these students. Finally, we wanted to be sure that our findings were due to students' misinterpreting where and what *data* are depicted instead of misinterpretations of variability (part of the statistical concept of distribution, Box 2).

Discussing our contribution, we note that, parallel to our studies, research attention of the statistics education research community has shifted—partly driven by the emergence of new forms of data and 'big data'—to data literacy. Gould argues that data literacy is statistical literacy with more focus on being "a critical consumer of data, controlling [ones…] personal data trial, finding meaning in data, and taking action based on data" (2017, p. 23). Others regard data literacy as part of evidence literacy (e.g., LaPointe-McEwan et al., 2017). Evidence literacy includes the ability to "critically evaluate both the meaning and strength of evidence that are used to support the claims and arguments of experts and other commentators in the media" (Gal & Geiger,

---

[63] Possibly due to too much focus on the shape of distributions, regardless of the type of data.

2022, p. 19). Both data and evidence literacy seem to be encompassed by the critical stance in statistical literacy and focus on specific aspects of it. We embrace the recent awareness of the importance of new forms of data—as encompassed in data literacy—and suggest extending it to all existing forms of data given students' difficulties with the key concept of data.

---

2   The key concept of distribution

Distribution is a lens through which statisticians look at variations in data, setting aside individual cases (Wild, 2006). Wild describes that distributions can reveal patterns in the data (explained variation) ignoring random variation (unexplained variation called noise).

The key concept (big idea) of distribution of quantitative data encompasses center, variability, gaps, clusters, shape, and outliers (e.g., Garfield & Ben-Zvi, 2004), density, spread, and skewness (e.g., Bakker & Gravemeijer, 2004), relative frequency, probability, proportionality, and causality (e.g., Reading & Canada, 2011) but also the difference between an empirical versus a theoretical distribution and between a distribution of a sample, a population, and a sampling distribution (e.g., Reading & Canada, 2011; Wild, 2006). Variability includes pattern, variation, randomness, deviation, signal, noise, and range (e.g., Engel et al., 2008).

In addition, statistical confidence and significance depend on this concept of distribution (Reading & Canada, 2011). Theoretical distributions come with "considerations of 'robustness' and 'goodness of fit' [of] the data" (Wild, 2006, p. 13). Drawing graphs is important for considering variation (Pfannkuch & Reading, 2006). Distribution is an organizing conceptual entity to grasp the overall aggregate (Bakker & Gravemeijer, 2004).

7.2.2   Task-specific gaze patterns on statistical graph tasks

A substantial part of the research described in this dissertation involves eye-tracking (Chapters 3–5). A major advantage of eye-tracking is that it can make students' strategies visible at a much greater level of detail compared to, for example, thinking aloud (e.g., Kaakinen, 2021). In addition, it can make strategies visible that participants are unaware of or are unable to articulate. However, there is not a simple relation between gaze patterns in general and students' strategies for specific topics (e.g., Kok & Jarodzka, 2017). Moreover, not every gaze is part of a student's strategy (e.g., Schindler & Lilienthal, 2019). Therefore, research is needed to reveal how gaze patterns relate to students' task-specific strategies (e.g., Schindler et al., 2021). Therefore, a form of

triangulation, for example through cued recall, will be needed until clear patterns have been found for specific tasks and topics and in different communities.

Contributing to statistics education research, we revealed that specific perceptual forms of students' gaze patterns are related to specific strategies for estimating and comparing means from histograms and case-value plots for university students (Boels et al., 2018), teachers (Boels et al., 2019b), and high school students (Chapter 3) in the Netherlands. For example, in estimating means from single histograms (and case-value plots), a horizontal gaze pattern indicates a strategy for interpreting the graph at hand as if it were a case-value plot.

7.2.3 Interpreting gaze patterns as attentional anchors

Our contribution is that we theoretically interpreted the *perceptual form* of students' gaze patterns on the graph area (horizontal or vertical lines) of statistical graphs of univariate data as attentional anchors. As such, we elaborated and applied the notion of attentional anchors for the case of histograms and case-value plots. For this interpretation, we draw on insights from theories of enactivism and embodied cognition that suggest that cognition emerges from interaction with the environment (e.g., Davis et al., 1996; Rowlands, 2010). The focus of an actor's interaction with this environment is called an attentional anchor (Hutto et al., 2015; Hutto & Sánchez-García, 2015). An attentional anchor is "a real or imagined object, area, or other aspect or behavior [...] that emerges to facilitate motor-action coordination" (Abrahamson & Sánchez-García, 2016, p. 203). The ones we found in our research facilitated students' imagined actions (strategies for finding the mean)—regardless of whether these strategies were correct. For students, these attentional anchors were like imaginary lines. For example, they referred to these imaginary lines as "making all bars equal" and their eyes moved horizontally on the graph area. Other students showed vertical gazes and referred to a point on the horizontal axis "where the graph is in balance."

7.2.4 Application of embodied instrumentation in statistics education

Our research contributes to the theory of embodied instrumentation by showing how more complicated artifacts (e.g., histograms) can be reinvented from actions with simpler ones (e.g., positioning dots on a scale). Specifically new is the explicit attention during the design phase for building on artifacts that students are already familiar with (e.g., horizontal scale coming from previous experiences with the number scale and the cartesian grid) and that once constituted the actions that are nowadays consolidated in the to-be-acquired artifact (e.g., a histogram). Using these already familiar artifacts,

students reinvent artifacts that are new to them or that they did not fully grasp yet in previous schooling (e.g., a histogram, estimating the mean from a graph).

Our motivation for applying an embodied instrumentation perspective arose from previous research that demonstrated that several students still misinterpreted histograms even when they were talked "through the *data creation process*" and had been prepared through dotplots (Bakker, 2004a, p. 272). The literature (Chapter 2) offered little clue to appropriate interventions and, foremost, revealed persistent misinterpretations. Previous studies gave us the impression that students lacked experience with dotplots and sufficient attention to how artifacts—histograms, dotplots—become tools in statistical reasoning. We suspected that students' education might have lacked an embodied grounding of how histograms are constructed. Therefore, using an embodied instrumentation approach (Drijvers, 2019) as a theoretical lens, we designed a learning trajectory with students reinventing the role of the axes in dotplots and histograms through specific tasks and constraints in the software, as described in the intervention study in Chapter 6.

Discussing the contribution of that study, we note that tasks designed from an embodied *instrumentation* perspective are still rare. To the best of our knowledge, our study is the first within statistics education. Two types of task designs from an embodied cognition approach are currently described in the literature: an action-based design (for an overview of recent examples in mathematics education, see Alberto et al., 2022) and a perception-based design (e.g., Abrahamson, 2009). In an action-based design, students are confronted with a motor-control problem such as keeping a screen green while moving one or two points or bars with their hands. Through solving this motor-control problem, students develop new ways of moving that help them to understand a mathematical concept. In perception-based designs, students solve a perceptual problem. A new type of task is currently being developed: incorporation-based tasks.

For an incorporation-based task, students are first invited to solve a sensorimotor task with feedback from some artifacts (e.g., an action-based task) or observe perceptual qualities enabled by an artifact (e.g., a perception-based task), and are then invited to perform the same task without the artifact, just with their body (Bos et al., 2021, p. 4).

Our design is not an action-based design, as none of the tasks require motor-control problems to be solved. Students also do not use their bodies to solve the tasks and feedback was not imagined, therefore, we do not consider it an incorporation-based design, either. Whether it can be considered a perception-based design or a new design genre is currently under debate.

Although, in embodied designs, attentional anchors are usually

introduced as artifacts in tasks as a consolidation of students' actions, in hindsight, we might have done this too early for estimating the mean from a histogram. We infer that this introduction was too early for students as they did not have any struggle with finding the balance point for a histogram and linking this point to the mean. Note that in Dutch education, the mean as a balance point is absent from the mathematics curriculum for elementary and secondary education.

7.2.5 Task design guidelines from an embodied instrumentation perspective

We worked out a design framework that helps when designing specific mathematical and statistical topics. The generalization and value of design guidelines lie in the iterative process of letting the guidelines do the actual work (Bakker, 2018). Based on the empirical tryout (Chapter 6), we reformulated our theory-driven design guidelines:

- Find the actions—through a logical-historical reconstruction—that could have constituted the target artifact
- Design motor-control or perception tasks to which these actions are the answer
- Create productive struggle for crucial steps in the HLT
- Have students perform the (digital) actions with feedback
- Have students reflect on their actions
- Create possibilities for transfer by varying contexts and environments.

The guideline to create productive struggle was added as a separate guideline during the revision. Theoretically, this guideline means that the tasks that the students' emerging functional systems are solving (Shvarts et al., 2021) should be new enough for the learning process—rather than simple recollection—to happen. In addition, a functional system needs to be flexible and adaptable to various environments. Therefore, transfer is also desirable within each crucial step and not only at the end of learning. Unlike other researchers who work on general principles of enactive pedagogy (e.g., Abrahamson et al., 2021), our design guidelines are for specific topics.

7.2.6 Contributing to a local instruction theory on teaching and learning of histograms

Our HLT can be seen as a further step from a general theory of embodied instrumentation toward a *domain-specific instructional framework* on teaching how data and their distribution are depicted in univariate graphs such as histograms, dot, box, stem-and-leaf, as well as hatplots (Konold, 2007) and histodots (Chapters 2, 6). A hypothetical learning trajectory "consists of three

components, a learning goal, a set of learning tasks, and a hypothesized learning process" (Simon, 2020, p. 355). Besides the need for more design cycles to test and revise our HLT in practice, an HLT must always be adjusted to local circumstances (e.g., Barab & Kirshner, 2001).

From our proposed learning trajectory (Chapter 6), we would like to discuss two things. First, the importance of reflection during and after task completion, which is in line with insights from recent literature on embodied designs (Abrahamson et al., 2021; Alberto et al., 2022; Shvarts et al., 2021). The results of our eye movement research also seem to underline this importance, as we suspect that a part of the learners gained insights about a correct strategy as a result of reflection during the cued recall in which they explained which strategy they used (Chapter 5). Second, the "balance [model] is a critical mathematical characteristic of the mean" (Mokros & Russell, 1995, p. 33) and can be linked to the algorithm for finding the equilibrium (mean) of moments (forces times distance) in physics. Some students started to reason about the heights of the bars being somehow proportional to the distance to the mean. This was not foreseen. The literature suggests having students explore the characteristics of the mean and how it is affected by different types of datasets and distributions (Garfield & Ben-Zvi, 2008a). Garfield and Ben-Zvi suggest bringing students from seeing the mean as a process (a computation, algorithm) to an object, a signal in a noisy process. In this sense, the mean can be regarded as a precursor for assessing variability in the data (e.g., the standard deviation is compared to the arithmetic mean), and further study is needed to develop students' notions on this key concept as part of the key concept of distribution.

The introduction (Chapter 1) pointed out that descriptive statistics, such as the mean, provide limited information—for example, when comparing groups—due to several factors, including variability in the data. This is one of the reasons why it is important to jointly examine measures of center and variability (cf. Shaughnessy, 2007; Bargagliotti et al., 2020). Variability or "spread are connected to 'spread around what'—[with the what being] some value indicating a measure of center" (Burrill, 2019, p. 133). Nevertheless, we focused in most of our studies on measures of center only. One reason for this is captured in the above statement that variability is always a variation relative to something, often being the mean. If students do not understand how to estimate the mean from a histogram, it is likely that they cannot assess variability either. Another reason is that variability is not part of the Dutch curriculum, except for some technical skills such as using a calculator to calculate a standard deviation.

Although our HLT also provided opportunities to work toward the algorithm for the mean (Chapter 6), such a route does not seem desirable for most students (e.g., with Mathematics A or C). Connecting the algorithm to a histogram is an example that is likely to fall under what Ridgway and Nicholson describe as a practice that "harks back to the days when calculations had to be done by hand with the result that students are required to learn techniques that are always automated in professional work" (2019, p. 2). In addition, such a route might even hinder the development of students understanding of the concept of center (mean) (cf. Tall & Vinner, 1981).

7.3   Methodological contributions

We applied three methods that are relatively new in statistics education research: eye-tracking, a transparent interpretable model, and MLAs. Contributing to eye-tracking research, we show how spatial eye-tracking measures can reveal task-specific strategies. Contributing to statistics education research, we discuss how several visualizations of gaze data, spatial eye-tracking measures, and application of MLAs and an interpretable model can be used for analyzing and classifying students' task-specific strategies. Contributing to the discipline of software design, we discuss insights gained by applying an embodied instrumentation approach. We finish with methodological limitations.

7.3.1   Spatial eye-tracking measures as a means to reveal task-specific strategies

A methodological contribution is that spatial measures can reveal task-specific strategies. After watching videos of students' gaze behaviors in more than 600 trials (with histograms or case-value plots)—combined with students' cued recall data—we discovered, in a qualitative study, (Chapter 3) that the *perceptual form* of students' gaze patterns within one AOI—the graph area— was relevant for students' task-specific strategies on these items. In later chapters, we used saccadic magnitude and direction (vectors) for this aim (Chapters 4 and 5), ignoring the also necessary alignment. Although vectors have been used for some time (e.g., Holmqvist et al., 2011), their use in education is rare. In addition, we applied them in a new way. When we conducted our study in 2018, we knew of only one other empirical study that also used vectors for educational purposes (Dewhurst et al., 2012). However, in that study, scanpath similarities are compared and related to task difficulty instead of inferring students' strategies from scanpaths as we did. For other research questions, for example, on general task performance, the 2012 study

used time measures. In contrast to Dewhurst et al., we did not use the position of the vector on the screen.

Discussing this contribution, we note that, although eye-tracking has been around for some time, its use in statistics education is still in its infancy. Most recent eye-tracking studies, also in statistics education, use quantitative gaze measures such as total time spent on an area of interest (AOI) (e.g., Cohen & Staub, 2015; Fleig et al., 2017) or number of fixations within an AOI (Schreiter & Vogel, 2023), although the latter also used vectors. Fixations are the positions on the screen that students looked at. Quantitative measures are used in cognitive sciences that usually aim for more general strategies such as self-regulated learning. These measures are temporal (e.g., total fixation duration or dwell time[64], reaction times, time to first fixation, total reading time), counts (fixation count, number of saccades between relevant or irrelevant parts of the stimuli; Godau et al., 2014), or both (e.g., Kaakinen, 2021; King et al., 2019; Lai et al., 2013). The advantage of these measures is that they can be computed easily. However, quantitative measures can hide visual scanning patterns (Goldberg & Helfman, 2010).

Spatial measures (e.g., scanpath, fixation position) can disclose the kind of detailed information Kaakinen (2021) refers to. Spatial measures seem better suited for providing detailed information about students' thinking processes (Hyönä, 2010; Schreiter & Vogel, 2023). However, spatial measures are still quite uncommon in eye-tracking research:

> Parallel to the findings of Lai et al.'s (2013) study, spatial measurements were the least common measure in the reviewed studies. Spatial scale comprises fixation positions, fixation sequence, and scan path patterns (Lai et al., 2013). It requires mostly a qualitative analysis of the scan paths to obtain these measurements. Although the scan path analyses reveal how learning occurs from moment to moment (Hyönä, 2010), few studies investigate them in detail (Krejtz, Duchowski, Krejtz, Kopacz, & Chrząstowski-Wachtel, 2016). This could be due to the difficulty in qualitative analysis and synthesis of the scan paths obtained from different participants. (Alemdag & Cagiltay, 2018, p. 419)

Moreover, when people refer to scanpaths, they usually mean a sequence of or transitions between AOIs (e.g., Garcia Moreno-Esteva et al., 2018, 2020;

---

[64] Orquin and Holmqvist (2017) suggest no longer using dwell time—total fixation duration—as other measures may be more suitable for measuring the different constructs underlying dwell time.

Krejtz et al., 2016). For students' solution strategies, both from the qualitative study (Chapter 3) and the quantitative machine learning analysis (Chapter 4), it appeared not to be relevant when the horizontal gaze pattern we found was a bit higher or lower on the graph area, if a student looked from left to right or vice versa, or if there was a slight slope in this gaze pattern or not, as long as the pattern was mainly horizontal. The irrelevance of such specific order and position on the screen has a potential advantage for future webcam usage. Webcams are still imprecise in their calibration. If a horizontal or vertical shift of a horizontal gaze pattern still results in a horizontal gaze pattern on the graph area, this might still be recognized by an MLA or interpretable model.

Schreiter and Vogel (2023) may seem to have used the same features in the same way as we did: vectors (saccadic magnitude and directions) within two AOIs for comparing graphs (they used one AOI for each dotplot including the axis; transitions between AOIs were excluded). However, they used the magnitude of saccades to distinguish between short (local) and long (global) viewing, and similarly used saccadic direction (horizontal for global; vertical for local). We used these same features for distinguishing students' task-specific strategies (e.g., a histogram interpretation strategy). For our single and double graph items, our approach reveals students' strategies and whether they interpreted the graphs correctly or incorrectly. In the approach of Schreiter and Vogel, a correct and incorrect strategy could both be classified as local. In addition, for our single items, interpreting vertical saccades as indicative of a local view would not make sense as that would be a correct (global) strategy when estimating a mean from a histogram.

7.3.2 Tools for analyzing eye-tracking data: heatmap, raw data, videos, and static gazeplots

Contributing to statistics education research, in the studies in this dissertation, we showed how raw gaze data (Chapters 4 and 5) and videos and—to a lesser extent--static gazeplots (Chapter 3) can be used for analyzing gaze data to reveal students' strategies. In addition, in a pilot study, we used heatmaps (Boels et al., 2018). Here, we reflect on our contribution and provide some guidance for methodological choices in future research.

In our studies, we analyzed the collected gaze data in two ways: a qualitative analysis of the videos of the gazes on fifteen tasks (see Chapters 3, 4, and 5) and quantitative analyses using the raw data of, in total, seven of these tasks (five single histogram tasks and two double histograms tasks, Chapters 4 and 5). For these analyses, we used two types of data obtained through *data moves* (Erikson et al., 2019) either from the Tobii Studio software (n.d.-a) or by us. In the quantitative studies, we used 'raw' data that consist of *x*- and *y*-coordinates of the gazes on the graph area only (sampled with 60 Hz,

Figure 7.1 left); fixations are indicated by short lines in a star-like form. For the qualitative analysis, we found videos of the gazes to be the best approach (Figure 7.1 middle left; also called 'dynamic gaze visualizations' e.g., Kok et al., 2017). In this way, we could see the order of the gazes and what students paid attention to. As previously discussed, our attention was mostly on the saccades—the thin, long lines between fixations (the latter indicated by circles in the gazeplots).

Other possible data moves (all created with the Tobii software) are heatmaps (Figure 7.1, middle) and a static gazeplot (middle right and right). Changing representations of gaze data is part of transnumeration (Wild & Pfannkuch, 1999) and can help with understanding what gaze data can tell.

Figure 7.1 Examples of different ways to visualize and use gaze data
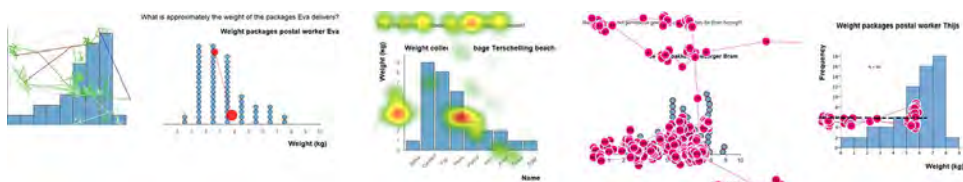
Note. Raw scanpath (left), video still of smoothed gaze pattern (middle left), heat map (middle), gaze plot (middle right), and scanpath in gaze plot (right) indicating an imaginary horizontal line (here superimposed for the reader).

Heatmaps have the advantage that they aggregate the gaze data and draw attention to the fixations (for example, Schindler et al., 2021) but have the disadvantage that time and spatial information gets lost (e.g., the order of the fixations or saccades). Fixations on locations where the student spent little time are green, and coloring goes to yellow and then red when more time is spent in total.

Static gazeplots (Figure 7.1, middle right) have the advantage that both fixations and saccades are shown but the pattern and item can get hidden behind the fixations. The most relevant part of the pattern can be isolated (Figure 7.1, right) but requires days of manual work as this needs to be done for every student for every item separately and requires the judgment of what belongs to this pattern and what not. Future research is needed to find out if it is possible to infer our students' strategies from heatmaps or static gazeplots in qualitative and quantitative analysis (e.g., through MLAs, Schindler et al., 2021).

7.3.3   Machine learning algorithms application in statistics education

We used our gaze data in combination with machine learning algorithms (MLAs) in Chapters 4 and 5. Contributing to the application of data sciences tools in statistics education, we showed how gaze data can be used for task-

specific strategy identification (classification) on tasks with statistical graphs (Chapter 4) and for finding strategy-relevant differences in gaze data between two similar tasks (Chapter 5). Moreover, we showed that despite the idiosyncrasy of gaze patterns, general gaze patterns on tasks can emerge that are relevant for teaching *specific* content or a topic. We showed that qualitatively found patterns (Chapter 3) can be captured by MLAs and used for identification of similar patterns in gazes of new students, and on new, similar tasks (Chapters 4, 5). Similarly, we showed that an interpretable mathematical model (IMM, Chapter 4) can be used that is transparent on how it came to its decision for an individual student. Furthermore, we showed the importance of using expert (teacher, researchers) information—together with students' cued verbal reports—on what part of the gaze pattern is relevant for students' task-specific strategies (namely, the pattern on the graph area).

Discussing the MLA application, we note that when applying a machine learning algorithm (MLA) in an educational context, the focus can be either on the educational application (Chapter 4) or on tailoring the MLA to the educational context (Chapter 5).

7.3.4 Software for developing statistical literacy—an embodied instrumentation perspective

Teachers can choose the software they use in their classrooms. Often, a distinction is made between software for *doing* statistics (InZight[65], Minitab[66], SPSS[67], R and RStudio[68]), and software for *learning to reason* in statistics education (Fathom[69], CODAP[70], TinkerPlots[71], VUStat[72]). From an embodied instrumentation perspective, we want to add software for *initial* learning.

Software designed from embodied instrumentation design principles
A contribution of our study is that it makes it plausible that software designers need to think carefully about what kind of actions (crystallized in artifacts) they outsource to the software for *initial* learning. For example, most software can create histograms but lacks possibilities for students to reinvent them. Moreover, the way graphs are presented provokes different comparisons. Side-by-side comparison elicits the (un)equal heights of bars or dotplots (and seems nonsensical for horizontal boxplots), whereas displaying graphs above

---

[65] https://www.stat.auckland.ac.nz/~wild/iNZight/
[66] https://www.minitab.com/en-us/
[67] https://www.ibm.com/products/spss-statistics
[68] https://posit.co/products/open-source/rstudio/
[69] https://Fathom.concord.org/
[70] https://CODAP.concord.org/
[71] http://tinkerplots.com/
[72] https://www.vustat.eu/apps/

each other elicits the comparison of the positions of the data along the horizontal axis (which seems more relevant for dotplots, horizontal boxplots, and histograms). Software designers often have already chosen the position of the graphs. Perhaps they might include an option for students to drag two graphs to a position that students find suitable for comparing the graphs. This is not to say that such an option is better than given positions, but only that when actions are outsourced to the software too early, students might never become aware of their relevance until they fail in practice. Moreover, passive tasks should be avoided as then "students need just [to] gaze at technological elements (no coordination required)" (Alberto et al., 2022, p. 18). This includes "readymade examples (students will just imitate them)" (p.18). In addition, when "elements of problem solving [are outsourced] to the technology […] students will ignore them" (p. 18).

Furthermore, as tasks designed from an embodied instrumentation perspective have students reinvent artifacts before these are incorporated into the system in a later stage, this seems to require route-type software tools (Bakker, 2002) that are designed for a particular learning trajectory. Therefore, we conjecture that for *initial learning*, route-type software will prevail over landscape-type software. A counterargument may be that learning is not linear and that routes will differ for various students. In such cases, a tree structure with branches for some routes might be more appropriate. In addition, from our experience, it is easier for teachers and students who are not so familiar with the software to use software with limited possibilities, as there usually is little time in classrooms to learn the software (cf. Bakker, 2004a; Van Dijke-Droogers, 2021). An online tool that has been built as a landscape tool, but with the advantage of limited possibilities in each part, is VUstat, as this has different apps to fulfill different teaching aims. We hope that the insights from the present research will inspire software developers to think about incorporating ideas from our learning trajectory as well as from other embodied designs.

Discussing these insights, in the section that follows, we describe some important characteristics of software tools from an embodied design perspective. We neither review all software (e.g., Abbasnasab Sardareh et al., 2021; Biehler et al., 2013; Chance et al., 2007; McNamara, 2016, 2018) nor compare in-depth Fathom, CODAP, and TinkerPlots (Frischemeier et al., 2023). Instead, we compare our design with that of others on *who* does the statistical thinking, in line with our fourth design guideline for embodied instrumentation design: *Have students perform the (digital) actions with feedback*. The purpose of this comparison is to provide the reader with an understanding of some differences between tools designed from an embodied instrumentation

perspective and those designed from a static and dynamic visualization perspective. We concentrate on graph construction and interpretation.

Static and dynamic visualization tools

In static visualization tools, students press a button and then observe the result of a hidden, statistical action. Minitools (Bakker, 2004a), InZight, and Minitab are examples of such tools for *doing* statistics "where the computer package is treated as a black box. After parameters have been entered, the outcomes are immediately shown, leaving the user [for example] to imagine the process of building the sampling distribution" (Meletiou-Mavrotheris, 2003, p. 269).

In much software literature, the term 'dynamic' actually means coupled (e.g., Frischemeier et al., 2023): changing data in one visualization (e.g., data card) affects other representations in real time (e.g., dotplot). We use a slightly different meaning of *dynamic* here: when a student manipulates something in a representation (clicks an option, moves a slider, drags a point), the student sees something happening—such as the transition to another representation—or a trace of the changes, for example, in the average (see also Wei et al., 2022). In CODAP, for example, students see the dots move after they dragged a variable (attribute) to the horizontal axis. Both in statistic and dynamic visualization tools, students see the consequences (results) of their actions, in line with the "action/consequence principle" (Burrill, 2019, p. 128). In dynamic visualization tools the process toward the consequence is visible, in static tools it is not.

TinkerPlots, Fathom, CODAP, and GeoGebra[73] are examples of dynamic software tools that mostly are (or can be) dynamic in this sense. Some statistical processes are still hidden actions, similar to those in statistic visualization tools, (e.g., calculation of the mean in CODAP), but the overall design principles are based on our definition of dynamic, although the developers use the word 'dynamic' also in the sense of 'coupled' (e.g., Finzer, 2006). There is also software that has a mixture of static and dynamic features. VUstat consists of several educational web-based apps that contain static (e.g., data analysis) or dynamic visualizations (e.g., sampling distribution). Similar to the minitools design, VUstat has the possibility to have a histogram overlay the dotplot in the data analysis tool (a variant of a histodot, see Chapter 2) as well as a boxplot.

Dynamic visualizations once were a major step forward. However, in contrast to embodied designs, students' actions in dynamic visualization software are (most often) not directly coupled with statistical actions, which

---

[73] https://www.geogebra.org/

might hinder statistical thinking and conceptualization. For example, sliding a dot to the right side of the screen (in TinkerPlots to make more bins) is not an action that places this dot on its correct value on the horizontal scale. The software does this for you for all dots. The same holds true for the binning in histograms. Although students can slide the border of a bin in a histogram, the adjustments (for all other bins as well as the height of the bars) are done (and shown) by the software. This is not to say that dynamic visualizations lack value for education. However, for persistent problems, such as students' conceptual difficulties with histograms, it is worth looking at who is doing the statistics: the student or the software. In embodied instrumentation design, *students* are doing the statistics.

7.3.5   Methodological limitations

A limitation of our review study is that a geographical selection bias seems to exist, as many studies were conducted in the United States and Europe. This might partly be due to the language (English) and the absence of or only recent attention to statistics in many countries (e.g., Burrill & Ben-Zvi, 2019). We expect that the misinterpretations we found will also hold for Asian, African, Latin-American, and Oceanic peoples. For example, Malaysian 7th and 10th graders had similar difficulties (Ismail & Chan, 2015; Lim et al., 2022; Saidi & Siew, 2019).

A limitation of our eye-tracking data collection is the sample size of 50 students. Although this is considered quite large for an eye-tracking study, for applying statistical tests (Chapter 5), it is considered relatively small. Limiting the generalizability of our findings about strategies (Chapter 3) is also that we mainly tracked the gazes of 10–12 graders—although similar scanpath patterns were found for university students (Boels et al., 2018) and STEM teachers (Boels et al., 2019b)—and that our sample consisted of Dutch students from only one school.

For the MLA application, we showed that both an 'off the shelf' tool (random forest implemented in the Mathematica Classify Function, Chapter 4) and a tailored MLA (random forest, Chapter 5) worked in classifying which students used what strategy (Chapter 4) and whether differences in gaze data occurred on histogram items after solving a series of dotplots items (Chapter 5). A limitation is that we mainly used only single histogram items for this aim (Chapter 4), although we used one pair of double histogram items in the second study that used an MLA (Chapter 5). A study using double histogram items in a similar way to the single histogram items with an MLA is foreseen. We consider our three eye-tracking studies as first steps toward uncovering students' task-specific scanpath patterns (Chapter 3), the possible application of gaze data in a future intelligent tutoring system in statistics education

(Chapters 4 and 5), and revealing micro-level changes in students' strategies that could point at learning—depending on how learning is defined—during homework or assessment (Chapter 5).

As described (Chapter 6), our lesson materials were tried out in a first cycle of design, implementation, and evaluation as a first step toward a design study. As this first cycle was tried out in a multiple case study, the small number of students in this multiple case study and their varying previous experiences with histograms is a limiting factor. Moreover, in most classrooms, there will not be time for spending 4–5 lessons (3.5 hours) on 'one' topic. On the one hand, a reduction in time for a next cycle of the HLT can be achieved by taking out problems that did not work yet, and, therefore, consumed relatively much of the teaching time (e.g., about 20 minutes for one task). Moreover, when applied in practice, the pretest will most likely be left out. On the other hand, when applied in schools, more time will be needed as starting and closing lessons consumes time. In addition, our mini-interviews might have made students reflect on their work, which could have speeded up subsequent tasks. Future design cycles need not only concentrate on how to further develop students' interpretation of data and distribution in histograms but also on what is minimally necessary for such development.

7.4   Implications and recommendations for future research

In the next sections, we offer implications and recommendations for statistics education research on the need for interventions and design research, how gaze data could be used, and on how an IMM, MLAs, and an embodied instrumentation approach could be applied. We end with some implications for eye-tracking research regarding the application of quantitative and spatial eye-tracking measures.

7.4.1   Need for interventions and design research instead of mostly surveys

An implication of our work is that more research discussing interventions is needed. When conducting our literature review (Chapter 2), we noticed that only a few studies reported interventions and even fewer of them explicitly discussed what such interventions should look like. This finding still stands since we completed our review study (e.g., Amaro & Sánchez, 2019; Burrill, 2019; Rodríguez-Muñiz et al., 2022). Most recent publications *assessed* students' conceptions (e.g., Cooper, 2018; Reinhart et al., 2022; Setiawan & Sukoco, 2021) or reviewed the literature on ensemble perception, visualizations, and statistics education with histograms as an example (Cui & Liu, 2021). One study only described a possible intervention without applying it

(Delport, 2020). However, intervention studies got more attention: "As seen in this volume […], research has now moved to trying different strategies and inventions that can help address these misconceptions. The hope is that these potential strategies will be further researched in different contexts and other countries" (Franklin, 2019, p. v).

Our proposed learning trajectory can be considered an answer to this call. Furthermore, the trajectory can help researchers to understand, for example, why "students mistake the bar heights in a histogram as the observed values in a dataset, and the number of bars as the number of observations" (Reinhart et al., 2022, p. 107).

Research on students' histogram misinterpretations is hard to find and spread across several disciplines. Our overview may assist researchers during the design stage of interventions to anticipate students' interpretations. In addition, it may assist them in developing new teaching materials that address misinterpretations more broadly—as different manifestations of the same underlying conceptual problem—rather than treating or remedying them one at a time.

7.4.2 Gaze data, an interpretable model, and MLAs as research tools in statistics education

Eye-tracking: lessons learned and warning
We are quite excited about what gaze data can tell. However, before continuing to possible future applications in the next sections, we would like to highlight two points of attention for those who want to start with eye-tracking research. First, the substantial time investment it takes to initiate such research as an early researcher—in our case, roughly nine months full-time for preparation and data collection and then over half a year for qualitative data analysis. Second, the already mentioned necessity to combine data, preferably through a cued retrospective think-aloud protocol (own perspective, McIntyre et al., 2022; Van Gog et al., 2005) as time on task and eye movements can be influenced by concurrent thinking aloud (Van Gog & Jarodzka, 2013). Researchers often want to know what students are paying attention to. Posing questions during an intervention or experiment can shift students' attention from where they were at that moment to what they think the researcher is asking for. Viewing patterns can potentially provide similar insight into what students pay attention to without disrupting students' thought processes.

Other lessons from reviews of eye-tracking studies in other fields (e.g., finance, communication) provided several insights relevant to eye-tracking research in statistics education. First, research on task-specific strategies is rare, as most studies in education contribute to general theories such as on information processing or multimedia learning (e.g., Beach & McConnel, 2019;

King et al., 2019). Second, eye-tracking can be used for observing changes in students' strategies that may point at learning as changes in gaze behavior occur during learning (e.g., Ashraf et al., 2018). More lessons for newcomers to eye-tracking research are provided in Appendix A of Chapter 3.

Students' gaze patterns on other graphs or tasks

An implication of our research is that scanpaths can potentially shed new light on tenacious didactical problems in mathematics teaching for other misinterpreted graphs, including boxplots (Lem et al., 2013b, 2014a), density curves (Batanero et al., 2004), stacked dotplots (Lyford, 2017), function graphs (Leinhardt et al., 1990), interpreting the slope or direction field when learning to solve differential equations (for an example in physics education, Klein et al., 2018), scatterplots (Estepa & Batanero, 1996), and violin plots (a kind of density plots), but also for other topics where scanpaths may play an important role: increase diagrams, frequency polygons, network topologies, line and point symmetry in functions, and the relation between a straight line, functions, and axis scales (logarithmic, linear, normally distributed). In addition, this could also hold for other mathematical topics, including the congruency of triangles, and maybe even the representation of a cube and hexagon.

Future research is needed to find out if the scanpath patterns we found for estimating and comparing means from histograms and case-value plots are similar in different cultural settings and educational systems around the world. Also left for future research is the analysis of some of the other tasks on which we collected data. As data were collected on 25 tasks, there are still ten tasks left that could be analyzed qualitatively and if successful, analyzed through an IMM and MLAs as done in Chapter 4: six messy dotplot tasks, two stacked dotplot tasks, and two horizontal histogram tasks. From students' answers, we suspect that students had no difficulties with single dotplots (Boels & Van Dooren, 2023) but that comparing arithmetic means from two dotplots was more difficult. When looking at students' gaze data during the assessment, we got the impression that some students used the heights of the stacks in stacked dotplots. We also got the impression that at least one student used the height of the 'bump' in a messy dotplot to estimate the arithmetic mean. Future research is needed to investigate students' strategies for solving the dotplot tasks. For example, is the gaze pattern on dotplots similar to the gaze patterns found for histograms and case-value plots? In hindsight, our messy dotplots might not be high enough to induce large vertical eye movements, which are typical for one correct strategy for estimating the mean from histograms. For future research, it is advised to make sure that the highest point in the graph is at least 200 (4.1°), preferably 250 (5.1°) pixels away from

the horizontal axis. Students' gaze data when solving dotplot tasks could possibly shed further light on why the research with dotplots had mixed results (e.g., Bakker, 2004a; delMas & Liu, 2005; Lem et al., 2013a; Lyford, 2017) and on how dotplots could be used in educational designs so that students benefit from their use. In addition, training an MLA with double histogram tasks is left for future research.

Revealing productive strategies

An implication of our research is that gaze data can potentially reveal correct reasoning. Many studies infer students' reasoning toward their answers from students' answers (e.g., Bolch & Jacobbe, 2019; Lovett & Lee, 2019). However, students could be using a productive strategy for solving the task at hand and still answer incorrectly, or vice versa. As correct reasoning is valued in statistics education, this could provide researchers with a new tool to discover such correct reasoning. From our experience, it can sometimes even be possible to infer from gaze data that a student started with a correct or incorrect strategy that was abandoned for some reason.

A future line of research could be to see how gaze patterns change within and over tasks, for example, when students develop a sense of a topic (e.g., Schindler & Lilienthal, 2020). Research suggests that a combination of students' actions, perceptions, and reflections results in a change in gaze patterns (Abrahamson et al., 2015; Alberto et al., 2022). A delay between a change in gaze pattern and students' verbal reflections can also indicate readiness for learning (Church & Goldin-Meadow, 1986). Vygotsky associated specific eye movements with thinking processes (1926/1997). Incongruencies between gaze and verbal data may be an indication of approaching or getting into the zone of proximal development (see also Chapters 3 and 5) where collaboration with a more knowledgeable person leads to joint actions and mutual understanding (e.g., Shvarts & Abrahamson, 2019). In one study, we also found indications of changes in gaze patterns that, combined with other data, suggested changes in strategies that could point at a learning effect of solving dotplot tasks (Chapter 5). A related future line of research is mismatches between gazes, gestures (e.g., students' hand movements on a screen or in a discussion), and speech. In addition, differences in gaze patterns between novices and experts could be studied (cf. Brunyé et al., 2019; Khalil, 2005; Van Marlen et al., 2022). Although we collected some data from experts for our tasks, we need to collect more data to investigate differences and guarantee anonymity. In addition, the gazes of students, teachers, and experts could be combined both in a qualitative study and in a machine learning analysis. Along with supervised MLAs, unsupervised MLAs can be considered, such as clustering or a combination of both. Furthermore, results from

machine learning analysis could be used to validate qualitative eye-tracking studies.

Further possible research directions

A future line of research could be to find out if students can be grouped meaningfully purely based on their scanpaths on the graph area through unsupervised clustering (an MLA) or latent class or profile analysis (e.g., Hickendorff et al., 2018). Another possible line of research could be the joint attention of teacher-student or student-student pairs (Chisari et al., 2020; Shvarts & Abrahamson, 2019). Gaze data from such pairs wearing glasses can show whether the gazes of a pair have a joint focus (e.g., Shvarts, 2018).

Future research could also be to investigate how the process of interpreting their own or other students' gaze data can help students' reasoning with data, data representations, center, variability, and so on (see Figure 7.2 for green fixations of a student interpreting the red fixations and saccades of another student who incorrectly estimated the mean weight from a histogram).

Figure 7.2 Video stills of one student's gazes (using glasses; green dots) looking at another student's gazes (red dots)
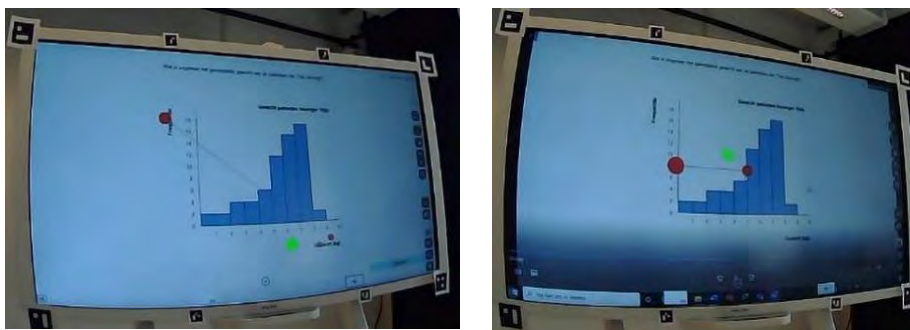
7.4.3 An embodied instrumentation approach in statistics education research

An implication of our research is that it is a first step toward a *domain-specific instructional framework* as described in section 7.2.6. We showed how an embodied instrumentation approach can be used to design tasks and develop an HLT which was evaluated in a first cycle of a design study (Bakker, 2018; McKenney & Reeves, 2012). An implication is also that we call researchers to question all pre-given aspects of the artifacts they use (e.g., height in histograms) and to reveal artifacts' origins (see Chapter 6).

Left for future research are follow-up cycles of scaling up in size (number of students, teachers, schools involved) and (re)design,

implementation, and evaluation. Some suggestions for redesign can be found in Chapter 6. A new cycle could work toward understanding variability from univariate graphs or toward the algorithm for finding the arithmetic mean from histograms. These are probably two different routes between which a similar tension may exist as between an uncertainty-based approach within statistics and a deterministic approach in mathematics (delMas, 2004; Meletiou-Mavrotheris & Stylianou, 2004; Ridgway & Nicholson, 2019). Since most students will become *consumers* of data and statistical models—including graphs—the first route seems to fit them better.

7.4.4   Quantitative and spatial measures in eye-tracking research

An implication of our research is that spatial measures (saccadic magnitude, direction, and for some strategies also alignment) do seem to provide task-specific guidance for developing a local instruction theory for learning or teaching a specific topic. Other successful attempts with sequences of AOIs within statistics and biology education show that *compressed* scanpaths are meaningful to researchers (Garcia Moreno-Esteva et al., 2018, 2020; see also our IMM in Chapter 4), and might, therefore, be more informative for educational aims than each single student's scanpath. A possible implication of our work is that a specific—*uncompressed*—order of AOIs and scanpath similarity or idiosyncrasy may be less important for uncovering task-specific solution strategies. The decision on how, when, and what spatial measures are relevant for task-specific strategies in statistics education is also left for future research.

Another implication for eye-tracking research is that, so far, quantitative measures in eye-tracking research do not seem to provide task-specific guidance for the teaching of a specific statistics education topic. Left for future research is the question of which of the quantitative metrics (if any)—including temporal metrics and counts—are relevant for statistics education research.

Scanpaths are idiosyncratic (Noton & Stark, 1971) and several studies found that "an individual's scanpath was […] more similar within an individual than between individuals" (Anderson et al., 2015, p. 1378). At first glance, our research seems to contradict this, as we look at similarities *between* individuals. Let us state up front that we do not dispute that scanpaths are idiosyncratic. However, there are several reasons why this idiosyncrasy can be ignored (or compressed, Garcia Moreno-Esteva et al., 2020). First, for students' solution strategies, we are not interested in differences that are irrelevant to those strategies, such as, for example, whether a student first looked at the graph title, and then at the graph area or vice versa. Second, for students' solution strategies, our qualitative study revealed that only the gaze pattern on

the graph area is relevant. Third, our students were not just looking at a scene but were looking to find an answer to the question posed. This is different from most studies in, for example, the review of Anderson et al. Although scene viewing can also be guided by the question to later describe the picture from memory (e.g., Johansson et al., 2006), such a question usually is more general than our question to estimate or compare means from the graphs. Fourth, if there were no similarities in students' scanpath patterns, it would not have been possible to train an MLA or make an IMM to find such similarities. We proved that this was possible with relatively high accuracy for single histograms.

In addition, an implication of our work could be that reading axes and graph titles is less important for students' task-specific strategies than we thought. A first indication stems from our first attempts with machine learning algorithms (MLAs)—not further reported in Chapters 4 and 5. Adding gaze data on graph and axes titles seemed to add noise and reduce MLA accuracy. Second, we saw in the videos of the gazes that students sometimes explicitly checked the axes titles and then still misinterpreted the graph. More research is needed to figure out whether—and if so, how—looking at axes titles is related to students' task-specific strategies.

7.5   Implications and recommendations for educational practice and design

In the next sections, we offer implications and recommendations for the role of histograms in the statistics education curriculum for teacher professional development and for future applications of eye-tracking in statistics education. In line with the quote of Marie Curie at the beginning of this chapter, we notice only what remains to be done.

7.5.1   Histograms in the statistics education curriculum

Dutch teachers in Grades 4 and 8 are the most frequent users of textbooks according to a Trends in International Mathematics and Science Study (TIMSS) (Foxman, 1999). In the Netherlands, similarly to many other countries, "textbooks […] are the supporting backbone for most teachers (whether or not one believes this should be the case)" (Leinhardt et al., 1990, p. 47). Therefore, in the following sections, we provide implications and recommendations for teachers, textbook authors, and curriculum designers.

Histograms are needed for learning key concepts
We recommend a central role for histograms in a curriculum that also includes other graphs of univariate data (e.g., dot-, stem-and-leaf, and boxplots). Histograms may play a pivotal role in learning statistical key concepts such as

data, distribution, variability or variation, and central tendency (Garfield & Ben-Zvi, 2008a). Histograms prepare for other key concepts such as probability distribution and density in probability theory (Batanero et al., 2004). Each key concept relies on other concepts (e.g., distribution relies on center, density, skewness, relative frequency, Bakker & Gravemeijer, 2004). Therefore, a histogram can be regarded as a spider in a web of knowledge (Chapter 1), see the network of statistical concepts relevant to interpreting histograms (Chapter 2). Unfortunately, we cannot learn those key concepts without signs (e.g., histograms), as the representation of data as well as how their distribution manifests itself (through its shape) strongly depends on the specific type of graph (Chapters 1, 2). The underlying conceptual difficulties become manifest when students interpret histograms, making histograms a good diagnostic instrument for teachers and researchers.

Given students' persistent difficulties with histograms, one might wonder whether we can do without them in research and education. We think we cannot. First, histograms are suitable for large amounts of data because they aggregate them, and as such, they can reveal aspects of distributions that most other graphs often do not (e.g., Pastore et al., 2017). Second, histograms are omnipresent in research and society, and should, therefore, be learned. Third, students also exhibit comparable misinterpretations of alternatives such as boxplots and stacked dotplots (Bakker et al., 2004; Lem et al., 2013a, 2013b, 2014a; Lyford, 2017). Fourth, it is the key concepts underlying histograms that are hard to grasp.

Focus on key concepts and bring data more upfront
An implication of our research is that within the key concepts it is plausible that *data* need to be put more to the forefront in Dutch secondary statistics education, and most likely also in other countries (Chapter 2). This plea for more emphasis on the key concept of data is in line with developments in data science and statistics. These developments required "Re-thinking learners´ reasoning with non-traditional data," which was the theme of an international conference (SRTL-12) on statistics education that was co-organized by Ben-Zvi, Boels, Makar, and Van Dijke-Droogers. In addition, recent statistics education research argues for adding data literacy to statistical literacy. Gould considers data literacy to be statistical literacy with more focus on being "a critical consumer of data, controlling [ones…] personal data trial, finding meaning in data, and taking action based on data" (2017, p. 23). Furthermore, analysis of media items during the pandemic brought to light that our students need to understand that statistics and predictions are tentative and that data, analyses, and results can be debatable or may need revision (Gal & Geiger, 2022). This might challenge the view that scientific findings or statistics are the

truth. They suggest adding *evidence literacy* to the statistics curriculum as part of critical thinking (see section 7.21). Teaching examples for incorporating critical interpretation of data presented in statistical graphs can be found on the website[74] What's Going On in This Graph?

The review study made us change our focus from how to interpret the sign or artifact (histogram) to key concepts in statistics. However, several educational systems are still dealing with a previous change from teaching students how to construct histograms to how to interpret histograms: "didactical research suggests that the emphasis of teaching is often put on the construction of such graphs, with little attention to their interpretation" (González et al., 2011, p. 188). Since then, little has changed (e.g., Burrill, 2020). For example, in United Kingdom assessments "there is very little emphasis on statistical skills such as interpreting data and drawing conclusions, and a great deal of emphasis on technical skills" (Ridgway & Nicholson, 2019, p. 1). The choice of software for classroom use can either support or hinder refocusing on key concepts (see the methodological contributions section). The Netherlands unfortunately is no exception to this international tendency for teaching practice to focus primarily on how, for example, to draw a histogram from a frequency table or calculate measures of centers (Chapter 1) rather than interpreting graphs and developing an understanding of key concepts through histograms. This stresses the importance of supporting teachers and authors of textbooks in such a refocusing, as textbooks may have a cumulative impact on students' achievement (Van den Ham & Heinze, 2018).

An implication of our research is also that it confirmed that estimating means from graphs can demonstrate students' conceptual knowledge (e.g., Gal, 1995). In addition, our eye-tracking studies showed that students used correct strategies for estimating means from case-value plots (Chapter 3) and most non-stacked dotplots (Boels & Van Dooren, 2023). Therefore, it is most likely not the estimation of means from graphs itself that is causing students' difficulties with histograms. Instead, it is how the *data* are presented in histograms (Chapter 2). We consider estimating and comparing means of data in graphs as a first step toward assessing and comparing variability, as variability is assessed against some measure, often the mean. For example, the standard deviation is a (non-linear) deviation from the arithmetic mean. Variability is part of the key concept of distribution. A next step could, therefore, be to develop students' knowledge about distribution.

---

[74] https://www.nytimes.com/2020/06/10/learning/over-60-new-york-times-graphs-for-students-to-analyze.html

Histograms in the investigative cycle

Research recommends having graphs play a central role in innovative curricula (Garfield & Ben-Zvi, 2008a) and in initial data analyses. This is in line with the advice to always graph data first before applying any statistical test (e.g., Anscombe, 1973; Matejka & Fitzmaurice, 2017; Pastore et al., 2017). In innovative curricula, the investigative cycle (Figure 7.3) plays an important role[75]. In this cycle, histograms may assist during planning, data collection and cleaning (e.g., finding outliers), and analysis (e.g., exploration, hypothesis generation). So far, we have not found any dataset being used in Dutch Grades 7–12 textbooks that needs, for example, data cleaning. In addition, histograms could be used during inferential reasoning such as testing a hypothesis and drawing conclusions (e.g., interpretations, generating new ideas). Garfield and Ben-Zvi (2008a) state:

> Today's more innovative curriculum and courses have students constantly revisit and discuss *graphical* representations of data, before any data analysis [e.g., calculating means and standard deviations] or inferential procedure [occurs]. In a similar vein, the ideas of distributions having characteristics of [...] center, and spread can be revisited when students encounter theoretical distributions and sampling distributions later [on]. (p. 168, emphasis added)

Discussing these recommendations, we note that histograms are part of what is often called descriptive statistics. This classification belittles the role of histograms in inferential reasoning. For inferential reasoning, research suggests starting with qualitative inferences (Van Dijke-Droogers, 2021) instead of computations. An example is asking students to *estimate* center (e.g., arithmetic mean) and variation from a graph (Gal, 1995) instead of *calculating* mean and standard deviation from a graph or frequency table. It is plausible that the investigative cycle and statistical key concepts, up to now, received little attention in Dutch statistics education. For example, if teachers follow mathematics textbooks, students will rarely collect data themselves. Instead, they will work either with representations of these data (e.g., tables, graphs) or with given datasets. As the choice of graphs depends on what data are collected (number of variables, measurement level), students' experiences with it are important. However, as it can be time-consuming, *talking* students through the data creation process can be an alternative:

---

[75] Recently, attempts have been made to incorporate insights from data science (IDSSP, 2019; Fry & Makar, 2021).

> [Several researchers] stress the importance of talking through the process of data creation as necessary preparation to seeing data as numbers in context. [It can also] address the measurement and sampling issues: what variable exactly is measured and how? However, such guided discussions alone may not suffice; in our view, students should also experience a whole investigative cycle. (Bakker, 2004a, pp. 256–257)

Figure 7.3 Investigative cycle (redrawn from Wild & Pfannkuch, 1999)
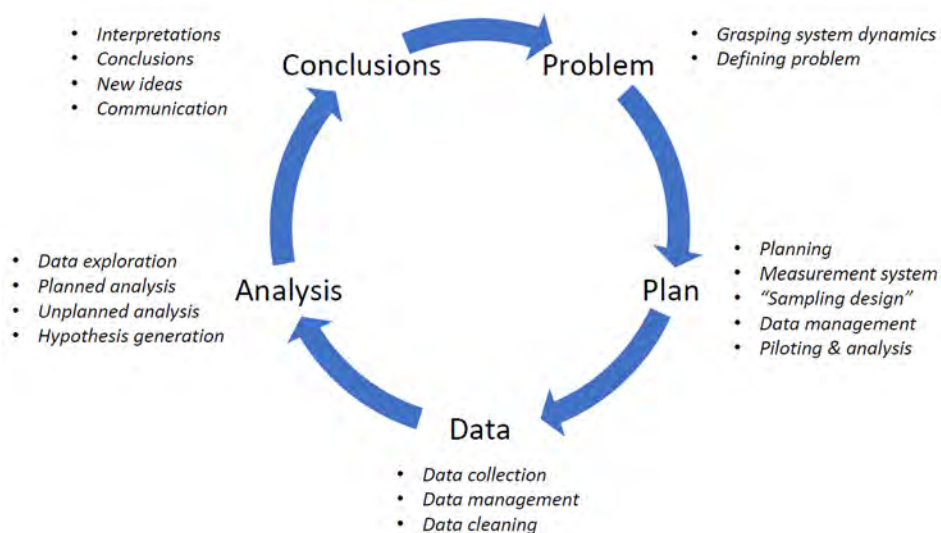
Note. Some research adapted this cycle to include data exploration with "data from disparate sources, some of which may not have been mindfully collected or may have been collected for a purpose different from the current application" (Gould, 2021, S21; see also Wise, 2020). Based on key practices of data scientists, Lee et al. (2022), use "consider and gather data" instead of data collection and separate this step from the processing of the data (e.g., data management and cleaning) (p. 11). For image-based data, Kazak et al. (2022) developed a version of the investigative cycle that starts with "data familiarization" and includes an "Identification/Generation of Variables" step (p. 5).

In our quick scan of textbooks (e.g., Moderne Wiskunde, 2015), we did not find examples of *talking* through this data collection process. Although, in recent textbooks, it seems that some improvements have been made, such as including a task in which students collect data themselves (e.g., Moderne Wiskunde, 2019), the full 'talking through the data creation process' for the many datasets given in the textbooks does not seem to occur. McClain and Cobb (2001) provide an example of how this could look for the braking distance of cars. This process then includes "what data would be needed to

make a good decision or how that data could be generated [...]. How one might test the effectiveness of a car's brakes or how that data could be gathered" (2001, p. 113). Given the vast amount of data that are nowadays collected by others, emphasis on how, why and by whom data are collected becomes more and more important for statistical literacy and critical citizenship (Fry, 2019; Gal & Geiger, 2022).

Postponing the introduction of histograms

Based on our research we suggest postponing the introduction of histograms to Grades 9 or 10 as we think that the introduction of histograms comes too early in curricula. Many students' difficulties with histograms continue to exist up to tertiary education and in fact, even some teachers have difficulties with histograms (e.g., Lovett & Lee, 2019). Bakker (2004a) already advised "against introducing histograms in early middle school grades" as students in Grades 7–8 needed a lot of time to understand center and spread in dotplots; spending time on histograms (or boxplots) would take time away from developing these notions (p. 261). Currently, the Common Core State Standards for Mathematics (CCSSM) introduce histograms (together with dotplots and boxplots) in Grade 6, and the Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II) gives examples of histogram test items for level A and B ("roughly equivalent to elementary, [and] middle [...] school "(Bargagliotti et al., 2020, p. 2; more assessment examples in Chance et al., 2018; Tintle & Vander Stoep, 2018). We think that non-stacked ('messy') dotplots (Chapter 6), and hatplots (e.g., Allmond & Makar, 2014; Konold, 2002) are preferred for Grades 6–9 (cf. Bakker, 2004a; Fielding-Wells & Hillman, 2018). In addition, we expect that histodots (Chapter 2) and stem-and-leaf plots can be introduced in Grades 8 or 9 as preparation for histograms. These univariate graphs have in common that they prepare for proportion-based reasoning (Frischemeier, 2019). Non-stacked dotplots and hatplots can also be used to build students' intuitions for boxplots (cf. Makar & Confrey, 2003) and prepare for quartile-based reasoning (Frischemeier, 2019). Boxplots share with histograms that they depict univariate data and that many misinterpretations occur (e.g., Bakker et al., 2004; Lem et al., 2013a, 2013b, 2014a). Therefore, boxplots need to be carefully introduced; we speculate in Grades 10 or 11, and *after* introducing histograms.

> There are hidden conventions and conceptual elements in histograms and boxplots: in histograms, the area of the bars is relative to the number of values it signifies, and in boxplots conceptual elements such as median and quartiles are depicted. (Bakker, 2004a, p. 13)

Another implication of our eye-tracking study (Chapter 3) is that just telling students to carefully read the axes and graph titles will probably not be enough (see also 7.4.4). An implication of testing our initial HLT is that pre-university track (vwo) students in Grades 10–12 seem to be capable of correctly estimating means from histograms when this is carefully prepared in targeted tasks. We expect students will also be able to compare means from histograms and dotplots when properly prepared. We consider such tasks as preparation for qualitative assessing and comparing variability from dotplots, histodots, and histograms.

Use different names for different types of graphs with bars
We recommend using different names for different types of graphs with bars. As pointed out in the introduction, in English, different names are used for different types of graphs (e.g., Cooper, 2018). When correctly applied, these different names can help students to distinguish different types of graphs with bars from each other. Unfortunately, our native language—Dutch—only distinguishes histograms and bar charts, and some Dutch textbooks do not use the word histogram at all (e.g., Moderne Wiskunde, 2015, 2019). Therefore, Dutch words were introduced (Boels, 2019) for case-value plots (casus-staafdiagrammen), distribution bar charts (verdelings-staafdiagrammen), and time-plots (tijd-diagrammen). We beg textbook authors worldwide, and in the Netherlands specifically, to start using them correctly and avoid ambiguous words such as bar graphs or bar charts (in Dutch: 'staafgrafieken' or 'staafdiagrammen')[76].

7.5.2 Professional development of mathematics teachers and textbook authors

Provide teachers with professional development opportunities
In discussing the implications of our work, we note that a change in the curriculum implemented in textbooks is not enough to incorporate insights from research. Making teachers aware of key concepts in statistics, students' misinterpretations related to these key concepts, their own misinterpretations, and effective teaching strategies to prevent such misinterpretations is necessary (cf. Pareja Roblin et al., 2018). For current teachers and future statistics education curriculum reform, we, therefore, recommend implementing it in teacher training.

First, teachers are the ones who deliver the curriculum. Therefore, it is important to not only look at how insights from research could be

---

[76] The distinction is important because they need to be analyzed and interpreted differently. Similar to a triangle and a square which are both polygons, but importantly different for reasoning.

implemented in the curriculum in, for example, mathematics textbooks as we did in the previous section, but also how teachers can be supported to enact the *intended* curriculum. Teaching statistics is "less popular among many mathematics teachers [… and several of those] in the early years of secondary education are inexperienced and not trained to teach inferential statistics" (Van Dijke-Droogers, 2021, p. 171). The same seems to apply to those teaching statistics in upper grades in the Netherlands. As we will argue below, we think this is partly due to a lack of training, not providing topic-specific support and overviews for teachers, and a lack of possibilities for many teachers to cross boundaries between research and teaching.

Second, in the Netherlands, as in many other countries, statistics in middle and high school is often taught by mathematics teachers. The deterministic approach in the mathematics curriculum and the inherent uncertainty that exists within statistics do not always align (e.g., Groth, 2015). Moreover, "many mathematics and science teachers in the USA have not benefitted from the sufficient opportunity to learn statistics in a sense-making manner" (Burrill & Ben-Zvi, 2019, p. xiv). The same holds true for several other countries, including the Netherlands (Van Dijke-Droogers, 2021). Although probability has been part of mathematics education for several decades now, statistics was first introduced as a topic in the elective part of the mathematics curriculum in 1985 (e.g., Wijers & De Haan, 2020). From 2007 on, a new curriculum reform resulted, on the one hand, in statistics playing a larger role for students who chose humanities and social studies as a continuation. On the other hand, for future science students, statistics became part of an elective subject (Mathematics D) that was chosen by only a limited proportion of them. Many future mathematics teachers are, therefore, still only introduced to statistics for the first time during their training as teachers. It is unclear to what extent knowledge about key concepts in statistics is covered, or whether this training focuses mainly on procedural or instrumental knowledge, including for example, how to find the median by hand or by using a calculator.

Third, despite in-service teacher training provided during the first years of the last reform in 2007, STEM teachers still hold several misinterpretations related to key concepts of statistics (e.g., for estimating and comparing means from histograms and case-value plots, Boels et al., 2019b). This situation is not unique to the Netherlands. High school mathematics teachers with at least five years of experience had similar misinterpretations to students about the key concept of variability (Vermette & Savard, 2019). Middle- and high-school teachers found it difficult to coordinate graphs with calculation procedures, such as absolute deviation for the mean (Peters & Stokes-Levine, 2019)—a measure that is not part of the Dutch statistics curriculum for secondary education.

Fourth, design research is needed on how best to organize such training in the Netherlands and internationally. That this is not an easy job is illustrated by research involving preservice teachers who compared data of groups (e.g., monthly income for males and females). After attending a course on "developing statistical thinking and reasoning with TinkerPlots" (Frischemeier, 2019, p. 292), teachers concentrated "on the production of displays and the calculation of summary statistics but [they] do not interpret their findings" (p. 301).

Fifth, given teachers' dependency on mathematics textbooks, we speculate that reforming textbooks might be another effective route to increasing the level of knowledge of both teachers and students.

Supporting teachers in implementing curriculum reform

Given that teaching statistics does not seem very popular for (Dutch) mathematics teachers, we recommend supporting teachers in implementing the statistics curriculum after each reform. Currently, we do not know how well the implemented statistics curriculum for Dutch secondary schools matches the intended curriculum (see also Verschut & Bakker, 2010), although there is some analysis of mathematics textbooks (e.g., Huang, 2022; Rodríguez-Muñiz et al., 2018). As many teachers rely on textbooks, research is needed on how the intended statistics education curriculum is implemented in textbooks and in teaching practice. Providing teachers with information about the matches and gaps between the intended and implemented statistics education curriculum in textbooks could help with closing knowledge gaps:

> Providing support for teachers as they form the intended curriculum and enact it could help ensure that the intended spirit of the curriculum materials [...] is not lost. Additionally, as curriculum writers interact with teachers, they may find that some adaptations teachers make to the written curriculum help to improve it. (Groth, 2015, p. 14)

Providing teachers with exemplary teaching materials (cf. cTWO, 2007) during curriculum reform most likely needs to be part of that but will probably not be enough. What teachers also need is to know "the intentions of the authors of exemplary teaching materials [...] Suggestions in the materials for classroom activities that stimulate coherent knowledge and make efficient use of time" (Verschut & Bakker, 2011, pp. 921–922), explication of the structure of exemplary materials, and their connection to the curriculum. We, therefore, recommend that the intentions of the materials (e.g., which key concepts and misinterpretations they aim to address) and suggestions for classroom activities be part of it. Left for future research is also how to best provide

teachers with supporting materials during and after reforming the statistics education curriculum.

Crossing boundaries between research and teaching and provide teachers with overviews of students' conceptual difficulties

We recommend developing supporting materials for teachers teaching statistics (cf. cTWO, 2007) as these materials could help cross boundaries. Crossing boundaries—here from teaching to research and back—can be difficult (Akkerman & Bakker, 2011). Before we started this study, many teachers and textbook authors were not aware of students' difficulties with interpreting histograms (Reinhart et al., 2021). In addition, research insights are not always used in teaching practice, not only in the Netherlands but in many countries (Bakker et al., 2021). We speculate this is partly due to inaccessible jargon, a fragmented landscape with *detailed* studies on specific topics, secondary school teachers who instead teach a *broad* curriculum, and the limited time available to teachers to explore topics in depth. Research is needed to find out what teachers' support should look like. One effective way to improve teaching practice could be to provide teachers with (replace science with statistics):

> …extensive lesson directions and […] activities [that] were designed to elicit students' ideas and many possible misconceptions […] questions were provided and teachers were given suggestions on how to help […]. The curriculum consists of instructional materials for both students and teachers. The suite of teacher support materials helps to deepen teachers' knowledge of science content and practices related to the unit. These materials include (among other supports) *Background Content Knowledge*, which provides teachers with more advanced information on the science content, important observations students should make, and any observations teachers might emphasize/deemphasize. (Pareja Roblin et al., 2018, p. 276, emphasis in original)

7.5.3   Future applications of eye-tracking in statistics education

Before continuing to possible future applications of eye-tracking research in statistics education, in the next section, we first would like to highlight an ethical point of attention regarding the usage of students' gaze data.

Ethical considerations

Although data collection can have advantages, such as music websites offering music that you might like based on your previous choices, it has a downside, too. Fry (2019) provides several examples of improper use of data to train

machine learning algorithms that decide who is to be invited for a job interview or who is turned down for a loan. Therefore, an ethical discussion needs to be started about gaze data.

Today, gaze data are already being used in the gaming community (e.g. EyewareBeam[77]) and its introduction into education will probably only be a matter of time. Currently, it seems impossible to retrace gaze data to a specific person, as for most software, only the gaze position on the screen is registered (e.g., Gorilla.sc, n.d.), although some *webcam* eye trackers do store videos with faces on servers as well. In addition, no general relationship seems to exist between eye movements and thought processes; this relation needs to be established for each situation. However, it is conceivable that in the future, faster methods will become available for analyzing data and that the idiosyncrasy of gazes could make them retraceable to specific persons.

It is, therefore, important that an ethical discussion be held *now* about who may collect and use students' eye movement data. Are we going to hand this over to large tech companies—just as we did with earlier data—or will this remain reserved for non-commercial parties only? Can students—and teachers—refuse to make their data available, something that currently seems impossible when using, for example, Google Classroom? We strongly recommend thinking about such questions now.

Gaze data in a feedback or information tool

We discuss several implications for how gaze data can be used as part of a feedback or information tool in education. A first possibility is to provide students with their own gaze data after solving one or more tasks and ask them to describe the strategy they used. Our research and analysis (Chapters 3 and 5) suggest this may also help students reflect on their chosen strategy. Our students seemed to have no difficulties interpreting their own gaze data during cued recall. Students' gazes were shown by illuminating the location where students looked—through a kind of spotlight—while making the rest of the graph darker. However, this way of having students individually look back at their eye movements is time-consuming and not (yet) feasible for regular use in classrooms.

A second possibility is to provide students with other students' gaze data. We have done this with two students in our intervention study, but do not report on that in Chapter 6. We have not yet analyzed these data. Both possibilities required a time delay between data collection and replay.

A third possibility is to provide students with immediate, personalized feedback based on their gaze data (e.g., Król & Król, 2019). Such automatic

---

[77] https://beam.eyeware.tech/games/

feedback could become possible if a number of conditions are fulfilled. First, there need to be distinctive eye movement patterns that can be linked to specific strategies (Chapter 3). Second, after developing an IMM or training an MLA, such tools need to be able to extract these patterns from the gaze data of new students. This condition was met for single histogram tasks (Chapter 4). Third, inexpensive equipment is needed to measure eye movements. Further research is needed to investigate if webcams could be used (e.g., Knoop-Van Campen et al., 2021) as these have less accuracy. Only if sufficient accurate eye-tracking becomes inexpensive, can large-scale applications for gaze-based personalized feedback become feasible in classrooms, during homework or distance learning, and in MOOCs. For all three possibilities mentioned above, the question remains in what way feedback should be given: should students see their own or others' eye movements (e.g., Król & Król, 2019), or is another form of feedback needed (cf. Tacoma et al., 2019)?

A fourth possibility is to provide teachers with information based on students' gaze data. Several questions still remain open in that case. Is it better if such a system reports back which students are using a correct strategy, which students are not, and for which students the strategy is unclear so that the teacher can intervene in a targeted way? Is it useful or necessary to then provide the teacher with a record of students' eye movements, and if so, in what form? This also raises the question of whether teachers can identify students' strategies—from students' gaze data—when students are interpreting a statistical graph. Do they need an instruction for that and if so, what should such an instruction look like? Teachers could not only be asked whether they think a student had performed a correct strategy—or what strategy they think the student used—but also, if the strategy was inappropriate, what kind of intervention they would do. This is what bachelor students did—(re)using students' gaze data of the study described in Chapter 3 for single histograms and case-value plots (Benson et al., 2020). To the best of our knowledge, this is one of the first studies that provided secondary school teachers with the opportunity to interpret and thus reason with this 'non-traditional' data. As this data collection was relatively small and hindered by the COVID-19 pandemic, further investigation is needed.

Gaze data in tertiary education courses
We suggest developing tertiary education courses that focus on task-specific strategies in mathematics and statistics education inferred from gaze data. In these courses, students (and teachers) will then collect and analyze gaze data themselves. There are already courses on eye-tracking, often taught in neurosciences or psychology departments. Students collecting data in these courses are, for example, interested in memory and cognitive load theory, self-

regulated learning, anomalous viewing patterns related to certain health problems, or metacognitive skills. In marketing, gaze data are used for inferring decision behavior. Currently, courses that focus on the kind of task-specific strategies found in our research seem to be rare.

Gaze data to revise instructional design

Future research can also use gaze data to revise the instructional design. Examples of using multimodal data for revising the design can be found in several studies (Alberto et al., 2022), such as on proportions (Shayan et al., 2017) and trigonometry (Shvarts et al., 2021).

8
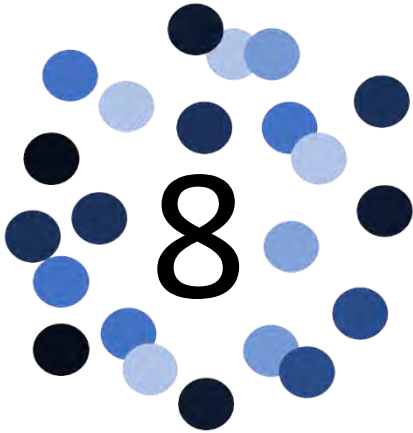
Teacher-researcher's reflections on conducting research

"Insanity is doing the same thing over and over again and expecting different results." [78]
Rita Mae Brown

---

[78] This quote is often misattributed to Einstein. Instead, this version was from Rita Mae Brown in *Sudden Death*, 1983, in which she rephrased a quote from Narcotics Anonymous, 1981. In Van Wayenburg, B. (2015). Vijf quotes die Einstein nooit heeft uitgesproken [Five quotes Einstein never spoke]. *Kijk*. https://www.kijkmagazine.nl/artikel/vijf-quotes-die-einstein-nooit-heeft-uitgesproken/

8.1 Reflections on conducting design research

As described in the introduction, when I started this research, I intended to conduct a design study with input from literature and from an eye-tracking study. However, from the literature review it became clear that, in 2016, there was not yet an effective intervention for secondary education waiting to be tested and tailored to Dutch education. In addition, although eye-tracking had been around for several decades, its application in the field of statistics education was and is still in its infancy (e.g., Strohmaier et al., 2020). At the start of our pilot eye-tracking study, we, therefore, had no idea what to expect. I remember very well that I was concerned that we would not find anything at all as we constructed graphs (histograms and case-value plots) that contained a clear context (weights of packages and garbage) excluding any of the known confusing contexts; we used 'easy' numbers (below 20), we used axes scales and titles, and we explicitly indicated graph titles. It came as a small surprise that even master students teaching statistics were sometimes misinterpreting histograms, or were overgeneralizing histograms to case-value plots, and that clear solution patterns seemed to emerge for single graphs (Boels et al., 2018).

Conducting a proper eye-tracking study is a lot of work and required me to also dive into the literature of that discipline. We benefitted most from the data collection by adding extra tasks to the first twelve tasks intended for uncovering students' solutions processes—strategies—for histograms and their look-alikes (case-value plots). These rich data formed the basis of three studies included in this dissertation and opened up opportunities for taking a step toward automatic feedback based on students' gazes (Chapter 4). Although Enrique Moreno-Esteva and Alex Lyford—whom I met at international conferences—executed the machine learning analysis, writing up the work we did required that I acquired at least some basic knowledge on yet another discipline, namely how machine learning algorithms (MLAs) function. Getting that work published was not an easy task either, as it is—again—on an intersection of disciplines: eye-tracking research related to cognitive sciences, artificial intelligence, and statistics education. Since educational research journals usually lack specific knowledge about MLAs, a specialist reviewer is frequently called in for this purpose. These reviewers often focus on tailoring an MLA to the situation, which originates from the discipline of artificial intelligence. However, our approach described in Chapter 4 used black-box software with an MLA as a tool, as we focused on the educational application. It is important to keep boundaries between disciplines permeable, as the dialogue between different disciplines can be fruitful for all:

> Sustained boundary interactions [here: between statistics and mathematics education] are vital to preventing insularity from contributing to the stagnation of interrelated communities of practice (Wenger, 2000). When boundary interactions occur, borders between disciplines can become exciting sites for learning rather than prohibitive barriers. (Groth, 2015, p. 5)

By crossing boundaries, practices from different disciplines can be combined and lead to new tools, concepts, models and new practices (Akkerman & Bakker, 2011). The disadvantage of working at boundaries is that results are reported in so many different journals and platforms that this can hinder further development (Groth, 2015).

Although the main contribution of this dissertation lies in the important "front-end work" (e.g., McKenney in Bakker, 2018, p. 142), as a teacher I could not live with this front-end work only. Therefore, I ended this dissertation with a first cycle of a design study, using an approach that was rather new in the field of statistics education—embodied cognition and instrumentation—requiring again getting acquainted with a new body of literature. I am very proud that we were able to introduce so many new tools and approaches (e.g., eye-tracking, task-specific gaze patterns, machine learning analysis of gaze data using vectors, embodied instrumentation design) into the field of statistics education.

8.2 Personal reflections as a teacher-researcher

I conclude with some personal reflections on the combination of being a teacher and a researcher. When I started this research, I dropped several tasks at school, including being a 'technator' (coordinator) at the Technasium[79]. Although this allowed me to focus on my teaching, it also meant that I had fewer connections with colleagues, a process that intensified during the COVID-19 pandemic. At the same time, compared to full-time PhD-students, it is a luxury to have your own classes for piloting. For example, from this experimenting, I learned that even the best-performing pre-university track students ('vwo', Mathematics D, Grade 11) found it difficult to indicate which and how many variables were along the axes for graphs with bars, as well as the measurement level of the variables' attributes. I kept it as a suitable task to utilize in professional development courses with high school teachers.

---

[79] In a Technasium, students in Grades 6–12 undertake projects in which they conduct STEM research or design STEM products. External companies act as principals and bring in existing scientific or technological problems from their own practice for students to solve.

I have always experimented in my classes, but now I was increasingly inspired by specific research. Based on embodied designs for trigonometry (Alberto et al., 2019)[80], I created a paper task—due to the lack of computers with touchscreens in this class—to review the subject in a Grade 12 (6 'vwo') Mathematics A class (Boels, 2022a, 2022b). The same research helped me choose simulations of the relationship between unit circle and sine more carefully (Boels, 2022a). In Grade 10 (4 'havo'[81]), I had students discover the relationship between the unit circle and sine graph through embodied tasks (cf. Alberto et al., 2019). In the absence of touchscreens, I placed one student at a time behind my own touchscreen laptop connected via an online meeting environment via the desktop computer to the digital board—digital skills I learned during the COVID-19 pandemic. I asked the class to discover the hidden rule, and if they knew it, not to tell but to demonstrate it on the computer. My impression after the test was that these students understood this topic better than other students.

As an early researcher, I wondered about a number of things as well. First, why are mathematics teachers almost absent at conferences on research in mathematics and statistics education? Such conferences (e.g., ICME-13) opened up a new world to me, as so many problems were discussed that I regularly encountered in my classroom. One answer is that the amount of travel, lodging, and participation fees for such conferences largely exceeds a teacher's annual training budget, a budget I tended to almost fully spend on visiting the national conference for mathematics teachers (NWD[82]). In addition, most schools do not allow teachers to be away for five days or more for a conference. I am very grateful that my school did, several times. I wish that every mathematics teacher could go—at least once every two years—to an international conference on mathematics or statistics education. Here are a few that seem appropriate, although I am aware that the scientific English jargon can be difficult: ICME[83], ICOTS[84], PME[85], CERME[86].

Second, conducting didactic research requires skills from research in education and psychology. These are usually not included in science education

---

[80] See also: https://embodieddesign.sites.uu.nl/activity/

[81] Havo is a pre-college track for Grades 10 and 11.

[82] https://www.uu.nl/en/research/freudenthal-institute/impact/conferences

[83] International Congress on Mathematics Education. Content: all mathematical domains. Website: https://www.mathunion.org/icmi

[84] International Conference on Teaching Statistics. Content: statistics, data science and probability education. website: https://iase-web.org/Conferences.php

[85] Psychology of Mathematics Education. Content: all mathematical domains. Website: https://www.igpme.org/

[86] Conference of the European Society for Research in Mathematics Education

(e.g., Bakx et al., 2016). I learned these skills on my own time parallel to writing the research proposal. I had the luxury of taking off at least one day a week for this, in addition to my job, running my own company, and taking care of three teenagers. I would wish for future teachers to have the opportunity to gain these skills with a teacher grant in a kind of pre-promotion master's program (1–2 years, maximum 30–60 credits), in which they also write their research proposal.

Third, I was surprised that so much research at universities is done by relatively inexperienced people: bachelor, master, and PhD students, and postdocs. By the time they are sufficiently experienced, there is often little to no external funding available anymore through, for instance, NWO. As a result, experienced researchers mainly teach and supervise, and very rarely conduct their own research. Moreover, for the majority of researchers, there is simply no place at the university after their PhD or postdoc position. Their research experience then gets lost to the scientific world. This is a waste of money, talent, and human resources.

Fourth, I greatly appreciate that a follow-up postdoc grant exists for didactical research by secondary and vocational education science and mathematics teachers. My wonderment concerns the possible gap of sometimes more than a year after completing a PhD trajectory. Although certainly not the most important reason, this possible gap played a role in my decision not to continue teaching at a secondary school. I was eager to continue in didactical research right away, and wanted to write my research proposal for continuation as part of my job, instead of on my own time. My advice to NRO is to install a post-promotion proposal writing grant that can be applied for on an ongoing basis once a dissertation is submitted and a promotion date is scheduled.

I sometimes joke that teachers must be incredibly smart people as they conduct a full PhD research in about two full-time years. The grant my school got to replace me was for 0.4 FTE[87], initially for four years, then for five, and thanks to COVID-19, three more months were added. That worked out at 2.1 FTE. Internal PhD candidates at the university get almost double—four years full time, hence, my joke. Of course, there is some room for disagreement with this reasoning. To give just one example, the latter have (few) teaching duties, although these could include teaching others to conduct qualitative research— lessons from which young teachers might learn a lot themselves. In addition, switching between short-term requirements from school and long-term planning of research is challenging (Bakx et al., 2016), hard work and requires a lot of energy. I am extremely grateful that I was able to do this research. Still, I

---

[87] Full Time Equivalent. 1 FTE is equivalent to a full time job.

hope that my successors will have an easier job, after finishing the proposed pre-promotion master's program.

I end this personal reflection with three more wishes for the future. First, as a teacher, I would like to see an analysis of the implemented statistics education curriculum in textbooks—in a format accessible for teachers—compared to the intended curriculum. Textbook analyses have been made previously (e.g., for primary education, Van Zanten, 2020; for secondary education, Huang, 2022) but this information is scattered, mostly in English, difficult for teachers to find and access, contains scientific jargon, and is often outdated by new books due to the long duration of research. Second, I am curious about the association between Dutch textbooks and final exam results for specific domains, as well as students' results in subsequent education. My suspicion is that there is a difference between them depending on the textbooks, but I know of no recent research on this for Dutch education. Third, I advocate for research-informed advice based on didactical research in mathematics and statistics education in a form that has been found effective for *secondary* STEM education, such as an A4-sheet per topic listing the most important misinterpretations, best practices, specific points of interest for teaching, and so on (e.g., Pareja Roblin et al., 2018), as so-called evidence-based advice is regularly based on elementary and to a lesser extent, special secondary education (e.g., Mason & Otero, 2021) and, due to the necessity of meta-analysis and review studies, also sometimes based on past practices that might not always fit future needs and innovations. Fourth, I think it would be helpful if there were research schools for secondary and vocational education, affiliated with an (applied) university, where research is conducted on an ongoing basis. This could create a community of mathematics teachers who jointly address didactical problems in mathematics and statistics education through lesson study and action or design research. In addition, such a community could prepare for research-informed teaching and could introduce teachers to opportunities to conduct their own doctoral research. Similar to what has been done for health research[88], I advocate for funding for research addressing the didactical—not pedagogical—needs of mathematics teachers in that community and beyond.

---

[88] https://www.zonmw.nl/nl/onderzoek-resultaten/preventie/programmas/programma-detail/alledaagse-ziekten/

References

Abbasnasab Sardareh, S., Brown, G. T., & Denny, P. (2021). Comparing four contemporary statistical software tools for introductory data science and statistics in the social sciences. *Teaching Statistics*, *43*, S157–S172. https://doi.org/10.1111/test.12274

Abrahamson, D. (2006). The shape of things to come: The computational pictograph as a bridge from combinatorial space to outcome distribution. *International Journal of Computers for Mathematical Learning*, *11*(1), 137–146. https://doi.org/10.1007/s10758-006-9102-y

Abrahamson, D. (2008). Bridging theory: Activities designed to support the grounding of outcome-based combinatorial analysis in event-based intuitive judgment–A case study. In *Proceedings of the International Congress on Mathematical Education (ICME 11).* ICME. https://iase-web.org/documents/papers/icme11/ICME11_TSG13_01P_abrahamson.pdf

Abrahamson, D. (2009). Embodied design: Constructing means for constructing meaning. *Educational Studies in Mathematics*, *70*(1), 27–47. https://doi.org/10.1007/s10649-008-9137-1

Abrahamson, D., & Bakker, A. (2016). Making sense of movement in embodied design for mathematics learning. *Cognitive Research: Principles and Implications*, *1*, Article 33. https://doi.org/10.1186/s41235-016-0034-3

Abrahamson, D., & Cendak, R. M. (2006). The odds of understanding the law of large numbers: A design for grounding intuitive probability in combinatorial analysis. In *Proceedings of the Thirtieth Conference of the International Group for the Psychology of Mathematics Education*, *2* (pp. 1–8).

Abrahamson, D., & Sánchez-García, R. (2016). Learning is moving in new ways: The ecological dynamics of mathematics education. *Journal of the Learning Sciences*, *25*(2), 203–239. https://doi.org/10.1080/10508406.2016.1143370

Abrahamson, D., & Wilensky, U. (2007). Learning axes and bridging tools in a technology-based design for statistics. *International Journal of Computers for Mathematical Learning*, *12*(1), 23–55. https://doi.org/10.1007/s10758-007-9110-6

Abrahamson, D., Dutton, E., & Bakker, A. (2021). Towards an enactivist mathematics pedagogy. In S. A. Stolz (Ed.), *The body*, *embodiment*, *and education: An interdisciplinary approach.* Routledge. https://doi.org/10.4324/9781003142010

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The future of embodied design for mathematics teaching and learning. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.00147

Abrahamson, D., Shayan, S., Bakker, A., & Van Der Schaaf, M. (2015). Eye-tracking Piaget: Capturing the emergence of attentional anchors in the coordination of proportional motor action. *Human Development*, *58*(4–5), 218–244.

Afonja, T. (2017, December 8). Accuracy paradox. *TDS*. https://towardsdata science.com/accuracy-paradox-897a69e2dd9b

k

o *Graphing. USMES intermediate "how to" set.* -k@# #
o U - -
-)

o 7 " " *Review of Educational Research* 81

k " ‡ \ " h " K) h = U
V
@ y u K U = h U †
- *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* 7 8 7
@ y y -kU -

k o) h "
International Journal of Child-Computer Interaction

- # M
Computers & Education 125

o U M 7 u h o
@ M'U "
o k 8 - *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* @@@ o-
@\ uo - OOU \ V)

K \ o - o
@ 8 ") " – - *Topics and trends in current statistics education research. International perspectives* o

K k ")) o -
O *Psychological Science* 15

V # 7 M " ‡ 7
Behavior research methods 47

O h 7 = M k \ o
k *International Journal of Science and Mathematics Education* 13
7 K 8 o *The American Statistician* 27
#
8@ - #\ " K

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics*, *52*(3), 215–241. https://doi.org/10.1023/A:1024312321077

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher*, *40*(1), 62–69, https://doi.org/10.1080/0142159X.2017.1391373

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2001). Toward a model of learning data representations. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 45–50). Psychology Press. https://www.researchgate.net/publication/2367832_Toward_a_Model_of_Learning_Data_ Representations

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2002). The resilience of overgeneralization of knowledge about data representations. In *Annual Meeting of the American Educational Research Association* (pp. 1–10). https://files.eric.ed.gov/fulltext/ED465530.pdf

Bakker, A. (2002). Route-type and landscape-type software for learning statistical data analysis. In B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the Sixth International Conference on Teaching Statistics* (pp. 1–6). ISI/IASE. https://iase-web.org/documents/papers/icots6/7f1_bakk.pdf?1402524963

Bakker, A. (2003). The early history of average values and implications for education, *Journal of Statistics Education*, *11*(1). https://doi.org/10.1080/10691898.2003.11910694

Bakker, A. (2004a). *Design research in statistics education: On symbolizing and computer tools* [Doctoral dissertation, Utrecht University]. https://dspace.library.uu.nl/handle/1874/893

Bakker, A. (2004b). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*(2), 64–83. http://iase-web.org/documents/ SERJ/SERJ3(2)_Bakker.pdf?1402525004

Bakker, A. (2018). Design research in education. A practical guide for early career researchers. Routledge. https://doi.org/10.4324/9780203701010

Bakker, A., Biehler, R., & Konold, C. (2004). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education: International Association for Statistical Education 2004 Roundtable,* (pp. 163–173). International Statistical Institute. https://www.stat.auckland.ac.nz/~iase/publications/rt04/4.2_Bakker_etal.pdf

Bakker, A., Cai, J., & Zenger, L. (2021). Future themes of mathematics education research: An international survey before and during the pandemic. *Educational Studies in Mathematics*, *107*, 1–24. https://doi.org/10.1007/s10649-021-10049-w

Bakker, A., Cai, J., English, L. Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, *102*, 1–8. https://doi.org/10.1007/s10649-019-09908-4

k

Mathematical Thinking and Learning 13

The challenge of developing statistical literacy reasoning and thinking

Educational Studies in Mathematics 62,

Educational Studies in Mathematics 60

K *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education*

Approaches to qualitative research in mathematics education. Advances in mathematics education

Teaching and Teacher Education 60

The International Journal for Geographic Information and Geovisualization 3

The Journal of the Learning Sciences 10

Pre-K-12 guidelines for assessment and instruction in statistics education (GAISE) report II

The challenge of developing statistical

literacy, *reasoning and thinking* (pp. 257–276). Springer. https://doi.org/10.1007/1-4020-2278-6_11

Battaglia, O. R., Di Paola, B., & Fazio, C. (2017). A quantitative analysis of educational data through the comparison between hierarchical and not-hierarchical clustering. *EURASIA Journal of Mathematics*, *Science and Technology Education*, *13*(8), 4491–4512. https://doi.org/10.12973/eurasia.2017.00943a

Beach, P., & McConnel, J. (2019). Eye tracking methodology for studying teacher learning: A review of the research, *International Journal of Research & Method in Education*, *42*(5), 485–501, https://doi.org/10.1080/1743727X.2018.1496415

Behrens, J. T. (1997). Toward a theory and practice of using interactive graphics in statistics education. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics. Proceedings of the 1996 IASE Round Table Conference* (pp. 111–121). ISI. http://iase-web.org/documents/papers/rt1996/10.Behrens.pdf?1402524984

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles*, *Policy & Practice, 18*(1), 5–25, https://doi.org/10.1080/0969594X.2010.513678

Benson, T., Van Ogtrop, N., Hodzelmans, S. (2020). *Can teachers recognize students' strategies when students interpret histograms*? [Unpublished bachelor thesis] Utrecht University.

Ben-Zvi, D., & Garfield, J. (Eds.), (2004a). *The challenge of developing statistical literacy*, *reasoning*, *and thinking* (1st ed.). Springer. https://doi.org/10.1007/1-4020-2278-6

Ben-Zvi, D., & Garfield, J. (2004b). Research on reasoning about variability: A forward. *Statistics Education Research Journal*, *3*(2), 4–6. http://iase-web.org/documents/SERJ/SERJ3(2)_forward.pdf?1402525004

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473–502). Springer. https://doi.org/10.1007/978-3-319-66195-7_16

Ben-Zvi, D., Makar, K., & Garfield, J. (Eds.), (2017). *International handbook of research in statistics education* (1st ed.). Springer. https://doi.org/10.1007/978-3-319-66195-7

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. Pergamon Press. (Original work published 1940)

Bernstein, N. A. (1996). *Dexterity and its development*. (M. L. Latash, & M. T. Turvey, Transl./Eds.). Psychology Press. https://doi.org/10.4324/9781410603357

Berrar, D. (2019). Cross-validation. *Encyclopedia of bioinformatics and computational biology*, *1* (pp. 542–545). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Biehler. R. (1997). Students' difficulties in practicing computer-supported data analysis: some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and*

learning statistics: Proceedings of the 1996 Iase Round Table Conference (pp. 169–190). International Statistical Institute.

Biehler, R. (2007). Students' strategies of comparing distributions in an exploratory data analysis context. *56th session of the International Statistical Institute*, *Lisbon*, *Portugal*.
https://www.stat.auckland.ac.nz/~iase/publications/isi56/IPM37_Biehler.pdf

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third International Handbook of Mathematics Education. Springer International Handbooks of Education*, *27*. Springer. https://doi.org/10.1007/978-1-4614-4684-2_21

Boels, L. (2019). FIzier. Wat elke docent zou moeten weten over histogrammen [What every teacher needs to know about histograms, translation author]. *Euclides, 94*(4), 10–13. https://archief.vakbladeuclides.nl/bestanden/094_2018-19_04.pdf

Boels, L. (2022a). Kleintje didactiek. De samenhang tussen sinus en eenheidscirkel. *Euclides*, *97*(6), 24.

Boels, L. (2022b). Kleintje didactiek. De samenhang tussen sinus en eenheidscirkel II. *Euclides*, *97*(7), 30–31.

Boels, L., Alberto, R. & Shvarts, A. (2023). Actions behind mathematical concepts: A logical-historical analysis. In *Proceedings of the 13th Congress of the European Society for Research in Mathematics Education.*

Boels, L., Bakker, A., & Drijvers, P. (2019a). Eye tracking secondary school students' strategies when interpreting statistical graphs. In M. Graven, H. Venkat, A. A. Essien, & P. Vale (Eds.), *Proceedings 43rd Annual Meeting of the International Group for the Psychology of Mathematics Education,* 2 (pp. 113–120). PME. http://www.igpme.org/publications/

Boels, L., Bakker, A., & Drijvers, P. (2019b). Unravelling teachers' strategies when interpreting histograms: an eye-tracking study. In M. Graven, H. Venkat, A. A. Essien, & P. Vale (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* (pp. 888–895). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02411575/document

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2023). *Gaze, interview and other data of secondary school students when interpreting statistical graphs* [Dataset will be published]. Freudenthal Institute, Utrecht University.

Boels, L., Ebbes, R. Bakker, A., Van Dooren, W., & Drijvers, P. (2018). Revealing conceptual difficulties when interpreting histograms: An eye-tracking study. Invited paper, refereed. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back*, *looking forward*. *Proceedings of the Tenth International Conference on Teaching Statistics* (pp. 1-4). ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8E2.pdf

Boels, L., & Shvarts, A. (2023). Introducing density histograms to Grades 10 and 12 students: Design and tryout of an intervention inspired by embodied

instrumentation. In G. Burrill, L. de Oliveria Souza & E. Reston (Eds.), *Research on reasoning with data and statistical thinking: International perspectives. Advances in Mathematics Education* (pp. 143–167)*.* Springer. https://doi.org/10.1007/978-3-031-29459-4_14

Boels, L., & Van Dooren, W. (2023). *Secondary school students interpreting and comparing dotplots: An eye-tracking study*. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education, 2,* (pp.123–130). PME.

Bolch, C. A., & Jacobbe, T. (2019). Investigating levels of graphical comprehension using the locus assessments. *Numeracy: Advancing Education in Quantitative Literacy*, *12*(1).

Bor-de Vries, M., & Hoogland, K. (2020). *De HU in gesprek over statistiek en instroomniveaus van studenten* [Utrecht University of Applied Sciences in dialogue about statistics and student entry levels]. Utrecht University of Applied Sciences. https://www.hu.nl/onderzoek/publicaties/de-hu-in-gesprek-over-statistiek-en-instroomniveaus-van-studenten

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bos, R., Doorman, M., Drijvers, P., & Shvarts, A. (2021). Embodied design using augmented reality: The case of the gradient. *Teaching Mathematics and its Applications: An International Journal of the IMA*. https://doi.org/10.1093/teamat/hrab011

Bosnić, Z., & Kononenko, I. (2009). An overview of advances in reliability estimation of individual predictions in machine learning. *Intelligent Data Analysis*, *13*(2), 385–401. https://doi.org/10.3233/IDA-2009-0371

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage. https://us.sagepub.com/en-us/nam/transforming-qualitative-information/book7714

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Wadsworth and Brooks/Cole Advanced Books and Software.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Bruno, A., & Espinel, M. C. (2009). Construction and evaluation of histograms in teacher training. *International Journal of Mathematical Education in Science and Technology*, *40*(4), 473–493. https://doi.org/10.1080/00207390902759584

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, *4*(1), Article 7. https://doi.org/10.1186/s41235-019-0159-2

Burrill, G. (2019). Building concept images of fundamental ideas in statistics: The role of technology. In G. Burrill, & D. Ben-Zvi (Eds.), *Topics and trends in current statistics*

education research. International perspectives (pp. 123–152). Springer. https://link.springer.com/book/10.1007/978-3-030-03472-6

Burrill, G. (2020). Statistical literacy and quantitative reasoning: Rethinking the curriculum. In P. Arnold (Ed.), New Skills in the Changing World of Statistics Education Proceedings of the Roundtable conference of the International Association for Statistical Education. ISI/IASE.

Burrill, G., & Ben-Zvi (2019). Introduction. In G. Burrill & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research. International perspectives* (pp. v–vi). Springer Nature. https://doi.org/10.1007/978-3-030-03472-6

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics: Challenges for teaching and teacher education* (pp. 57–69). Springer. https://doi.org/10.1007/978-94-007-1131-0_10

Cai, J., Moyer, J. C., & Grochowski, N. J. (1999). Making the mean meaningful: An instructional study, *Research in Middle Level Education Quarterly*, *22*(4), 1–24. https://doi.org/10.1080/10848959.1999.11670153

Capraro, M. M., Kulm, G., & Capraro, R. M. (2005). Middle grades: Misconceptions in statistical thinking. *School Science and Mathematics*, *105*(4), 165–174. https://doi.org/10.1111/j.1949-8594.2005.tb18156.x

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75–100. https://doi.org/10.1037/1076-898x.4.2.75

Carrión, J., & Espinel, M. (2006). An investigation about translation and interpretation of statistical graphs and tables by students of primary education. In A. Rossman & B. Chance (Eds.), *The Seventh International Conference on Teaching Statistics* (pp. 1–4). ISI/IASE. https://iase-web.org/documents/papers/icots7/C332.pdf

Catron, D. W. (1978). Immediate test-retest changes in WAIS scores among college males. *Psychological Reports*, *43*(1), 279–290. https://doi.org/10.2466/pr0.1978.43.1.279

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

CBS (2018). *Arbeidsdeelname en werkloosheid per maand* [Employment and unemployment rates by month]. https://www.cbs.nl/nl-nl/cijfers/detail/83933 NED?q=arbeidsdeelname

CBS (2021). Verdeling van gestandaardiseerd [jaar]inkomen [Distribution of standardized [annual] income]. https://www.cbs.nl/nl-nl/visualisaties/ inkomensverdeling

CBS (2022). Levend geboren kinderen; leeftijd moeder, volgorde geboorte uit de moeder [Live-born children; age of mother, order of birth from mother]. https://www.cbs.nl/nl-nl/cijfers/detail/37744ned?q=leeftijd%20moeder%20 geboorte%20eerste%20kind

Chan, S. W., & Ismail, Z. (2013). Assessing misconceptions in reasoning about variability among high school students. *Procedia-Social and Behavioral Sciences*, *93*, 1478–1483. https://doi.org/10.1016/j.sbspro.2013.10.067

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education*, *1*(1), Art 2. http://escholarship.org/uc/item/8sd2t4rr

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy*, *reasoning and thinking* (pp. 295–323). Springer. https://doi.org/10.1007/1-4020-2278-6_13.

Chance, B., Mendoza, S., & Tintle, N. L. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back*, *looking forward*, *Proceedings of the Tenth International Conference on Teaching Statistics* (pp. 1–6). International Statistical Institute.

Chase, K., & Abrahamson, D. (2015). Reverse-scaffolding algebra: Empirical evaluation of design architecture. *ZDM*, *47*(7), 1195–1209. https://doi.org/10.1007/s11858-015-0710-7

Chevallard, Y., & Bosch, M. (2014). Didactic transposition in mathematics education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education.* Springer. https://doi.org/10.1007/978-94-007-4978-8_48

Chisari, L. B., Mockevičiūtė, A., Ruitenburg, S. K., van Vemde, L., Kok, E. M., & van Gog, T. (2020). Effects of prior knowledge and joint attention on learning from eye movement modelling examples. *Journal of Computer Assisted Learning*, *36*(4), 569-579. https://doi.org/10.1111/jcal.12428

Chumachemko, D., Shvarts, A., & Budanov, A. (2014). The development of the visual perception of the Cartesian coordinate system: An eye tracking study. In C. Nicol, P. Liljedahl, S. Oesterle, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36*, *2* (pp. 313–320). PME. https://files.eric.ed.gov/fulltext/ED599779.pdf

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, *23*(1), 43–71. https://doi.org/10.1016/0010-0277(86)90053-3

Cito (2022). TiaEngine®, Toets en item analyse havo wiskunde A 2022 tijdvak 1, vraag 13 [Test and item analysis pre-college track mathematics A 2022 period 1, question 13]. Cito, Psychometrisch Onderzoek- en Kenniscentrum.

Clayden, A., & Croft, M. (1990). Statistical consultation—Who's the expert? *Annals of Mathematics and Artificial Intelligence*, *2*, 65–75. https://doi.org/10.1007/BF01530997

Cobb, P., Jackson, K., & Dunlap, C. (2016). Design research: An analysis and critique. In L. D. English & D. Kirshner (Eds.) *Handbook of international research in mathematics education*, *3* (pp. 481–503). Routledge.

References

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking,* 375–395. Springer. https://doi.org/10.1007/1-4020-2278-6_16

Cohen, A. L., & Staub, A. (2015). Within-subject consistency and between-subject variability in Bayesian reasoning strategies. *Cognitive Psychology*, *81*, 26–47. https://doi.org/10.1016/j.cogpsych.2015.08.001

Cohen, S. (1996). Identifying impediments to learning probability and statistics from an assessment of instructional software. *Journal of Educational and Behavioral Statistics*, *21*(1), 35–54. https://doi.org/10.2307/1165254

Cooper, L. L. (2002). *An assessment of prospective secondary mathematics teachers' preparedness to teach statistics* (Abstract of doctoral dissertation). University of Maryland.

Cooper, L. L. (2018). Assessing students' understanding of variability in graphical representations that share the common attribute of bars. *Journal of Statistics Education*, *26*(2), 110–124. https://doi.org/10.1080/10691898.2018.1473060

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, *16*(2), https://doi.org/10.1080/10691898.2008.11889559

Cooper, L. L., & Shore, F. S. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of Statistics Education*, *18*(2), https://doi.org/10.1080/10691898.2010.11889487

Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons and evaluative criteria. *Zeitschrift Für Soziologie*, *19*(6), 418–427. https://doi.org/10.1515/zfsoz-1990-0602

Corredor, J. A. (2008). *Learning statistical inference through computer-supported simulation and data analysis* [Unpublished doctoral dissertation]. University of Pittsburgh. http://d-scholarship.pitt.edu/6683/

Creswell, J. W. (2013). Qualitative inquiry and research design: Choosing among five approaches. 3rd Edition. Sage publications.

cTWO (2007). *Rijk aan betekenis. Visie op vernieuwd wiskundeonderwijs* [Rich in meaning. Vision for renewed mathematics education]. Commissie Toekomst Wiskunde Onderwijs. https://www.fisme.science.uu.nl/ctwo/publicaties/docs/Rijkaanbetekenisweb.pdf

Cui, L., & Liu, Z. (2021). Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review. *Attention*, *Perception*, *& Psychophysics*, *83*(3), 1290–1311.

Curcio, F. R. (1981). *The effect of prior knowledge, reading and mathematics achievement, and sex on comprehending mathematical relationships expressed in graphs* [Doctoral dissertation, Saint Francis Coll., Brooklyn, N.Y. Dept. of Education]. https://files.eric.ed.gov/fulltext/ED210185.pdf

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, *18*(5), 382–393. https://doi.org/ 10.2307/749086

Dabos, M. (2014). A glimpse of two year college instructors' understanding of variation in histograms. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1–4). ISI/IASE. https://icots.info/9/proceedings/pdfs/ICOTS9_C150_DABOS.pdf

Daemen, J., Konings, T., & van den Bogaart, T. (2020). Secondary School mathematics teacher education in the Netherlands. In M. Van den Heuvel-Panhuizen (Ed.), *National reflections on the Netherlands didactics of mathematics: Teaching and learning in the context of realistic mathematics education* (pp. 147–175). Springer. https://doi.org/10.1007/978-3-030-33824-4_9

Davis, A. B., Sumara, D. J., & Kieren, T. E. (1996). Cognition, co-emergence, curriculum. *Journal of Curriculum Studies*, *28*(2), 151–169.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79–95). Springer. https://doi.org/10.1007/1-4020-2278-6_4

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In *Proceedings of the fourth international research forum on statistical reasoning*, *literacy*, *and reasoning.* University of Auckland. http://apps3.cehd.umn. edu/artist/ articles/SRTL4_ARTIST.pdf

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58. http://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82. http://iaseweb.org/documents/ SERJ/SERJ4(1)_delMas_Liu.pdf

Delport, D. H. (2020). Teaching first-year statistics students with COVID-10 real-world data: Graphs. *Teaching Statistics*, *43*, 36–43. https://doi.org/10.1111/test.12245

Derouet, C., & Parzysz, B. (2016). How can histograms be useful for introducing continuous probability distributions? *ZDM Mathematics Education*, *48*(6), 757–773. https://doi.org/10.1007/s11858-016-0769-9

Dewhurst, R., Foulsham, T., Jarodzka, H., Johansson, R., Holmqvist, K., & Nyström, M. (2018). How task demands influence scanpath similarity in a sequential number-search task. *Vision Research*, *149*, 9–23. https://doi.org/ 10.1016/j.visres.2018. 05.006

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, *44*, 1079–1100. https://doi.org/10.3758/s13428-012-0212-2

References

Dickson, J., McLennan, J., & Omodei, M. M. (2000). Effects of concurrent verbalization on a time-critical, dynamic decision-making task. *The Journal of General Psychology*, *127*(2), 217–228.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system, *International Journal of Human-Computer Studies*, *70*(5), 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint* arXiv:1702.08608v2 [stat.ML] https://arxiv.org/abs/1702.08608

Drijvers, P. (2019). Embodied instrumentation: combining different views on using digital technology in mathematics education. In U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the 11th Congress of the European Society for Research in Mathematics Education* (pp. 8–28). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02436279v1

Dunn, P. K. (1999). A simple data set for demonstrating common distributions. *Journal of Statistics Education*, *7*(3). http://jse.amstat.org/jse_archive.htm#1999

Dunn, P. K. (2018). *Babyboom* [data file]. https://ww2.amstat.org/publications/jse/jse_data_archive.htm

Dzeroski, S., & Zenko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, *54*(3), 255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics, 9*(3), 586–596. https://www.jstor.org/stable/2240822

Eivazi, S., & Bednarik, R. (2010). Inferring problem solving strategies using eye-tracking: System description and evaluation. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research* (pp. 55–61). Association for Computing Machinery.

Eisenhart, M., & Howe, K. (1992). Validity in educational research. In M. LeCompte, W. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 64–680). Academic Press. http://www.elsevier.com/ wps/find/ bookdescription. cws_home/674919/description#description

Enders, F. (2013). Do clinical and translational science graduate students understand linear regression? Development and early validation of the regress quiz. *Clinical and Translational Science*, *6*(6), 444–51. https://doi.org/10.1111/cts.12088

Engel, J., Sedlmeier, P., & Wörn, C. (2008). Modelling scatter plot data and the signal-noise metaphor: towards statistical literacy for pre-service teachers. In C. Batanero, G. Burrill, C. Reading, & A. Rossman. (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference.* Springer.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research*, *41*(12), 1561–1574. https://www.sciencedirect.com/science/article/pii/S004269890000256X

Erickson, T. Wilkerson, M. , Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education*, *12*(1). https://doi.org/10.5070/T5121038001

Eshach, H., & Schwartz, J. L. (2002). Understanding children's comprehension of visual displays of complex information. *Journal of Science Education and Technology Journal of Science Education and Technology*, *11*(4), 333–346. https://doi.org/10.1023/A:1020690201324

Estepa, A., & Batanero, C. (1996). Judgments of correlation in scatterplots: Students' intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education*, *4,* 21–41.

Fabbri, S., Stubbs, K. M., Cusack, R., & Culham, J. C. (2016). Disentangling representations of object and grasp properties in the human brain. *The Journal of Neuroscience*, *36*(29), 7648–7662. https://doi.org/10.1523/JNEUROSCI.0313-16.2016

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals, *Journal of Clinical and Experimental Neuropsychology*, *28*(7), 1095–1112. https://doi.org/10.1080/13803390500205718

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letter*, *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feng, S., & Law, N. (2021). Mapping artificial intelligence in education research: A network-based keyword analysis. *International Journal of Artificial Intelligence in Education*, *31*, 277–303. https://doi.org/10.1007/s40593-021-00244-4

Fielding-Wells, J., & Hillman, J. (2018). Supporting young students emerging understandings of centre through modelling. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (pp. 1–6). ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_2B1.pdf?1531364242

Finzer, W. (2006). What does dragging this do? The role of dynamically changing data and parameters in building a foundation for statistical understanding. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1–4). ISI/IASE. https://iase-web.org/documents/papers/ icots7/7D4_FINZ.pdf?1402524965

Fisher, R. (1947). The analysis of covariance method for the relation between a part and the whole. *Biometrics*, *3*(2), 65–68. https://doi.org/10.2307/3001641

Fishman, B. J., Penuel, W. R., Allen, A. R., Cheng, B. H., & Sabelli, N. O. R. A. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *Teachers College Record*, *115*(14), 136–156.

References

Fleig, H., Meiser, T., Ettlin, F., & Rummel, J. (2017). Statistical numeracy as a moderator of (pseudo)contingency effects on decision behavior. *Acta Psychologica*, *174*, 68–79. https://doi.org/10.1016/j.actpsy.2017.01.002

Foster, C. (2011). Productive ambiguity in the learning of mathematics. *For the Learning of Mathematics*, *31*(2), 3–7.

Foxman, D. (1999). Mathematics textbooks across the world. Some evidence from the third international mathematics and science study. National Foundation for Educational Research.

Franklin, C. (2019). Foreword. In G. Burrill & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research. International perspectives* (pp. v–vi). Springer Nature. https://doi.org/10.1007/978-3-030-03472-6

Freedman, D., Pisani, R., & Purves, R. (1978). *Statistics.* W. W. Norton & Co.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). Springer series in statistics. https://link.springer.com/book/10.1007/978-0-387-21606-5

Friel, S. N., & Bright, G. W. (1995). Graph knowledge: Understanding how students interpret data using graphs. In *Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1–22). PME. https://eric.ed.gov/?id=ED391661

Friel, S. N., & Bright, G. W. (1996). Building a theory of graphicacy: How do students read graphs. In *Annual Meeting of the American Educational Research Association* (pp. 1–20). AERA. https://eric.ed.gov/?id=ed395277

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124–158. https://doi.org/10.2307/749671

Friendly, M. (2008). The golden age of statistical graphics. *Statistical Science*, *23*(4), 502–535. http://dx.doi.org/10.1214/08-STS268

Frischemeier, D. (2019). Statistical reasoning when comparing groups with software—Frameworks and their application to qualitative video data. In G. Burrill & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research. International perspectives* (pp. 283–305). Springer. https://doi.org/10.1007/978-3-030-03472-6_13

Frischemeier, D., & Biehler, R. (2016). Preservice teachers' statistical reasoning when comparing groups facilitated by software. In *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education* (pp. 643–650). ERME. https://hal.archives-ouvertes.fr/hal-01287058

Frischemeier, D., Podworny, S., & Biehler, R. (2023). Data visualization packages for non-inferential civic statistics in high school classrooms. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement: Teaching Civic Statistics to develop informed citizens.* Springer. https://doi.org/10.1007/978-3-031-20748-8_9

Fry, E. (1981). Graphical literacy. *Journal of Reading*, *24*(5), 383–389. https://www.jstor.org/stable/40032373

Fry, H. (2019). Hello world. How to be human in the age of the machine. Black Swan.

Fry, K., & Makar, K. (2021). How could we teach data science in primary school? *Teaching Statistics. 43*, S173–S181. https://doi.org/10.1111/test.12259

Fuchs, A. F. (1967). Saccadic and smooth pursuit eye movements in the monkey. *The Journal of Physiology*, *191*(3), 609–631. https://doi.org/10.1113/jphysiol.1967.sp008271

Gal, I. (1995). Statistical tools and statistical literacy: The case of the average. *Teaching Statistics*, *17*(3), 97–99. https://doi.org/10.1111/j.1467-9639.1995.tb00720.x

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education.* International Statistical Institute and IOS press. https://iase-web.org/Books.php?p=book1

Gal, I & Geiger, V. (2022). Welcome to the era of vague news: a study of the demands of statistical and mathematical products in the COVID-19 pandemic media. *Educational Studies in Mathematics*, *111*, 5–28. https://doi.org/10.1007/s10649-022-10151-7

Garcia Moreno-Esteva, E., Kervinen, A., Hannula, M. S., & Uitto, A. (2020). Scanning signatures: A graph theoretical model to represent visual scanning processes and a proof of concept study in biology Education. *Education Sciences*, *10*(5), Article 141. https://doi.org/10.3390/educsci10050141

Garcia Moreno-Esteva, E., White, S. L. J., Wood, J., & Black, A. (2016). Mathematical and computational modeling of eye-tracking data to predict success in a problem solving task. In *Proceedings of the 40th PME* (Vol. 1, p. 163). PME.

Garcia Moreno-Esteva, E., White, S. L. J., Wood, J. M., & Black, A. A. (2018). Application of mathematical and machine learning techniques to analyse eye-tracking data enabling better understanding of children's visual-cognitive behaviours. *Frontline Learning Research*, *6*(3), 72–84. https://doi.org/10.14786/flr.v6i3.365

Garfield, J. (2002). Histogram sorting. *Statistics teaching and resource library* (STAR). https://amser.org/index.php?P=AMSER--ResourceFrame&resourceId=8554

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy*, *reasoning and thinking* (pp. 397–409). Springer. https://doi.org/10.1007/1-4020-2278-6_17

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x

Garfield, J. B., & Ben-Zvi, D. (2008a). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer. https://doi.org/10.1007/978-1-4020-8383-9

Garfield, J. B., & Ben-Zvi, D. (2008b). Learning to reason about distribution. In J. Garfield & D. Ben-Zvi (Eds), *Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 165–186). Springer. https://link.springer.com/content/pdf/10.1007/978-1-4020-8383-9_8.pdf

Garfield, J., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review*, *67*(1), 1–12. https://doi.org/10.2307/1403562

Garnett, P. J., & Treagust, D. F. (1992). Conceptual difficulties experienced by senior high school students of electrochemistry: Electrochemical (galvanic) and electrolytic cells. *Journal of Research in Science Teaching*, *29*(10), 1079–1099. https://doi.org/10.1002/tea.3660291006

Gehrke, M., Kistler, T., Lübke, K., Markgraf, N., Krol, B., & Sauer, S. (2021). Statistics education from a data-centric perspective, *Teaching Statistics*, *43*, S201–S215. https://doi.org/10.1111/test.12264

Gerard, L., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K–12 education. *Educational Research Review*, *15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001

Getal en Ruimte [Numbers and Space] (2014). Vwo A/C, part 2, 11th edition.

Getal en Ruimte [Numbers and Space] (2015). 3 vmbo-KGT, part 1, 10th edition.

Gill, R. D. (2022). An Italian CSI drama: Social media. *Nieuw Archief voor Wiskunde*, *23*(5), 37–42. http://www.nieuwarchief.nl/serie5/pdf/naw5-2022-23-1-037.pdf

Gill, R. D., Groeneboom, P., & De Jong, P. (2018). Elementary statistics on trial—the case of Lucia de Berk. *Chance*, *31*(4), 9–15. https://doi.org/10.1080/09332480.2018.1549809

Gillespie, C. S. (1993). Reading graphic displays: What teachers should know. *Journal of Reading*, *36*(5), 350–354. https://www.jstor.org/stable/40033324

Gilmartin, K., & Rex, K. (2000). *Student toolkit: More charts, graphs and tables*. Open University. https://ahpo.net/assets/more-charts-graphs-and-tables-toolkit.pdf

Godau, C. Haider, H., Hansen, S., Schubert, T., Frensch, P. A., Gaschler, R. (2014). Spontaneously spotting and applying shortcuts in arithmetic—a primary school perspective on expertise. *Frontiers in Psychology – Cognition 5*, art. e556, 1664–1078. https://doi.org/10.3389/fpsyg.2014.00556

Goderis, B., Van Hulst, B., & Hoff, S. (2019). Waar ligt de armoedegrens? In *Armoede in kaart: 2019* [Where is the poverty line? Poverty mapped out: 2019]. https://digitaal.scp.nl/armoedeinkaart2019/waar-ligt-de-armoedegrens

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–606. https://nsuworks.nova.edu/tqr/vol8/iss4/6

Goldberg, J. H., & Helfman, J. I. (2010). Comparing information graphics: A critical look at eye tracking. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization* (pp. 71–78). Association for Computing Machinery. https://doi.org/10.1145/2110192.2110203

Goldberg, J., & Helfman, J. (2011). Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization*, *10*(3), 182–195. https://doi.org/10.1177/1473871611406623

González, O. (2014). Secondary mathematics teachers' professional competencies for effective teaching of variability-related ideas: A Japanese case study. *Statistique Et Enseignement*, *5*(1), 31–51. http://publications-sfds.fr/index.php/StatEns/article/view/299

González, O., & Chitmun, S. (2019). It's a good score! Just looks low: Using data-driven argumentation to engage students in reasoning about and modelling variability. In U. T. Jankvist, M. Van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education*. Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02411589

González, M. T., Espinel, M. C., & Ainley, J. (2011). Teachers' graphical competence. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education: A joint ICMI/IASE study*, (pp. 187–197). Springer. https://doi.org/10.1007/978-94-007-1131-0_20

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633. https://doi.org/10.1525/aa.1994.96.3.02a00100

Gorilla.sc (n.d.). *Eye tracking 2 zone*. https://support.gorilla.sc/support/tools/legacy-tools/task-builder-1/task-builder-zones#eyetracking2

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews.* Sage.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics, 43,* S11-S22.

Gravemeijer, K., & Doorman, M. (1999). Context problems in realistic mathematics education: a calculus course as an example. *Educational Studies in Mathematics*, *39*, 111–129.

Green, H. J., Lemaire, P., & Dufau, S. (2007). Eye movement correlates of younger and older adults' strategies for complex addition. *Acta Psychologica*, *125*(3), 257–278. https://doi.org/10.1016/j.actpsy.2006.08.001

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, *38*(5), 427–437. https://www.jstor.org/stable/30034960

Groth, R.E. (2009). Characteristics of teachers' conversations about teaching mean, median, and mode, *Teaching and Teacher Education*, *25*(5), 707–716. https://doi.org/10.1016/j.tate.2008.11.005

Groth, R. E., (2013). Characterizing key developmental understandings and pedagogically powerful ideas within a statistical knowledge for teaching framework. *Mathematical Thinking and Learning*, *15*(2), 121–145.

References

Groth, R. E. (2015). Research commentary: Working at the boundaries of mathematics education and statistics education communities of practice. *Journal for Research in Mathematics Education*, *46*(1), 4–16. https://www.jstor.org/stable/10.5951/jresematheduc.46.1.0004

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1253–1262). Association for Computing Machinery. https://doi.org/10.1145/1124772.1124961

Guerra-Carrillo, B. C., & Bunge, S.A. (2018). Eye gaze patterns reveal how reasoning skills improve with experience. *Science of Learning 3*, art. 18. https://doi.org/10.1038/s41539-018-0035-8

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), Article 93. https://doi.org/10.1145/3236009

Haddaway, N. R., Collins, A. M., Coughlin, D., Kirk, S. (2015). The role of google scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE 10*(9). Article e0138237. https://doi.org/10.1371/journal.pone.0138237

Hammer, D. (1996). More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research. *American Journal of Physics*, *64*(10), 1316–1325. https://doi.org/10.1119/1.18376

Hancox-Li, L. (2020). Robustness in machine learning explanations: Does it matter? In *Proceedings of the 2020 Conference on Fairness*, *Accountability*, *and Transparency*. ACM. https://doi.org/10.1145/3351095.3372836

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence 12*, 993–1001. https://doi.org/10.1109/34.58871

Harsh, J. A., Campillo, M., Murray, C., Myers, C., Nguyen, J., & Maltese, A. V. (2019). "Seeing" data like an expert: An eye-tracking study using graphical data representations. *LSE*, *18*(3), Article 32. https://doi.org/10.1187/cbe.18-06-0102

Harteis, C., Kok, E. M., & Jarodzka, H. (2018). The journey to proficiency: Exploring new objective methodologies to capture the process of learning and professional development. *Frontline Learning Research*, *6*(3), 1–5 https://doi.org/10.14786/flr.v6i3.435

Hawkins, A. (1997). Myth-conceptions! In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics. Proceedings of the 1996 IASE Round Table Conference* (pp. 1–14). ISI. https://www.stat.auckland.ac.nz/~iase/publications/8/1.Hawkins.pdf

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts, *The Clinical Neuropsychologist*, *24*(8), 1267–1278. https://doi.org/10.1080/13854046.2010.526785

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. C. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*(8), 1–23. https://doi.org/10.1098/rsos.180502

Heyes, S. D., Sinclair, S. J., Hoebee, S. E., & Morgan, J. W. (2020). How widespread are recruitment bottlenecks in fragmented populations of the savanna tree Banksia marginata (Proteaceae)? *Plant Ecology*, *221*, 545–557. https://doi.org/10.1007/s11258-020-01033-0

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis, *Learning and Individual Differences*, *66*, 4–15. https://doi.org/10.1016/j.lindif.2017.11.001.

Hinton-Bayre, A. D. (2010). Deriving reliable change statistics from test–retest normative data: Comparison of models and mathematical expressions. *Archives of Clinical Neuropsychology*, *25*(3), 244–256. https://doi.org/10.1093/arclin/acq008

Hoffmann, M. H. (2011). "Theoric transformations" and a new classification of abductive inferences. *Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy*, *46*(4), 570-590. https://doi.org/10.2979/trancharpeirsoc.2010.46.4.570

Hohn, R. W. (1992). *An analysis of the components of curriculum-based assessment* [Doctoral dissertation, University of Denver]. https://www.proquest.com/openview/abb4ea5179410900d6e2af9a7473f0e9/1?pq-origsite=gscholar&cbl=18750&diss=y

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: what it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 45–52). ACM Press.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press. https://global.oup.com/academic/product/eye-tracking-9780199697083?cc=nl&lang=en&

Holmqvist, K., Örbom, S. L., Hooge, I. T. C., Niehorster, D. C., Alexander, R. G., Andersson, R., Benjamins, J. S., Blignaut, P., Brouwer, A-M., Chuang, L. L., Dalrymple, K. A., Drieghe, D., Dunn, M. J., Ettinger, U., Fiedler, S., Foulsham, T., Van der Geest, J. N., Witzner Hansen, D., Hutton, S., … Hessels, R. S. (2023). Eye tracking: Empirical foundations for a minimal reporting guideline. *Behavior Research Methods, 55,* 364–416. https://doi.org/10.3758/s13428-021-01762-8

Huang, L. (2022). *Inquiry-based learning in lower-secondary mathematics education in China (Beijing) and the Netherlands* [Doctoral dissertation, Utrecht University].

Huck, S. W. (2016). Statistical misconceptions: Classic edition. Routledge.

Humphrey, P. B., Taylor, S., & Mittag, K. C. (2014). Developing consistency in the terminology and display of bar graphs and histograms. *Teaching Statistics: An*

International Journal for Teachers, *36*(3), 70–75. https://doi.org/10.1111/test. 12030

Hutto, D., & Abrahamson, D. (2022). Embodied, enactive education: Conservative versus radical approaches.

Hutto, D. D., & Sánchez-García, R. (2015). Choking RECtified: Embodied expertise beyond Dreyfus. *Phenomenology and the Cognitive Sciences*, *14*(2), 309–331.

Hutto, D., Kirchhoff, M.D., & Abrahamson, D. (2015). The enactive roots of STEM: Rethinking educational design in mathematics. *Educational Psychology Review*, *27*, 371–389. https://doi.org/10.1007/s10648-015-9326-2

Hwang, G-J., & Tu Y-F., (2021). Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics*, *9*(6), Article 584. https://doi.org/10.3390/math9060584

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction*, *20*(2), 172–176. https://doi.org/10.1016/j.learninstruc. 2009.02.013

IDSSP. International Data Science in School's Project Curriculum Team. (2019). *Curriculum frameworks for introductory data science*. http://www.idssp.org/ pages/about.html

Ioannidis, Y. (2003). The history of histograms (abridged). In *Proceedings 2003 VLDB Conference: 29th International Conference on Very Large Databases* (pp. 19–30). VLDB. https://doi.org/10.1016/B978-012722442-8/50011-2

Ismail, Z., & Chan, S. W. (2015). Malaysian students' misconceptions about measures of central tendency: An error analysis. In *Proceedings of the AIP Conference* (pp. 93–100). AIP. https://doi.org/10.1063/1.4907430

Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data* (Vol. 117). Sage. https://dx.doi.org/10.4135/9781412985963

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). Springer.

Janßen, T., Vallejo-Vargas, E., Bikner-Ahsbahs, A., & Reid, D. A. (2020). Design and investigation of a touch gesture for dividing in a virtual manipulative model for equation-solving. *Digital Experiences in Mathematics*, *6*, 166–190. https://doi.org/ 10.1007/s40751-020-00070-8

Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research*, *10*(1). https://doi.org/10.16910/jemr.10.1.3

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 211–218). ACM.

Jarodzka, H., Van Gog, T., Dorr, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction* 25, 62–70. https://doi.org/10.1016/j.learninstruc.2012.11.004

Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2019). What multimodal data can tell us about the students' regulation of their learning process? *Learning and Instruction*, Article 101203. https://doi.org/10.1016/j.learninstruc.2019.04.004

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, *30*(6), 1053–1079. https://doi.org/10.1207/s15516709cog0000_86

Kaakinen, J. K. (2021). What can eye movements tell us about visual perception processes in classroom contexts? Commentary on a special issue. *Educational Psychology Review*, *33*, 169–179. https://doi.org/10.1007/s10648-020-09573-7

Kang, J., Han, X., Song, J., Niu, Z., & Li, X. (2020). The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data. *Computers in Biology and Medicine*, *120*, Article 103722. https://doi.org/10.1016/j.compbiomed.2020.103722

Kaplan, J., Fisher, D. G., & Rogness, N. T. (2009). Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education*, *17*(3). https://doi.org/10.1080/10691898.2009.11889535

Kaplan, J. J., Gabrosek, J. G., Curtiss P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education*, *22*(2), 1–30. https://doi.org/ 10.1080/10691898.2014.11889701

Kaplan, J.J., Lyford, A., & Jennings, J. K. (2018). Effects of question stem on student descriptions of histograms. *Statistics Education Research Journal*, *17*(1), 85–102. https://iase-web.org/documents/SERJ/SERJ17(1)_Kaplan.pdf

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107

Karagiannakis, E. (2013). *The effect of an online course on social sciences university students' understanding of statistics* [Master's thesis, Utrecht University]. https://studenttheses.uu.nl/handle/20.500.12932/14448

Kazak S., Fielding, J., & Zapata-Cardona, L. (2022). Investigation cycle for analysing image-based data: perspectives from three contexts. Invited paper: refereed. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the Eleventh International Conference on Teaching Statistics* (pp. 1–6). ISI/IASE. https://doi.org/ 10.52041/iase.icots11.T8D1

Kelly, A. E., Sloane, F., & Whittaker, A. (1997). Simple approaches to assessing underlying understanding of statistical concepts. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 85–90). IOS Press. https://www.stat.auckland.ac.nz/~iase/publications/assessbk/chapter07.pdf

References

Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*, *1*. https://doi.org/10.3389/fdata.2018.00006

Khalil, K. A. I. (2005). *Expert-novice differences: Visual and verbal responses in a two-group comparison task* [Master's thesis, University of Massachusetts]. https://scholarworks.umass.edu/theses/2428

King, A. J., Bol, N., Cummins, R. G., & John, K. K. (2019). Improving visual behavior research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. *Communication Methods and Measures*, *13*(3), 149–177. https://doi.org/10.1080/19312458.2018.1558194

Klein, P., Becker, S., Küchemann, S., & Kuhn, J. (2021). Test of understanding graphs in kinematics: Item objectives confirmed by clustering eye movement transitions. *Physical Review Physics Education Research*, *17*(1), 013102.

Klein, P., Viiri, J., Mozaffari, S., Dengel, A., & Kuhn, J. (2018). Instruction-based clinical eye-tracking study on the visual interpretation of divergence: How do students look at vector field plots? *Physical Review Physics Education Research*, *14*(1), 010116. https://doi.org/10.1103/PhysRevPhysEducRes.14.010116

Knoop-Van Campen, C. A. N., Kok, E., Doornik, R. V., Vries, P. D., Immink, M., Jarodzka, H., & Van Gog, T. (2021). How teachers interpret displays of students' gaze in reading comprehension assignments. *Frontline Learning Research*, *9*(4), 116–140. https://doi.org/10.14786/flr.v9i4.881

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education*, *51*(1), 114–122. https://doi.org/10.1111/medu.13066

Kok, E. M., & Knoop-Van Campen, C.A.N. (2022). *Using webcam-based eye-tracking to uncover reading strategies*. Presentation at the EARLI SIG 27 conference, Southampton, UK.

Kok, E. M., Aizenman, A. M., Võ, M. L.-H., & Wolfe, J. M. (2017). Even if I showed you where you looked, remembering where you just looked is hard. *Journal of Vision*, *17*(12), 1–11. https://doi.org/10.1167/17.12.2

Konold, C. (2002). *Hat plots*? Unpublished manuscript, University of Massachusetts, Amherst.

Konold, C. (2007). Designing a data analysis tool for learners. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 267–291). Taylor & Francis. https://doi.org/10.4324/9780203810057

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics*, *88*(3), 305–325. https://doi.org/10.1007/s10649-013-9529-8

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289. https://www.jstor.org/stable/749741

Konold, C., & Pollatsek, A. (2004). Conceptualizing an average as a stable feature of a noisy process. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 169–199). Springer. https://doi.org/10.1007/1-4020-2278-6_8

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics. Proceedings of the 1996 IASE Round Table Conference* (pp. 151–167). ISI. https://iase-web.org/documents/papers/rt1996/13.Konold.pdf?1402524984

Kragten, M., Admiraal, W., & Rijlaarsdam, G. (2015). Students' learning activities while studying biological process diagrams. *International Journal of Science Education*, *37*(12), 1915–1937. https://doi.org/10.1080/09500693.2015.1057775

Kramarski, B. (1999). The study of graphs by computers: Is easier better? *Education Media International*, *36*(3), 203–209. https://doi.org/10.1080/0952398990360306

Kramarski, B. (2004). Making sense of graphs: Does metacognitive instruction make a difference on students' mathematical conceptions and alternative conceptions? *Learning and Instruction*, *14*(6), 593–619. https://doi.org/10.1016/j.learninstruc.2004.09.003

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686–5697). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858529

Krejtz, K., Duchowski, A. T., Krejtz, I., Kopacz, A., & Chrząstowski-Wachtel, P. (2016). Gaze transitions when learning with multimedia. *Journal of Eye Movement Research*, *9*(1), 1–17. https://doi.org/10.16910/jemr.9.1.5.

Król, M., & Król, M. (2019). Learning from peers' eye movements in the absence of expert guidance: A proof of concept using laboratory stock trading, eye tracking, and machine learning. *Cognitive Science*, *43*(2), e12716. https://doi.org/10.1111/cogs.12716

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kukar, M., & Kononenko, I. (2002). Reliable classifications with machine learning. In T. Elomaa, H. Mannila, H. Toivonen (Eds.), *European Conference on Machine Learning. Lecture notes in computer science*, *2430* (pp. 219–231). Springer.

Kukliansky, I. (2016). Student's conceptions in statistical graph's interpretation. *International Journal of Higher Education*, *5*(4), 262–267. https://eric.ed.gov/?id=EJ1119488

Kulm, G., Dager Wilson, L., & Kitchen, R. (2005). Alignment of content and effectiveness of mathematics assessment items. *Educational Assessment*, *10*(4), 333–356. https://doi.org/10.1207/s15326977ea1004_2

Lai, M., Tsai, M., Yang, F., Hsu, C., Liu, T., Lee, S. W., Lee, M., Chiou, G., Liang, J., & Tsai, C. (2013). A review of using eye-tracking technology in exploring learning from 2000

to 2012. *Educational Research Review*, *10*, 90–115. https://doi.org/10.1016/j.edurev.2013.10.001

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI*, *Ethics*, *and Society* (pp. 131–138). AAAI/ACM. https://doi.org/10.1145/3306618.3314229

Lakoff, G., & Núñez, R. (2000). Where mathematics comes from: How the embodied mind brings mathematics into being. Basic Books.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

LaPointe-McEwan, D., DeLuca, C. and Klinger, D. A. (2017), Supporting evidence use in networked professional learning: The role of the middle leader, *Educational Research*, *59*(2), 136–153. https://doi.org/10.1080/00131881.2017.1304346

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100. https://doi.org/10.1016/s0364-0213(87)80026-5

Lashley, K. S. (1951). *The problem of serial order in behavior*. Bobbs-Merrill.

Lawson, M. J. (1990). The case for instruction in the use of general problem-solving strategies in mathematics teaching: A comment on Owen and Sweller. *Journal for Research in Mathematics Education*, *21*(5), 403–410. https://doi.org/ 10.2307/749397

Leavy, A. (2006). Using data comparison to support a focus on distribution: Examining preservice teacher's understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal*, *5*(2), 89–114. http://iase-web.org/documents/SERJ/SERJ5(2)_Leavy.pdf

Lee, J. T. (1999). It's all in the area. *The Mathematics Teacher*, *92*(8), 670. https://www.jstor.org/stable/27971168

Lee, C., & Meletiou-Mavrotheris, M. (2003). Some difficulties of learning histograms in introductory statistics. In *Joint Statistical Meetings - Section on Statistical Education* (pp. 2326–2333). Amstat. http://www.statlit.org/pdf/2003leeasa.pdf

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal, 21*(2), Article 3. https://doi.org/10.52041/serj.v21i2.41

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research*, *60*(1), 1–64. https://doi.org/10.3102/00346543060001001

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2011). Coordinating between histograms and box plots. In *IASE Satellite Conferences: Statistics Education and Outreach*. IASE. https://iase-web.org/documents/papers/sat2011/ IASE2011Paper Poster6Lemetal.pdf?1402524996

Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013a). External representations for data distributions: in search of cognitive fit. *Statistics Education Research Journal*, *12*(1), 4–9. https://doi.org/10.52041/serj.v12i1.319

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013b). The heuristic interpretation of box plots. *Learning and Instruction*, *26*, 22–35. http://dx.doi.org/10.1016/j.learninstruc.2013.01.001

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013c). On the misinterpretation of histograms and box plots. *Educational Psychology*, *33*(2), 155–174. https://doi.org/10.1080/01443410.2012.674006

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2014a). Experts' misinterpretation of box plots – a dual processing approach. *Psychologica Belgica*, *54*(4), 395–405. http://dx.doi.org/10.5334/pb.az

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2014b). Interpreting histograms. As easy as it seems? *European Journal of Psychology of Education*, *29*(4), 557–575. https://doi.org/10.1007/s10212-014-0213-x

Lem, S., Kempen, G., Ceulemans, E., Onghena, P., Verschaffel, L., & Van Dooren, W. (2015). Combining multiple external representations and refutational text: An intervention on learning to interpret box plots. *International Journal of Science and Mathematics Education*, *13*(4), 909–926. https://doi.org/10.1007/s10763-014-9604-3

Leontyev, A. N. (2009). Activity and consciousness. *Marxist Internet Archive*. http://www.marxists.org/archive/leontev/works/activity-consciousness.pdf

Levitt, H. M. (2021). Qualitative generalization, not to the population but to the phenomenon. *Qualitative Psychology*, *8*(1), 95–110. https://doi.org/10.1037/qup000018

Liaw A., & Wiener M (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*(6), 1672–1682. https://ink.library.smu.edu.sg/lkcsb_research/5693

Lilienthal, A., & Schindler, M. (2019). Eye-tracking research in mathematics education: A PME literature review. In M. Graven, H. Venkat, A.A. Essien, & P. Vale (Eds.), *Proceedings 43rd Annual Meeting of the International Group for the Psychology of Mathematics Education, 4*, (pp. 62). PME. Extended version of this article: https://arxiv.org/ftp/arxiv/papers/1904/1904.12581.pdf

Lim, H. L., Wun, T. Y., & Chew, C. M. (2022). Assessing lower secondary school students' common errors in statistics. *Pertanika Journal of Social Sciences & Humanities*, *30*(3), 1427–1450. https://doi.org/10.47836/pjssh.30.3.26

Lovett, J. N., & Lee, H. S. (2019). Preservice secondary mathematics teachers' statistical knowledge: a snapshot of strengths and weaknesses. *Journal of Statistics Education*, *26*(3), 214–222. https://doi.org/10.1080/10691898.2018.1496806

References

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, *27*(1), 251–280. https://doi.org/10.1146/annurev.ps.27.020176.001343

Lyford, A. J. (2017). Investigating undergraduate student understanding of graphical displays of quantitative data through machine learning algorithms [Doctoral dissertation, University of Georgia]. https://iase-web.org/documents/dissertations/17.AlexanderLyford.Dissertation.pdf

Lyford, A., & Boels, L. (2022). Using machine learning to understand students' gaze patterns on graphing tasks. Invited paper: refereed In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the Eleventh International Conference on Teaching Statistics* (pp. 1–6). ISI/ IASE. https://doi.org/10.52041/iase.icots11.T8D2

Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal, 29*, 861–878.

Madden, S. R. (2008). *High school mathematics teachers' evolving understanding of comparing distributions* [Doctoral dissertation, Western Michigan University]. https://scholarworks.wmich.edu/dissertations/792/

Makar, K., & Confrey, J. (2003). Chunks, clumps, and spread out: Secondary preservice teachers' informal notions of variation and distribution. In C. Lee (Ed.), *Reasoning about variability: A collection of current research studies*. Kluwer Academic Publisher. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.532.4111&rep=rep1&type=pdf

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353–374). Springer.

Martin, M. A. (2003). "It's like...you know": The use of analogies and heuristics in teaching introductory statistical methods. *Journal of Statistics Education*, *11*(2). https://doi.org/10.1080/10691898.2003.11910705

Martineau, H. (1859). *England and her soldiers.* Cambridge University Press. https://doi.org/10.1017/CBO9780511751301

Mason, L., & Otero, M. (2021). Just how effective is direct instruction? *Perspectives on Behavior Science*, *44*, 225–244. https://doi.org/10.1007/s40614-021-00295-x

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1290–1294). ACM. https://doi.org/10.1145/3025453.3025912

Mathplus. (2014). *Vwo/gymnasium 1, deel B* [Pre-university track/gymnasium 1, part B]. Malmberg.

McCampbell, S. M. (2014). *Pre-service teachers' self-efficacy for teaching mathematics* [Doctoral dissertation, University of New Mexico]. https://digitalrepository.unm.edu/educ_ifce_etds/28/

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics*, *45*(1), 103–129. https://doi.org/10.1023/A:1013874514650

McGatha, M., Cobb, P., & McClain, K. (2002). An analysis of students' initial statistical understandings: Developing a conjectured learning trajectory. *The Journal of Mathematical Behavior*, *21*(3), 339–355. https://doi.org/10.1016/S0732-3123(02)00133-5

McGee, J. R. (2012). *Developing and validating a new instrument to measure the self-efficacy of elementary mathematics teachers* [Doctoral dissertation, University of North Carolina]. http://libres.uncg.edu/ir/uncc/f/McGee_uncc_0694D_10291.pdf

McGovern, A., Lagerquist, R., Gagne, D.J. II, Jergensen, G.E., Elmore, K.L., Homeyer, C.R., & Smith, T. (2019). Making the black box more transparent. Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199 https://doi.org/10.1175/BAMS-D-18-0195.1

McIntyre, N. A., Draycott, B., & Wolff, C. E. (2022). Keeping track of expert teachers: Comparing the affordances of think-aloud elicited by two different video perspectives. *Learning and Instruction*, *80*, Article 101563, 1–12. https://doi.org/10.1016/j.learninstruc.2021.101563

McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research.* Routledge.

McKinney, M. (2015). The relationship between elementary teachers' self-efficacy for teaching mathematics and their mathematical knowledge for teaching [Master's thesis, Boise State University]. https://scholarworks.boisestate.edu/td/944/

McNamara, A. (2016). *On the state of computing in statistics education: Tools for learning and for doing*. arXiv preprint arXiv:1610.00984. https://doi.org/10.48550/arXiv.1610.00984

McNamara, A. (2018). Imagining the future of statistical education software. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (pp. 1–3). ISI/IASE. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_1B2.pdf

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Meletiou, M. (2000). *Students' understanding of variation: An untapped well in statistical reasoning* [Doctoral dissertation, University of Texas]. https://iase-web.org/documents/dissertations/00.Meletiou.Dissertation.pdf

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the Sixth International Conference on Teaching Statistics* (pp. 1–4). ISI/IASE. http://iase-web.org/documents/papers/icots6/10_19_me.pdf?1402524959

References

Meletiou-Mavrotheris, M. (2003). Technological tools in the introductory statistics classroom: Effects on student understanding of inferential statistics. *International Journal of Computers for Mathematical Learning*, *8*, 265–297. https://doi.org/10.1023/B:IJCO.0000021794.08422.65

Meletiou-Mavrotheris, M., & Lee, C. (2005). Exploring introductory statistics students' understanding of variation in histograms. In M. Bosch (Ed.), *Proceedings of the 4th Congress of the European Society for Research in Mathematics Education*. ERME. http://erme.site/wp-content/uploads/2021/06/CERME4_WG5.pdf

Meletiou-Mavrotheris, M., & Stylianou, D. A. (2004). On the formalist view of mathematics: Impact on statistics instruction and learning. In M. A. Mariotti (Ed.), *European Research in Mathematics Education III: Proceedings of the Third Conference of the European Society for Research in Mathematics Education*. University of Pisa and ERME. http://erme.site/wp-content/uploads/2021/06/cerme3_tg5.zip

Mevarech, Z. R., & Kramarsky, B. (1997). From verbal descriptions to graphic representations: Stability and change in students' alternative conceptions. *Educational Studies in Mathematics*, *32*(3), 229–263. https://doi.org/10.1023/A:1002965907987

Mitchell, T. M., Buchanan, B., De Jong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. *Annual Review of Computer Science*, *4*, 417–433. https://doi.org/10.1146/annurev.cs.04.060190.002221

Mitev, N., Renner, P., Pfeiffer, T., & Staudte, M. (2018). Towards efficient human–machine collaboration: Effects of gaze-driven feedback and engagement on performance. *Cognitive Research: Principles and Implications*, *3*, Article 51. https://doi.org/10.1186/s41235-018-0148-x

Moderne Wiskunde [Modern Math] (2019). Vwo AC, part 1, 12th edition. Noordhoff Uitgevers.

Moderne Wiskunde [Modern Math] (2015). 11th edition. Noordhoff Uitgevers.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26(1), 20–39. https://doi.org/10.2307/749226

Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2012). Brain regions with mirror properties: A meta-analysis of 125 human fMRI studies. *Neuroscience & Biobehavioral Reviews*, *36*(1), 341–349. https://doi.org/10.1016/j.neubiorev.2011.07.004

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/index.html

Najar, A. S., Mitrovic, A., & Neshatian, K. (2014). Utilizing eye tracking to improve learning from examples. In C. Stephanidis & M. Antona (Eds.), *Lecture Notes in Computer Science 8514. Proceedings of the Eighth International Conference on Universal Access in Human-Computer Interaction part 2* (pp. 410–418). Springer. https://doi.org/10.1007/978-3-319-07440-5_38

Nijdam, A. D. (2003). *Statistiek in onderzoek*, *1 Beschrijvende technieken* [Statistics in research, 1 Descriptive techniques]. Noordhoff.

Nolan, D., & Perrett, J. (2016). Teaching and learning data visualization: Ideas and assignments. *The American Statistician*, *70*(3), 260–269.

Norabiatul Adawiah binti Abd Wahid, Suzieleez Syrene binti Abdul Rahim, & Sharifah Norul Akmar binti Syed Zamri (2021). Pre-service teachers perceptions to outliers in statistical graphs. In *Proceedings Annual South East Asian Association for Institutional Research Conference. SEAAIR 2020/21: Diversity in Education. Vol. 1* (pp. 296–304). SEAAIR. https://mojes.um.edu.my/index.php/MOJES/article/ view/ 39418

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, *11*(9), 929–949. https://doi.org/10.1016/ 0042-6989(71)90213-6

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy*, *9*(1), article 4, 1–27. https://doi. org/10.5038/1936-4660.9.1.4

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature,* 506, 150–152. http://www.nature.com/news/scientific-methodstatistical-errors-1.14700

O'Dell, R. S. (2012). The mean as balance point. *Mathematics Teaching in the Middle School*, *18*(3), 148–155. https://doi.org/10.5951/mathteacmiddscho.18.3.0148

Obersteiner, A., Tumpek, C. (2016). Measuring fraction comparison strategies with eye-tracking. *ZDM Mathematics Education*, *48*, 255–266. https://doi.org/10.1007/ s11858-015-0742-z

Oehrtman, M. (2008). Layers of abstraction: Theory and design for the instruction of limit concepts. In M. Carlson & C. Rasmussen. (Eds.), *Making the connection: Research and teaching in undergraduate mathematics education.* (pp. 65–80). https://www.maa.org/sites/default/files/pdf/pubs/books/members/NTE73.pdf#p age=78

Olande, O. (2014). A case study on pre-service teacher students' interaction with graphical artefacts. *REDIMAT Journal of Research in Mathematics Education*, *3*(1), 73–102. https://doi.org/10.4471/redimat.2014.41

Oohira, A., Okamoto, M., & Ozawa, T. (1981). 正常人の衝動性眼球運動最大速度につ いて[Peak velocity of normal human saccadic eye movements (author's translation)]. 日限会誌 [*Journal of the Japanese Society of Ophthalmology*], *85*(11), 2001–2007. https://pubmed.ncbi.nlm.nih.gov/7337121/ or https://www. Research gate.net/publication/15862222_Peak_velocity_of_normal_human_saccadic_eye_ movements_author%27s_transl

Orquin, J. L., & Holmqvist, K. (2017). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods*, *50*(4), 1645–1656. https://doi.org/ 10.3758/s13428-017-0998-z

References

Pareja Roblin, N., Schunn, C., & McKenney, S. (2018). What are critical features of science curriculum materials that impact student and teacher outcomes? *Science Education*, *102*(2), 260–282. https://doi.org/10.1002/sce.21328

Pastore, M., Lionetti, F., & Altoè, G. (2017). When one shape does not fit all: A commentary essay on the use of graphs in psychological research. *Frontiers in Psychology*, *8*, 1666. https://doi.org/10.3389/fpsyg.2017.01666

Pearson, K. (1895). X. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, *186*, 343–424. https://doi.org/10.1098/rsta.1895.0010

Peebles, D., & Cheng, P. C. (2001). Graph-based reasoning: From task analysis to cognitive explanation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23.* https://escholarship.org/uc/item/9rz4r25j

Peirce, C. S. (1994). Volume 6: Scientific Metaphysics*. In C. Hartshorne & P. Weiss (Eds.), *The collected papers of Charles Sanders Peirce (1958-1966)* [Electronic edition]. Belknap Press of Harvard University. http://pm.nlx.com/xtf/view?docId=peirce/peirce.06.xml;chunk.id=div.peirce.cp1.24;toc.depth=1;toc.id=div.peirce.cp1.24;brand=default

Peters, S. A., & Stokes-Levine, A. (2019). Secondary teachers' learning: Measures of variation. In G. Burrill & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research. International perspectives* (pp. 245–264). Springer. https://doi.org/10.1007/978-3-030-03472-6_11

Pettibone, T. J., & Diamond, J. J. (1972). An incorrect index of skewness. *Resources in Education*, *8*(1–4), 1–8. https://eric.ed.gov/?id=ED068579

Pfannkuch, M., & Ben-Zvi, D. (2011). Developing teachers' statistical thinking. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 323–333). New ICMI Study Series *14.* Springer. https://doi.org/10.1007/978-94-007-1131-0_31

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal*, *5*(2), 4–9, http://www.stat.auckland.ac.nz/serj

Piaget, J. (1952). The origins of intelligence in children. New York.

Płomecka, M. B., Barańczuk-Turska, Z., Pfeiffer, C., & Langer, N. (2020). Aging effects and test–retest reliability of inhibitory control for saccadic eye movements. *eNeuro*, *7*(5) https://doi.org/10.1523/ENEURO.0459-19.2020

Prodromou, T., & Pratt, D. (2006). The role of causality in the co-ordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, *5*(2), 69–88. http://iase-web.org/documents/SERJ/SERJ5(2)_Prod_Pratt.pdf

Radford, L. (2010). The eye as a theoretician: Seeing structures in generalizing activities. *For the Learning of Mathematics*, *30*(2), 2–7. http://www.jstor.org/stable/20749442

Radford, L., Schubring, G., & Seeger, F. (2011). Signifying and meaning-making in mathematical thinking, teaching, and learning. *Educ Stud Math*, *77*, 149–156. https://doi.org/10.1007/s10649-011-9322-5

Reading, C., & Canada, D. (2011). Teachers' knowledge of distribution. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study*, (pp. 223–234). Springer.

Redfern, N. (2011, January 13). *Bar chart or histogram?* https://nickredfern.wordpress.com/2011/01/13/bar-chart-or-histogram/

Reinhart, A., Evans, C., Luby, A., Orellana, J., Meyer, M., Wieczorek, J., Elliott, P., Burckhardt, P., & Nugent, R. (2022). Think-aloud interviews: A tool for exploring student statistical reasoning. *Journal of Statistics and Data Science Education*, *30*(2), 100–113. https://doi.org/10.1080/26939169.2022.2063209

Reitemeyer, M. (2017). *Promoting productive struggle in middle school math* [Doctoral dissertation, University of Delaware.]

Ridgway, J., & Nicholson, J. (2019). Problematising high-stakes assessment in statistics. In J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística*. Universidad de Granada. www.ugr.es/local/fqm126/civeest.html

Ridgway, J., Nicholson, J., & McCusker, S. (2009). The next great leap–from official data to public knowledge. In Satellite: Next steps in statistics education. IASE/ISI. https://iase-web.org/documents/papers/sat2009/4_2.pdf

RIVM (2022). [data] Original link: https://data.rivm.nl/meta/srv/dut/catalog.search;jsessionid=8C7C4450AE5BB96DC4C6F9701E53BF7E#/metadata/2c4357c8-76e4-4662-9574-1deb8a73f724?tab=relations *Permanent link*: https://data.overheid.nl/dataset/34ff61d8-17a6-49a6-94db-90826562f8b5

Rodríguez-Muñiz, L. J., Díaz, P., & Muñiz-Rodríguez, L. (2018). Statistics and probability in the Spanish baccalaureate: intended curriculum and implementation in textbooks. In Y. Shimizu & R. Vithal (Eds.), *24th ICMI study conference. School mathematics curriculum reforms: challenges, changes and opportunities. Conference proceedings* (pp. 413–420). ICME.

Rodríguez-Muñiz, L. J., Muñiz-Rodríguez, L., García-Alonso, I., López-Serentill, P., Vásquez C., & Alsina, À. (2022). Navigating between abstraction and context in secondary school statistics education (Nadando entre dos orillas: Abstracción y context en educación estadística en Secundaria), *Culture and Education.* https://doi.org/10.1080/11356405.2022.2058794

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. WSPC. https://doi.org/10.1142/9097

Roth, J. A. (2019). *Making the struggle productive: Conceptualizing the role and impact of the mathematics teacher in episodes of productive struggle* [Doctoral dissertation, Kennesaw State University].

References

Roth, W. (2005). Mathematical inscriptions and the reflexive elaboration of understanding: An ethnography of graphing and numeracy in a fish hatchery. *Mathematical Thinking and Learning*, *7*(2), 75–110. https://doi.org/ 10.1207/ s15327833mtl0702_1

Roth, W.-M., & Bowen, G. M. (2001). Professionals read graphs: A semiotic analysis. *Journal for Research in Mathematics Education*, *32*(2), 159–194. https://doi.org/ 10.2307/749672

Rowlands, M. J. (2010). The new science of the mind. From extended mind to embodied phenomenology. The MIT Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rufilanchas, D. R. (2017). On the origin of Karl Pearson's term "histogram" [Acerca del origen del término 'histograma' acuñado por Karl Pearson]]. *Estadistica Espanola*, *59*(192), 29–35. https://dialnet.unirioja.es/ejemplar/478889

Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (1999). *Student science journals and the evidence they provide: Classroom learning and opportunity to learn* [Paper presentation]. The Annual Meeting of the National Association for Research in Science Teaching, Boston, MA, USA. https://eric.ed.gov/?id=ED431796

Rumsey, D.J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002. 11910678

Russo, J. E. (2010). Eye fixations as a process trace. In M. Schulte-Mecklenbeck, A. Kühberger, & R. Ranyard (Eds.), *Handbook of process tracing methods for decision research* (pp. 43–64). Psychology Press. https://doi.org/10.4324/9781315160559

Sahann, R., Müller, T., & Schmidt, J. (2021). Histogram binning revisited with a focus on human perception. In *IEEE Visualization Conference (VIS)* (pp. 66–70). IEEE. https:// doi.org/10.1109/VIS49827.2021.9623301

Saidi, S. S., & Siew, N. M. (2019). Reliability and validity analysis of statistical reasoning test survey instrument using the Rasch measurement model. *International Electronic Journal of Mathematics Education*, *14*(3), 535–546. https://doi.org/ 10.29333/iejme/5755

Saldanha, L., & Hatfield, N. (2021). Students conceptualizing the box plot as a tool for structuring quantitative data: a design experiment using TinkerPlots. *Canadian Journal of Science*, *Mathematics and Technology Education*, *21,* 758–782. https:// doi.org/10.1007/s42330-021-00184-0

Scharfen, J., Jansen, K., & Holling, H. (2018). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin & Review 25*, 2175–2199. https://doi. org/10.3758/s13423-018-1461-6

Scheiter, K., Schubert, C., Schülera, A., Schmidt, H., Zimmermann, G., Wassermann, B., Krebsa, M.-C., & Edera, T. (2019). Adaptive multimedia: Using gaze-contingent

instructional guidance to provide personalized processing support. *Computers & Education*, *139,* 31–47. https://doi.org/10.1016/j.compedu.2019.05.005

Schindler, M., & Lilienthal, A. J. (2019). Domain-specific interpretation of eye-tracking data: Towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics*, *101*, 123–139. https://doi.org/10.1007/s10649-019-9878-z

Schindler, M., & Lilienthal, A. J. (2020). Students' creative process in mathematics: Insights from eye-tracking-stimulated recall interview on students' work on multiple solution tasks. *International Journal of Science and Mathematics Education*, *18*, 1565–1586. https://doi.org/10.1007/s10763-019-10033-0

Schindler, M., Schaffernicht, E., & Lilienthal, A. J. (2021). Identifying student strategies through eye tracking and unsupervised learning: The case of quantity recognition. In M. Inprasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the Forty-fourth Conference of the International Group for the Psychology of Mathematics Education*, *4* (pp. 9–16).. PME. https://www.igpme.org/wp-content/uploads/2022/04/Volume-4_final.pdf

Schreiter, S., & Vogel, M. (2023). Eye-tracking measures as indicators for a local vs. global view of data. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.1058150

Setiawan, E. P., & Sukoco, H. (2021). Exploring first year university students' statistical literacy: A case on describing and visualizing data. *Journal on Mathematics Education*, *12*(3), 427–448. http://doi.org/10.22342/jme.12.3.13202.427-448

Setlur, V., Correll, M., & Battersby, S. (2022). OSCAR: A semantic-based data binning approach. *arXiv Computer Science*, *Human-Computer Interaction*. https://arxiv.org/abs/2207.07727

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. J. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Information Age Publishing.

Shayan, S., Abrahamson, D., Bakker, A., Duijzer, A., & Van der Schaaf, M.F. (2017). Eye-tracking the emergence of attentional anchors in a mathematics learning tablet activity. In C. Was, F. Sansosti, & B. Morris (Eds.), *Eye-tracking technology applications in educational research* (pp. 166–194). IGI-Global. https://doi.org/10.4018/978-1-5225-1005-5.ch009

Shvarts, A. (2017). Eye movements in emerging conceptual understanding of rectangle area. In Kaur, B., Ho, W. K., Toh, T. L., & Choy, B. H. (Eds.), *Proceedings of the 41st Conference of the International Group for the Psychology of Mathematics Education*, *Vol. 1*, (p. 268). PME.

Shvarts, A. (2018). A dual eye-tracking study of objectification as students-tutor joint activity appropriation. In *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education, Vol 4*, (pp. 171–178). PME.

Shvarts, A., & Abrahamson, D. (2019). Dual-eye-tracking Vygotsky: A microgenetic account of a teaching/learning collaboration in an embodied-interaction

technological tutorial for mathematics. *Learning*, *Culture and Social Interaction*, *22*, Article 100316. https://doi.org/10.1016/j.lcsi.2019.05.003

Shvarts, A., & Alberto, R. A. (2021). Melting cultural artifacts back to personal actions: embodied design for a sine graph. In M. Prasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education*, *Vol. 4*, (pp. 49–56). PME.

Shvarts, A., & Van Helden, G. (2021). Embodied learning at a distance: from sensory-motor experience to constructing and understanding a sine graph. *Mathematical Thinking and Learning*. https://doi.org/10.1080/10986065.2021.1983691

Shvarts, A., Alberto, R., Bakker, A., Doorman, M., & Drijvers, P. (2019). Embodied instrumentation: Reification of sensorimotor activity into a mathematical artifact. In B. Barzel, R. Bebernik, L. Göbel, M. Pohl, H. Ruchniewicz, F. Schacht, & D. Thurm (Eds.), *Proceedings of the 14th International Conference on Technology in Mathematics Teaching–ICTMT 14*. DuEPublico. https://doi.org/10.17185/due publico/70749

Shvarts, A., Alberto, R. A., Bakker, A., Doorman, M., & Drijvers, P. (2021). Embodied instrumentation in learning mathematics as the genesis of a body-artifact functional system. *Educational Studies in Mathematics 107*, 447–469. https://doi.org/ 10.1007/s10649-021-10053-0

Simon, M. (2020). Hypothetical learning trajectories in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 354–357). Springer. https://doi.org/10.1007/978-3-030-15789-0

Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: an elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning, 6*(2), 91–104. https://doi.org/10.1207/s15327833mtl0602_2

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching*, *77*, 20–26. http://www.davidtall.com/skemp/pdfs/ instrumental-relational.pdf

Slauson, L. V. (2008). *Students' conceptual understanding of variability* [Doctoral dissertation, Ohio State University]. http://rave.ohiolink.edu/etdc/view?acc_num= osu1199117318

Sorto, M. A. (2004). *Prospective middle school teachers' knowledge about data analysis and its application to teaching* [Doctoral dissertation, Michigan State University].

Spivey, M. J., & Dale, R. (2011). Eye movements both reveal and influence problem solving. In S. P. Liversedge, I. Gilchrist & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 551–562)*. Oxford University Press. https://doi.org/10.1093/ oxfordhb/9780199539789.013.0030

SRTL-12 (2020). *SRTL12 preliminary announcement*. http://www.srtl.info

Stevens, S. P., & Palocsay, S. W. (2012). Identifying addressable impediments to student learning in an introductory statistics course. *INFORMS Transactions on Education*, *12*(3), 124–139. https://doi.org/10.1287/ited.1120.0085

Stone, A. (2006). *A psychometric analysis of the statistics concept inventory* [Doctoral dissertation, University of Oklahoma]. https://hdl.handle.net/11244/1013

Strasser, N. (2007). Avoiding statistical mistakes. *Journal of College Teaching & Learning (TLC)*, *4*(7), 51–55. https://doi.org/10.19030/tlc.v4i7.1565

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. M. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics*, *104*, 147–200*.* https://doi.org/10.1007/s10649-020-09948-1

Susac, A. N., Bubic, A., Kaponja, J., Planinic, M., & Palmovic, M. (2014). Eye movements reveal students' strategies in simple equation solving. *International Journal of Science and Mathematics Education*, *12*(3), 555–577. https://doi.org/10.1007/s10763-014-9514-4

Sweller, J. (1990). On the limited evidence for the effectiveness of teaching general problem-solving strategies. *Journal for Research in Mathematics Education*, *21*(5), 411–415. https://doi.org/10.2307/749398

Tacoma, S. G., Heeren, B. J., Jeuring, J. T., & Drijvers, P. H. M. (2019). Automated feedback on the structure of hypothesis tests. In U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* (pp. 2969–2976)*.* Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal. archives-ouvertes.fr/hal-02428867v1

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education*, *29*(2), 185–208. https://doi.org/10.1080/17437270600891614

Tall, D. ., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics, 12,* 151–169.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*(4), 357–369. https://doi.org/ 10.1017/s1355617799544068

TI (2015). Building concepts: Mean as a balance point. https://education.ti.com/en/building-concepts/activities/statistics/sequence1/ mean-as-balance-point

Tiefenbruck, B. F. (2007). *Elementary teachers* [sic] *conceptions of graphical representations of categorical data* [Doctoral dissertation, University of Minnesota]. https://conservancy.umn.edu/handle/11299/91699

Tintle, N. L., & Vander Stoep, J. (2018). Development of a tool to assess students' conceptual understanding in introductory statistics. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (pp. 1–4). ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3B1.pdf?1531364253

References

Tobii (n.d.-a). *Tobii Studio. Users' Manual.* Version 3.4.5. https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf/?v=3.4.5

Tobii (n.d.-b). *Specification of gaze accuracy and gaze precision*, *Tobii X2-60 eye tracker.* https://eyetracking.ch/wordpress/wp-content/uploads/2011/12/Tobii_X2-60_Eye_Tracker_Technical_Specification.pdf

Tracy, S. J. (2013). *Qualitative Research Methods*. John Wiley & Sons.

Tufte, E. R. (2001). *The visual display of quantitative information.* Graphics Press. (Original work published 1983) https://www.edwardtufte.com/tufte/books_vdqi

Tukey, J. W. (1972). Some graphic and semigraphic displays. In T.A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor* (pp. 293–316). Iowa State University Press. https://www.edwardtufte.com/tufte/tukey

Turegun, M., & Reeder, S. (2011). Community college students' conceptual understanding of statistical measures of spread. *Community College Journal of Research and Practice*, *35*(5), 410–426. https://doi.org/10.1080/10668920903381854

Tversky, B. (1997). Cognitive principles of graphic displays. In *AAAI Fall Symposium on Reasoning with Diagrammatic Representations II* (pp. 8–10). AAAI Press. https://www.aaai.org/Papers/Symposia/Fall/1997/FS-97-03/FS97-03-015.pdf

Twining, P., Heller, R. S., Nussbaum, M., Tsai, C.-C. (2017). Some guidance on conducting and reporting qualitative studies, *Computers & Education*, *106*, A1–A9. https://doi.org/10.1016/j.compedu.2016.12.002

Van de Schoot, R. (2020). *Machines vervangen wetenschappers* [Machines replace scientists]. https://www.uu.nl/agenda/oratie-machines-vervangen-wetenschappers

Van den Ham, A. K., & Heinze, A. (2018). Does the textbook matter? Longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation*, *59,* 133–140.

Van der Gijp, A. Ravesloot, C. J., Jarodzka, H., Van der Schaaf, M. F., Van der Schaaf, I. C., Van Schaik, J. P. J., & Ten Cate, Th. J. (2017). How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education*, *22*(3), 765–787. https://doi.org/10.1007/s10459-016-9698-1

Van der Stigchel, S., Hessels, R. S., van Elst, J. C., & Kemner, C. (2017). The disengagement of visual attention in the gap paradigm across adolescence. *Experimental Brain Research*, *235*(12), 3585–3592.

Van Dijke-Droogers, M. (2021). *Introducing statistical inference: Design and evaluation of a learning trajectory* [Doctoral dissertation, Utrecht University]. https://www.fisme.science.uu.nl/publicaties/literatuur/2021_van_dijke_introducing_statistical_inferences.pdf

Van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143–156). Springer. https://doi.org/10.1007/978-1-4419-5546-3_10

Van Gog, T., Paas, F., Van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied*, *11*(4), 237–244. https://doi.org/10.1037/1076-898X.11.4.237

Van Marlen, T., Van Wermeskerken, M., Jarodzka, H., Raijmakers, M., & Van Gog, T. (2022). Looking through Sherlock's eyes: Effects of eye movement modelling examples with and without verbal explanations on deductive reasoning. *Journal of Computer Assisted Learning*, *38*(5), 1497–1506. https://doi.org/10.1111/jcal.12712

Van Meeuwen, L. W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P. A., de Bock, Jeano J. P. R., & van Merriënboer, J. G. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction*, *32*, 10–21. https://doi.org/10.1016/j.learninstruc.2014.01.004

Van Zanten, M. (2020). *Opportunities to learn offered by primary school mathematics textbooks in the Netherlands* [Doctoral dissertation, Utrecht University]. https://dspace.library.uu.nl/handle/1874/399577

Vermette, S., & Savard, A. (2019). Necessary knowledge for teaching statistics: example of the concept of variability. In G. Burrill & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research. International perspectives* (pp. 225–244). Springer. https://doi.org/10.1007/978-3-030-03472-6_10

Vermette, S., & Gattuso, L. (2014). High school teachers' pedagogical content knowledge of variability. Invited paper. Refereed. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1–6). ISI/IASE. http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_2F3_VERMETTE.pdf

Verschut, A., & Bakker, A. (2010). Towards evaluation criteria for coherence of a data-based statistics curriculum. Invited paper. Refereed. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics* (pp. 1–6). ICOTS. https://iase-web.org/documents/papers/icots8/ICOTS8_3E1_VERSCHUT.pdf

Verschut, A., & Bakker, A. (2011). Implementing a more coherent statistics curriculum. In M. Pytlak, T. Rowland, & E. Swoboda (Eds.), *Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education (CERME 7).* ERME.

Villagrá-Arnedo, C. J., Gallego-Durán, F. J., Llorens-Largo, F., Compañ-Rosique, P., Satorre-Cuerda, R., & Molina-Carmona, R. (2017). Improving the expressiveness of black-box models for predicting student performance. *Computers in Human Behavior*, *72*, 621–631. https://doi.org/10.1016/j.chb.2016.09.001

Voisin, S., Pinto, F., Morin-Ducote, G., Hudson, K. B., & Tourassi, G. D. (2013). Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography. *Medical Physics*, *40*(10), 101906-01–101906-10. https://doi.org/10.1118/1.4820536

References

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674576292

Vygotsky, L. S. (1997). *Educational Psychology* (R. H. Silverman Translation). CRC Press LLC. (Original work published 1926). https://www.taylorfrancis.com/chapters/mono/10.4324/9780429273070-9/

Wade, N. J., & Tatler, B. W. (2011). Origins and applications of eye movement research. In S. P. Liversedge, I. Gilchrist & S. Everling (Eds.), *The Oxford handbook on eye movements* (pp. 17–46). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0002

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*(2), 145–168.

Watson, J. M., & Shaughnessy, J. M. (2004). Proportional reasoning: lessons from research in data and chance. *Mathematics Teaching in the Middle School*, *10*(2), 104–109. http://www.jstor.org/stable/41182026

Watts, C., Moyer-Packenham, P. S., Tucker, S. I., Bullock, E. P., Shumway, J., Westenskow, A., Boyer-Thurgood, J., Anderson-Pence, K., Mahamane, S. & Jordan, K. (2016). An examination of children's learning progression shifts while using touch screen virtual manipulative mathematics apps. *Computers in Human Behavior*, *64*, 814–828. https://doi.org/10.1016/j.chb.2016.07.029

Webb, M.E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschêne, M., & Zagami, J. (2020). Machine learning for human learners: Opportunities, issues, tensions and threats. *Educational Technology Research and Development.* https://doi.org/10.1007/s11423-020-09858-2

Wei, H., Bos, R., & Drijvers, P. (2022). Embodied approaches to functional thinking using digital technology: A bibliometrics-guided review. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)*. ERME. https://hal.archives-ouvertes.fr/hal-03748999

Whitaker, D., & Jacobbe, T. (2017). Students' understanding of bar graphs and histograms: Results from the LOCUS assessments. *Journal of Statistics Education*, *25*(2), 90–102. https://doi.org/10.1080/10691898.2017.1321974

Whitaker, D., Foti, S., & Jacobbe, T. (2015). The levels of conceptual understanding in statistics (LOCUS) project: Results of the pilot study. *Numeracy*, *8*(2), 3. https://doi.org/10.5038/1936-4660.8.2.3

Wijers, M., & De Haan, D. (2020). Mathematics in teams—Developing thinking skills in mathematics education. In M. Van den Heuvel-Panhuizen (Ed.), *National Reflections on the Netherlands Didactics of Mathematics, ICME-13 Monographs*. Springer. https://doi.org/10.1007/978-3-030-33824-4_2

Wijnker, W., Smeets, I., Burger, P., & Willems, S. (2022). *Debunking strategies for misleading bar charts*. Preprint. https://doi.org/10.31235/osf.io/wm6te

Wild (2006). The concept of distribution. *Statistics Education Research Journal*, *5*(2), 10–26, http://iase-web.org/documents/SERJ/SERJ5(2)_Wild.pdf?1402525006

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science*, *5*(5), 249–252. https://doi.org/10.1111/j.1467-9280.1994.tb00621.x

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165–181. https://doi.org/10.1080/10508406.2019.1705678

Wong, C. (2009). *Kid's survey network: Teaching data literacy with multiplayer online games* [Master's thesis]. https://dspace.mit.edu/handle/1721.1/53130

WRI. (2020). *Mathematica* [computer software]. Wolfram Research.

Yilmaz, E., & Gompel, M. (n.d.) Automatic transcription of Dutch speech recordings. https://webservices.cls.ru.nl/

Yin, R.K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), 321–332. https://doi.org/10.1177/1356389013497081

Yuan, L., Haroz, S., & Franconeri, S. (2019). Perceptual proxies for extracting averages in data visualizations. *Psychonomic Bulletin & Review*, *26*(2), 669–676. https://doi.org/10.3758/s13423-018-1525-7

Yun, H. J., & Yoo, Y. J. (2011). Students' misconceptions and obstacles in generating histograms using variables collected during statistical investigation. In *The 8th Korean Women in Mathematical Sciences International Conference.* Springer.

Yun, H. J., Ko, E., & Yoo, Y. J. (2016). Students' misconceptions and mistakes related to measurement in statistical investigation and graphical representation of data. In D. Ben-Zvi, & K. Makar (Eds.), *The teaching and learning of statistics* (pp. 119-120). Springer. https://doi.org/10.1007/978-3-319-23470-0_14

Zaidan, A., Ismail, Z., Yusof, Y. M., & Kashefi, H. (2012). Misconceptions in descriptive statistics among postgraduates in social sciences. *Procedia-Social and Behavioral Sciences*, *46*, 3535–3540. https://doi.org/10.1016/j.sbspro.2012.06.100

Summary

Statistical literacy is an important learning goal for citizens. The studies in this dissertation focus on a specific part of it—graph literacy—which includes being able to comprehend and interpret graphs of statistical data. Many secondary school students are not well prepared to draw justified conclusions from statistical data in graphs. For example, in 2022, only 42 percent of Dutch Grade 11 students correctly selected a graph from which they could draw justified conclusions (Cito, 2022). Such problems occur even with seemingly simple graphical representations of data, such as histograms. As histograms are omnipresent in research, society and education, they are important for learning about key concepts such as probability distributions. Therefore, our main research question is: How can pre-university track students in Grades 10–12 be supported in understanding histograms?

We expected that a review of the literature and a small-scale eye-tracking study would be sufficient input for a larger design study (Bakker, 2018). However, the topic of our research turned out to be much tougher than initially expected. As histograms are used in numerous disciplines it was impossible to summarize all that is known about them for education. Moreover, few interventions in statistics education had been reported at the start of our research, which, in addition, were not very successful. Hence, the literature provided little basis for the design of a new intervention. More research was needed before a new approach to teaching histograms could be designed. The eye-tracking studies not only examined in more detail how students interpreted histograms but also how these interpretations changed after solving dotplots.

In Chapter 1, we elaborated on the important role of graphs in statistical literacy. As many people tend to misinterpret histograms, an introduction to histograms was also provided. Furthermore, some reflections on the position of histograms in the school curriculum as well as a motive for doing this research as a teacher were given. Figure 1 shows an overview of the studies in this dissertation. We started with a literature review (Chapter 2). We conducted several eye-tracking studies, in which we qualitatively (Chapter 3) and quantitatively (Chapters 4 and 5) analyzed students' gaze data. We finished with a design study (Chapter 6) for which we developed, empirically tested and evaluated our conjectured learning trajectory.

Figure 1 Overview of the studies in this dissertation
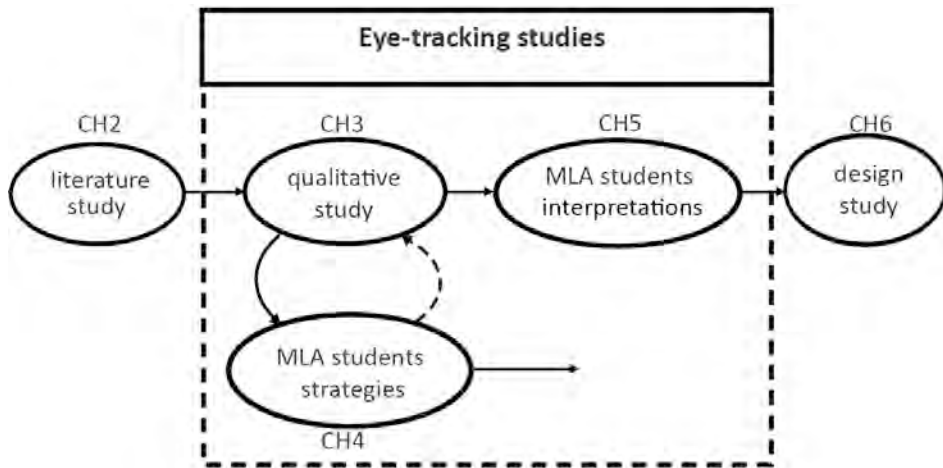
Review of literature on interpreting and constructing histograms

As an overview of the most common misinterpretations of histograms was lacking, in Chapter 2 we reviewed 86 publications on people's difficulties with histograms. Given the persistence of these misinterpretations, there is a need to reflect on what conceptual difficulties may lie at their basis through the first research question:

> *RQ1: What are the conceptual difficulties that become manifest in the common misinterpretations people have when constructing or interpreting histograms?*

The most common conceptual difficulties could be grouped into three categories labeled data, distribution, and miscellaneous. The first two each relate to a key concept in statistics: *data* and *distribution*. Difficulties that relate to the key concept of data are, for example, identifying the number of statistical variables and the measurement level of their attributes. Distribution-related difficulties include estimating or comparing centers (e.g., the mean) or comparing variation (variability). Although data-related misinterpretations are observed more often, research specifically addressing these misinterpretations is scarce. A third and more diverse category of misinterpretations is related to other conceptual difficulties. This includes having trouble linking the context to the histogram, not understanding the difference between a histogram of a sample and of a population and the influence of ICT. The analysis of the publications in our review also led to the identification of a network of statistical concepts specifically relevant to interpreting histograms (Chapter 2, Figure 2.2).

Students' strategies for statistical graphs tasks: an eye-tracking study

The review results allowed for more broadly addressing students' conceptual difficulties that become manifest in most common misinterpretations rather than focusing on a specific misinterpretation. Misinterpretations related to the statistical key concepts data and distribution can be observed when students confuse histograms with look-alikes, including case-value plots. As many studies draw conclusions from students' final answers (e.g., delMas et al., 2007; Whitaker & Jacobbe, 2017), little was known about students' strategies for reaching these answers, including their micro-level thinking processes. Therefore, it was unclear how to intervene effectively. The persistence of students' misinterpretations also called for a closer inspection of students' conceptual difficulties. Hence, in the second study (Chapter 3), we decided to figure out on a more detailed level what students' difficulties with histograms were through a larger *eye-tracking study*, as we thought that students' gaze patterns could provide insight into their approaches. In this second study, we posed the research question:

> RQ2: How and how well do Grades 10–12 pre-university track students estimate and compare arithmetic means of histograms and case-value plots?

Therefore, in Chapter 3, we observed students' actions by tracking their gazes while they were solving graph tasks, in particular, estimating and comparing arithmetic means of histograms and their look-alikes, case-value plots. By observing these actions, it becomes clear how students use their conceptual knowledge of the data in histograms, hence what strategies they employ. In this eye-tracking study, we investigated Grades 10–12 pre-university track (VWO) students' strategies ($N = 50$) when interpreting graphs. We recorded students' gazes while they solved 12 graph tasks and interviewed them right after. Students' gaze data were combined with verbal data from this cued recall to connect specific gaze patterns—the perceptual forms of gazes—to interpretation strategies.

In a qualitative analysis of students' scanpath patterns, we found five strategies. Two hypothesized most-common strategies for single graph tasks for *estimating* the mean as found in our pilot study (Boels et al., 2018) were confirmed: a typical case-value plot and a histogram strategy, the latter indicating that the student interprets the graph at hand as if it is a histogram (Figure 2). A vertical gaze pattern reflected this histogram strategy. A horizontal pattern was connected to a case-value plot strategy. In addition, a third, new, count-and-compute strategy was found that was only correct for case-value plots. Two more strategies were found for *comparing* case-value

plots and histograms—hence, for double graph items: a distribution-informed histogram strategy and a distribution-informed case-value plot strategy (Figure 3). Distribution informed means using specific features of the graph such as the same symmetry and positions of the bar, thus, the same mean, or similar shape but moved to the right, thus, higher mean; in short, they used 'shape' and 'shift' (cf. Frischemeier & Biehler, 2016).

Figure 2 Examples of the perceptual form of gaze patterns on single histogram tasks (estimating means) with a correct strategy for histograms in the top left and a correct strategy for case-value plots in the bottom right
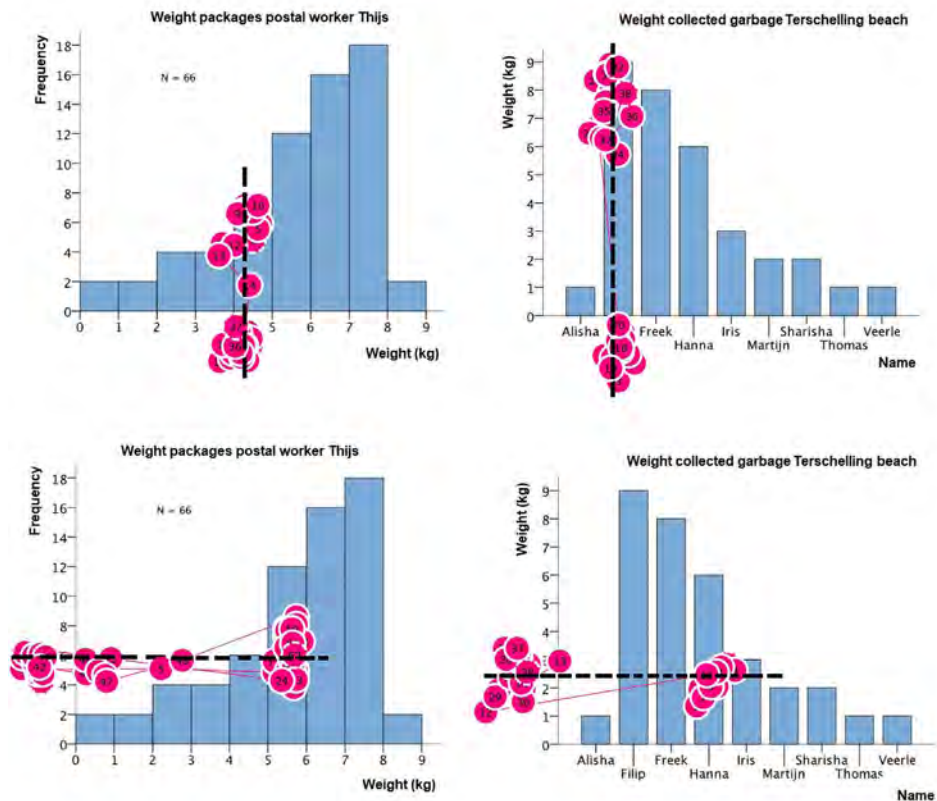

Note. The circles are places where a student looked longer. The dotted lines indicate the perceptual form of the scanpath patterns relevant to the strategy used.

The percentages of correct strategies varied between on average 43% for *single* histograms and 100% for case-value plots; the latter being distributed between a case-value plot strategy (71%) and a count-and-compute strategy (29%). These findings were in line with results from a pilot study (Boels et al., 2018) and a study with teachers rather than students (Boels et al., 2019b). The percentages of correct strategies varied between on average 50% for *double*

histograms and 90% for double case-value plots; the latter being mostly a case-value plot strategy (87% points). To our surprise, in on average 9% of the double case-value plots tasks, students used a distribution-informed *histogram* strategy that resulted in an incorrect answer, for example, by using the symmetry of the graphs. Furthermore, some students ignored bars with frequency or measured value zero even though they looked at them (cf. delMas & Liu, 2005).

Figure 3 A correct distribution-informed strategy for comparing the means of two histograms in Item09 using similar shape, shifted to the right and for comparing means of two case-value plots in Item07 using shape and number of bars
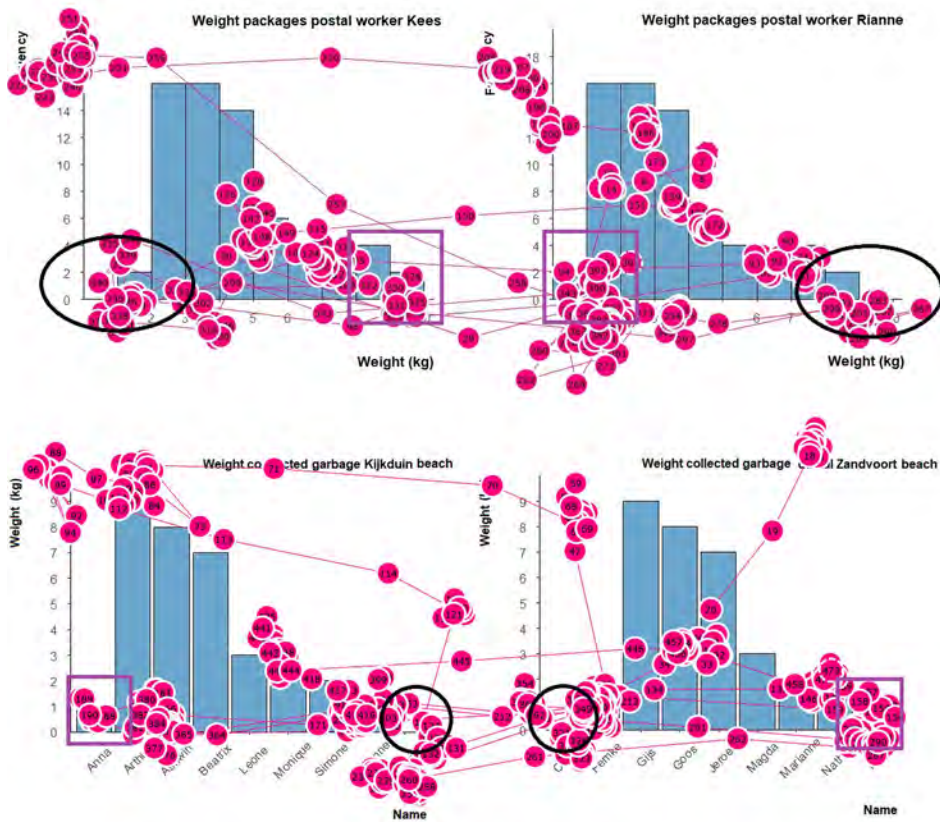
Note. Students specifically compared the position of the 'zero' bars (black ovals) and other bars on similar positions (e.g., purple squares). Correct answer top: Kees; bottom: same.

**Automatic, gaze-based identification of student strategies**

The patterns we found in students' gaze data for single histograms stimulated us to explore, in Chapter 4, whether automatic recognition of students' strategies through a machine learning analysis might be possible. Identification of student strategies is a prerequisite for targeted intelligent feedback, for example, during online learning. However, it was still unclear how to automate real-time identification of task-specific solution strategies based on students' gazes on histograms. The study in Chapter 4 is a first step in this automation process.

> *RQ3: How can gaze data be used to automatically identify students' task-specific strategies on single histograms?*

We used a software tool (Mathematica Classify Function) which automatically prepared the gaze data and fed these into its implementation of a supervised machine learning algorithm (MLA; random forest). This MLA was able to identify whether students used a histogram interpretation strategy or another strategy when estimating the mean from a single histogram. This other strategy most often was a strategy that would have been correct if the graph had been a case-value plot. The MLA performed acceptably (Figure 4), and accuracies varied from around chance level (38%) to well above (88%) depending on the validation procedure. Values above 70% are considered good as these are well above chance level. One disadvantage of the MLA is that it does not explain how it reached its decision for an individual student and we, therefore, consider it a black box. The results of the MLA provided a baseline for the transparent, interpretable mathematical model (IMM) we constructed. This IMM was theoretically meaningful and performed well with accuracies between 62% and 84%, acceptable sensitivity, and quite good specificity (Figure 4). We also succeeded in training our MLA when we used students' gaze data from one item and had the MLA identify strategies for all other items. These results indicate that students' strategies can be derived from their gaze data.

In the future, the results of such an automated strategy identification might be made available to teachers. Our method allows for the design of immediate, personalized feedback during online learning, homework, or massive open online courses (MOOCs)through measuring gazes with, for example, a webcam.

Figure 4 Results of the MLA and the IMM
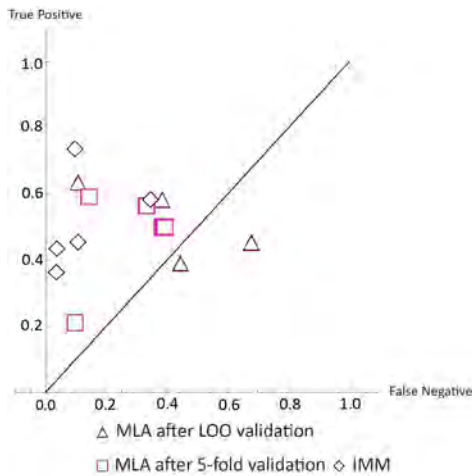
Note. Ideally, points should be concentrated in the upper left corner of the graph and close together for all items. The MLA provided a baseline for the IMM. Although the IMM worked well, the plot implies room for improvement.

Assessing students' interpretations of histograms before and after solving dotplot tasks

The previous studies revealed students' solution strategies when solving histogram tasks in more detail. A local instruction theory in statistics education suggests that solving dotplots can support interpreting histograms (e.g., Bakker & Gravemeijer, 2004; Garfield, 2002; Garfield & Ben-Zvi, 2008b) as dotplots can draw students' attention to the variable being presented along the horizontal axis in both graphs. We wondered whether interpreting dotplots would influence students' strategies on histogram tasks. Therefore, while collecting eye-tracking data, we included six dotplot tasks immediately after the first twelve tasks followed by three histogram tasks. In Chapter 5 we explored:

> RQ4: In what way do Grades 10–12 pre-university track students' histogram interpretations change after solving dotplot items?

Students' gaze data on four histogram items were used as inputs for an MLA (random forest). Our MLA can quite accurately classify whether students' gaze data belong to an item solved before or after solving the dotplot items. The results indicate that there is a change in students' gaze patterns. Moreover, we found that the direction (e.g., almost vertical) and magnitude (length) of saccades (fast transitions between positions where students looked) were different on the before and after items. For example, gazes contained more

vertical and less horizontal saccades on the histogram tasks after solving the dotplot tasks. These changes could indicate a change in strategies.

We found three indications that students' histogram interpretations changed after solving the dotplot items: a change in students' gaze patterns (from the MLA result), an improvement in students' estimations of the arithmetic means for single histogram items, and a shift in students' reported strategies for solving histogram items. However, the number of correct answers did not change significantly. For single histograms this number was sensitive to the researchers' choice of an answer range for correct answers. In addition, evidence that the change in gaze behavior indicates learning, that in turn can *only* be attributed to solving the dotplot items, is weak. We consider as a most likely explanation for the mixed results that the action of solving dotplot items creates readiness for learning (Church & Goldin-Meadow, 1986). Reflection on their strategy—induced by the stimulated recall with the adult interviewer—then made students realize their misinterpretation of, for example, the frequencies as the measured values. This study suggests that activities with dotplots may support students in understanding histograms. Moreover, Konold (2007) already noted that dotplots can support students' understanding of histograms through *actions* such as "separate, order, […] stack" and "fuse" the dots (p. 282). Fusing dots results in a bar that contains all these dots. Taken together, the results could point at a learning effect of solving the dotplots tasks—depending on how learning is defined.

Embodied design of a learning trajectory

The literature research (Chapter 2) also made clear that existing interventions were not sufficiently successful in teaching students to correctly interpret histograms. Students' solution strategies (Chapter 3) showed that many of these Dutch students lacked understanding of how and where data are represented in histograms. Interpreting dotplots may assist students' understanding of histogram (Chapter 5) as they draw students' attention to the axis along which the data are represented in both histograms and dotplots (being the horizontal axis). However, it was still unclear how an intervention could be designed that would support students' learning of statistical key concepts through interpreting dotplots and histograms.

Given the persistence of students' difficulties with interpreting histograms, we assumed that students' education might have lacked an embodied grounding of how histograms are constructed as well as sufficient attention to how these artifacts become tools in statistical reasoning. In embodied designs, students' actions play an important role, such as the actions described in the previous paragraph (Konold, 2007). Therefore, using an embodied instrumentation approach (Drijvers, 2019) as a theoretical lens,

we designed a learning trajectory (Chapter 6) using findings and insights from previous studies. This design study is a first cycle of a design research project on how to teach some of the most important aspects of the key concepts of statistics through teaching histograms. The research question for this study was:

> *RQ5: What sequence of tasks designed from an embodied instrumentation perspective can support students' understanding of histograms and the underlying key concepts?*

Figure 5 Saccades of magnitude 200 pixels or more of all participants on Item11 and Item21 (double-histograms, top) as well as Item02 and Item20 (single-histogram, bottom)
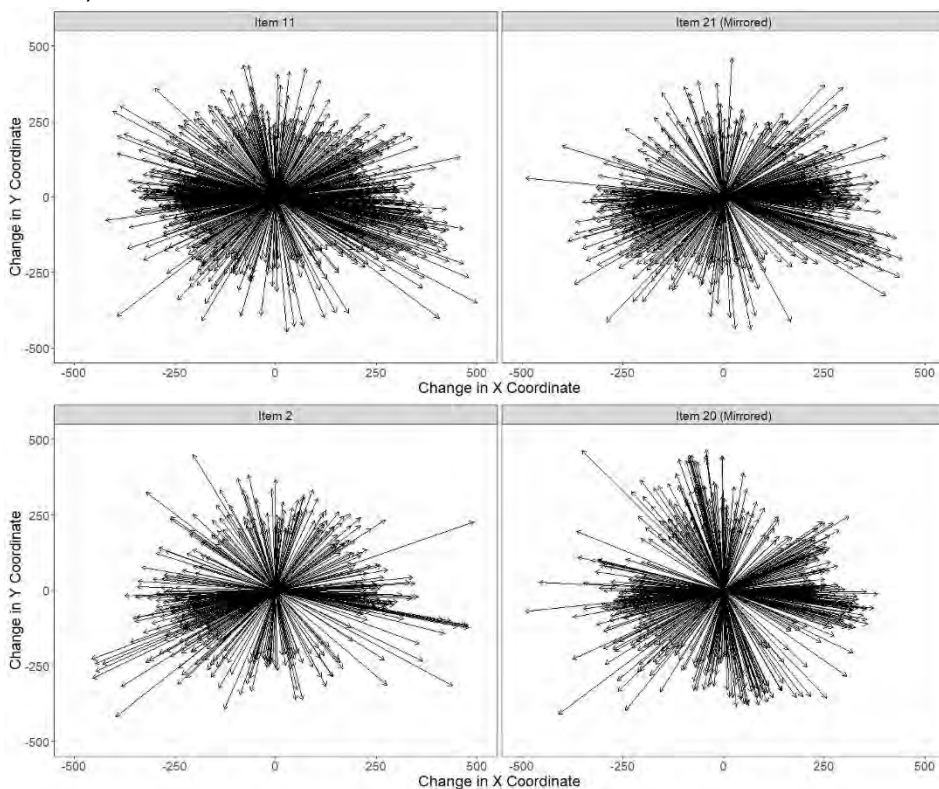

Note. Notice the difference in whiteness and blackness of students' saccade directions between the *before* items (left column) and *after* items (right column). Differences between rows are most likely mainly due to differences in tasks and, therefore, most probably irrelevant for our research question.

Our conjectured learning trajectory consists of five stages: (1) learning initiation—experiencing a lack of understanding, (2) reinventing the role of the horizontal scale in univariate graphs, (3) reinventing the role of the vertical

scale in histograms, (4) reinventing arithmetic means in histograms, and (5) confirming learning—transfer to other contexts and environments.

Our multiple case study with five students (Grades 10–12) suggested that most conjectures of the learning trajectory were met but transfer can be improved. Contributing to further theorization of embodied instrumentation, we discussed heuristics for the design process. In addition, we showed how more complicated artifacts (e.g., histograms) can be reinvented from actions with simpler ones (e.g., positioning dots on a scale, dotplots).
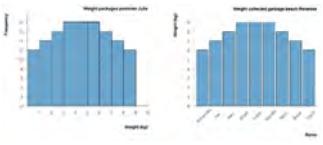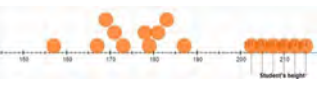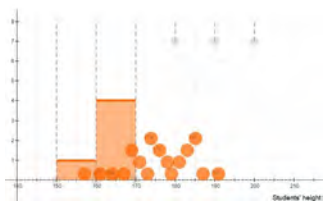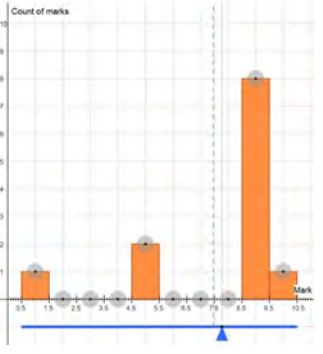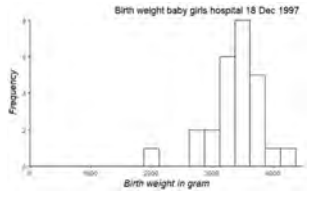
Figure 6 Example of attempts by students when reinventing that the bars' heights in histograms is equal to the number of measured values in each bar, incorrect (left) and correct (right)
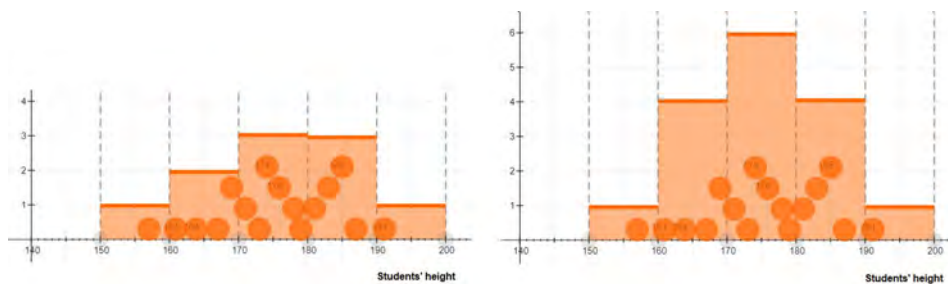
Our hypothetical learning trajectory (HLT, Table 1) was used to support students in understanding some of the most important aspects of the concepts of data and distribution presented in histograms. For data, these aspects were where and how the data are depicted in a histogram, including that the vertical axis does *not* represent measured values. For distribution, these aspects were how the mean is influenced by the spread in and shape of the histogram as a precursor for understanding variation in a histogram. In secondary schools, the focus is often on calculating measures of center and plotting histograms (e.g., Burrill, 2020). In our design, the focus was on the key concepts that we wanted students to grasp instead of only teaching them how to construct histogram.

What is also new in the design is that we added tasks that required instrumented actions (Shvarts et al., 2021). Instrumented actions can be understood as actions influenced by digital technology, hence by the specific way the digital tasks were designed. An example is the unit height for bars in histograms not being equal to the size of dots, which made students understand through a productive struggle (e.g., Kapur, 2014; Roth, 2019) that this height is equal to the number of measurements (in histograms with equal bin widths only).

Table 1 Overview of the hypothetical learning trajectory

Step	Example task	Activities	Example conjecture
1 Learning initiation: Experiencing lack of understanding		Compare means and variation in a histogram (left) and case-value plot (right).	H1b: By experiencing initial confusion or misunderstanding, students' intentionality and motivation for upcoming tasks are established.
2 Reinventing the role of the horizontal scale in univariate graphs		Slide dots to their position on a scale.	H2b: By horizontally moving dots to their correct position on a horizontal scale, students notice the *position* of a dot depicts the measured value.
3 Reinventing the role of the vertical scale in histograms		Build a histogram overlay from a dotplot.	H3b: By moving the (orange) sliders up, students notice the height of the bars is related to the number of cases in a bar when class intervals are equal.
4 Reinventing arithmetic means in histograms		Establish relation between data and mean; discover influence of outliers, gaps, distribution, on the mean.	H5a: By finding the balancing point of the graph, students perceive the mean can be seen as the point where the graph is "in balance."
5 Confirming learning: transfer to other situations		Construct and interpret histograms on paper. Sort histograms and look-alikes.	H18b: By drawing a histogram on paper from a frequency table, transfer to another environment (paper) is established.

Comparing students' performances with the conjectures from the anticipated HLT, our case study suggests most conjectures were met. Students experienced misunderstanding in the first step, had no trouble reinventing the role of the horizontal scale, struggled but reinvented the role of the vertical scale in histograms, seemed to have an easy task estimating the balance point of a histogram, and stated that it is the arithmetic mean. The final tasks showed that students were often able to transfer the acquired knowledge to paper, hence, to a different environment. Students' gestures indicated using actions from previous tasks to solve follow-up tasks. Taken together, the results suggest that embodied experiences followed by reflection contributed to overcoming some well-known misinterpretations. However, some improvements are suggested for future designs, such as also adding transfer tasks after the steps which are dedicated to horizontal and vertical actions of the HLT. To further develop students' notions of distribution and variability, the artifact "area" may need to be included in the design, and the artifact "interval" may need to be reinvented by students.

Conclusions and discussion

An important component of statistical literacy is graph literacy. The histogram can be regarded as a spider in a web of knowledge. For example, understanding histograms is a good preparation for key concepts such as probability distribution and density in probability theory. The aim of this research was to contribute to an empirically grounded theory on how to teach histograms as a means to contribute to students' statistical literacy.

We answered the question of how pre-university track students in Grades 10–12 can be supported in understanding histograms. The main answer is a hypothetical learning trajectory (HLT) (Simon, 2020) that intends to develop students' notions of some key aspects of first data and then distributions in graphs of univariate data. This HLT was based on an extensive review of literature and methodologically innovative eye-tracking studies. In addition, it was designed from an embodied instrumentation perspective. Our HLT is a step toward a *domain-specific instructional framework* on how to help students correctly interpret graphs of univariate data, including histograms, dot-, stem-and-leaf, and boxplots, hatplots (Konold, 2007), frequency polygons, and histodots (Chapter 2). For future designs, it could be investigated whether an Intelligent Tutoring System could be incorporated for automatic feedback based on scanpath patterns on only the graph area of histograms. Such a system would require webcams that can do eye-tracking.

A scientific contribution of our work is that we showed how theoretical (Chapter 2) and empirical (Chapters 3, 5) insights about students' difficulties with statistical concepts can be incorporated into a sequence of tasks designed

from an embodied instrumentation perspective (Chapter 6). This is the first design in statistics education using an embodied instrumentation approach. In addition, we developed, tested, and evaluated guidelines for an embodied *instrumentation* design.

A methodological contribution of our work is that we introduced and applied several new research tools in statistics education research: eye-tracking, machine learning algorithms (MLAs), and an Interpretable Mathematical Model (IMM) (Chapter 4). These tools can be used for investigating details of students' strategies and informing designs (eye-tracking, MLA) and for designing intelligent tutoring systems that provide feedback (MLA and IMM). In addition, we showed that the perceptual forms of scanpaths on the graph area only of statistical graphs can reveal students' strategies when comparing and estimating means from these graphs.

A methodological limitation of our work is the geographical selection bias that seems to exist in the review study (Chapter 2) and the number of students in the eye-tracking study (N = 50, Chapters 3–5) and multiple-case study (N = 5; Chapter 6). Still, the approach is open to further scaling up and the results seem independent of these specific settings.

An implication for research is that eye-tracking can potentially shed new light on tenacious didactical problems in mathematics teaching, as students' scanpaths can reveal correct reasoning even when answers are incorrect. In addition, gaze data combined with an MLA and IMM could be a powerful tool for validating qualitative research findings.

An implication for educational practice is that histograms may play a central role in learning statistical key concepts such as data, distribution, variability or variation, and central tendency, and that more attention is needed to the key concept of data. In addition, more emphasis is needed on interpreting histograms and less on technical skills such as *how* to draw them and how to *compute* means of data presented in graphs. For initial learning, an embodied instrumentation approach seems a fruitful route for developing students' graphical literacy as part of statistical literacy. With this in mind, we call on designers to use our guidelines for embodied instrumentation designs for tenacious didactical problems in mathematics teaching and to question for all aspects of the artifacts (axes, scale, area) whether the mathematical actions and 'thinking' should be done by the software or the student.
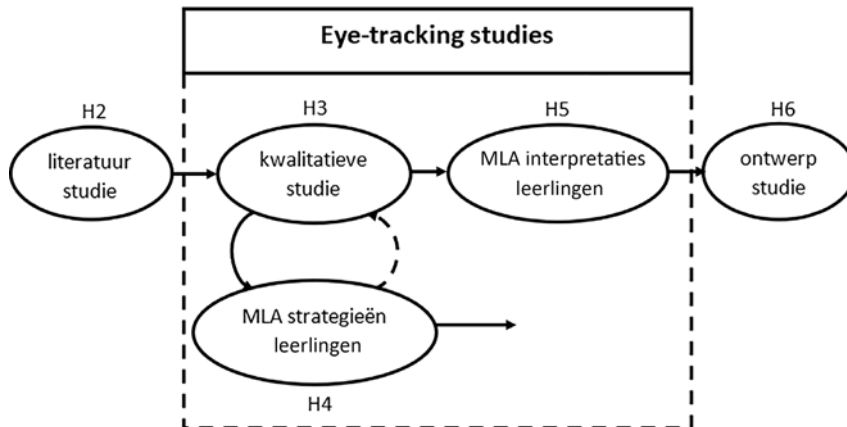
Samenvatting

Statistische gecijferdheid is een belangrijk leerdoel voor burgers om volwaardig mee te kunnen doen in de maatschappij. Dit onderzoek richt zich op een onderdeel hiervan: grafische gecijferdheid. Grafische gecijferdheid omvat het correct kunnen interpreteren van statistische data die zijn weergegeven in diagrammen. Veel leerlingen zijn niet goed voorbereid op het trekken van verantwoorde conclusies uit statistische data in diagrammen. In 2022 kon slechts 42 procent van de Nederlandse 5-havo-leerlingen met wiskunde A op hun eindexamen een correct diagram selecteren waaruit zij onderbouwde conclusies konden trekken (Cito, 2022). Dergelijke problemen doen zich ook voor bij ogenschijnlijk eenvoudige grafische weergaven van data zoals histogrammen. Histogrammen worden veel gebruikt in onderzoek, maatschappij en onderwijs en zijn daarom belangrijk om te leren. Daarnaast zijn histogrammen belangrijk om kernconcepten—zoals kansverdelingen—te leren. De hoofdvraag van dit onderzoek is daarom: *Hoe kunnen leerlingen in 4–6 vwo ondersteund worden bij het begrijpen van histogrammen?*

Aanvankelijk hadden we verwacht dat een literatuurstudie en een kleinschalige studie van oogbewegingen voldoende informatie zou opleveren voor een grotere ontwerpstudie (Bakker, 2018). Het onderwerp van ons onderzoek bleek echter veel lastiger dan aanvankelijk gedacht. Ten eerste worden histogrammen in tal van disciplines gebruikt. Ten tweede waren er weinig interventies in het statistiekonderwijs gerapporteerd, die bovendien niet erg succesvol waren. De literatuur bood dus weinig basis voor het ontwerp van een nieuwe interventie. Er was daarom meer onderzoek nodig voordat we een nieuwe aanpak voor het onderwijzen van histogrammen konden ontwerpen. In de oogbewegingsstudies is zodoende niet alleen nader onderzocht hoe leerlingen histogrammen interpreteerden maar ook hoe deze interpretaties veranderden na het oplossen van stippendiagramtaken.

Hoofdstuk 1 bespreekt de belangrijke rol van diagrammen bij statistische gecijferdheid. Aangezien veel mensen geneigd zijn histogrammen verkeerd te interpreteren, is ook een inleiding op histogrammen gegeven. Verder beschouwen we kort de plaats van histogrammen in het schoolcurriculum en motieven om als docent dit promotieonderzoek te doen. Figuur 1 biedt een overzicht van de studies in dit proefschrift. De eerste studie is een literatuurstudie (hoofdstuk 2). We voerden verschillende oogbewegingsstudies uit waarin we de oogbewegingsdata van leerlingen kwalitatief (hoofdstuk 3) en kwantitatief (hoofdstukken 4 en 5) analyseerden.

We eindigden met een ontwerpstudie (hoofdstuk 6) waarin we een hypothetisch leertraject ontwikkelden, empirisch testten en evalueerden.

Figuur 1 Overzicht van de studies in dit proefschrift

Eye-tracking studies

H2 — literatuur studie → H3 — kwalitatieve studie → H5 — MLA interpretaties leerlingen → H6 — ontwerp studie

H4 — MLA strategieën leerlingen

Opmerking. MLA betekent machine learning algoritme

Literatuuronderzoek naar het interpreteren en construeren van histogrammen

Aangezien een overzicht van de meest voorkomende misinterpretaties van histogrammen ontbrak, hebben we 86 publicaties bekeken waarin moeilijkheden van mensen met het interpreteren van histogrammen (hoofdstuk 2) voorkwamen. De hardnekkigheid van deze misinterpretaties maakte het nodig om na te gaan welke conceptuele moeilijkheden eraan ten grondslag liggen:

> *V1: Wat zijn de conceptuele moeilijkheden die tot uiting komen in veelvoorkomende misinterpretaties die mensen hebben bij het construeren of interpreteren van histogrammen?*

De meest voorkomende conceptuele moeilijkheden kunnen worden gegroepeerd in drie categorieën: data, verdeling en overige. De eerste twee zijn kernconcepten in de statistiek. Moeilijkheden die gerelateerd zijn aan het kernconcept data zijn bijvoorbeeld het bepalen van het aantal statistische variabelen en het meetniveau van de bijbehorende metingen. Moeilijkheden die gerelateerd zijn aan het kernconcept verdeling zijn bijvoorbeeld het bepalen of vergelijken van een centrummaat—zoals het gemiddelde—of het vergelijken van variabiliteit. Een derde en meer diverse categorie moeilijkheden houdt verband met andere conceptuele moeilijkheden zoals

problemen om de context aan een histogram te koppelen, het niet begrijpen van het verschil tussen een histogram van een steekproef en van een populatie, en de invloed van ICT. De analyse van de publicaties in ons overzicht leidde ook tot de identificatie van een netwerk van statistische concepten die specifiek relevant zijn voor de interpretatie van histogrammen (hoofdstuk 2, figuur 2).

Leerlingstrategieën voor statistische diagramtaken: een oogbewegingsstudie

Het literatuuronderzoek maakte een bredere aanpak mogelijk van de conceptuele problemen van leerlingen die tot uiting komen in de meest voorkomende misinterpretaties. Misinterpretaties gerelateerd aan de kernconcepten data en verdeling kunnen worden geobserveerd wanneer leerlingen histogrammen verwarren met hun evenbeelden, inclusief staafdiagrammen. Veel studies trekken conclusies uit de uiteindelijke antwoorden van studenten (bv. delMas et al., 2007; Whitaker & Jacobbe, 2017). Hierdoor was er weinig bekend over de strategieën van studenten om tot deze antwoorden te komen, inclusief hun denkprocessen op microniveau. Het was daarom onduidelijk hoe een effectieve interventie eruit kon zien. De hardnekkigheid van de misinterpretaties van leerlingen vroeg bovendien om een nadere inspectie van hun conceptuele moeilijkheden. In een tweede onderzoek is daarom op een gedetailleerder niveau uitgezocht wat deze moeilijkheden waren in een oogbewegingsonderzoek omdat we verwachtten dat de oogbewegingen van leerlingen inzicht konden geven in hun aanpak. In deze tweede studie stelden we de onderzoeksvraag:

> *V2: Hoe en hoe goed schatten en vergelijken bovenbouw vwo-leerlingen rekenkundige gemiddelden van histogrammen en casusstaafdiagrammen?*

Hiertoe observeerden we in hoofdstuk 3 de acties van leerlingen door hun oogbewegingen te volgen terwijl ze diagramtaken oplosten, in het bijzonder het schatten en vergelijken van rekenkundige gemiddelden van histogrammen en hun ogenschijnlijke evenbeelden, casusstaafdiagrammen. Door deze acties te observeren wordt duidelijk hoe leerlingen hun conceptuele kennis van data in histogrammen gebruiken en welke strategieën ze daarbij hanteren.

In deze oogbewegingsstudie onderzochten we de strategieën van leerlingen in 4–6 vwo (*N* = 50) bij het interpreteren van statistische diagrammen. We registreerden de oogbewegingen van leerlingen terwijl ze 12 diagramtaken oplosten en interviewden hen direct daarna. Daarbij lieten we

hen hun eigen oogbewegingen terugzien terwijl ze vertelden welke strategie ze hadden gebruikt. De oogbewegingsdata van de leerlingen werden gecombineerd met deze verbale data zodat specifieke kijkpatronen—de perceptuele vormen ervan—konden worden verbonden met interpretatiestrategieën.

Figuur 2 Voorbeelden van de kijkpatronen op enkelvoudige histogramtaken (schatten van gemiddelden) met een correcte strategie voor histogrammen linksboven en een correcte strategie voor casusstaafdiagrammen rechtsonder
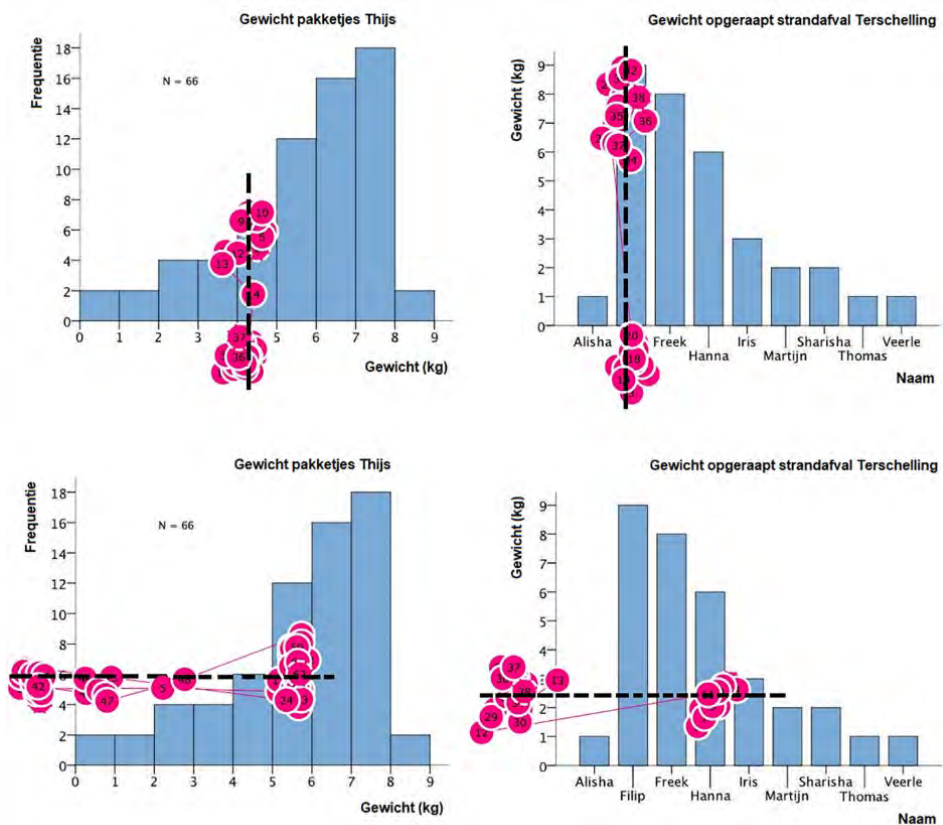
Opmerking. De cirkels zijn plekken waar een leerling langer keek. De stippellijnen geven de vorm van de voor de gebruikte strategie relevante kijkpatronen aan.

In een kwalitatieve analyse van de kijkpatronen van studenten vonden we vijf strategieën. Vooraf vermoedden we dat er twee meest voorkomende strategieën zouden zijn waarbij leerlingen het gemiddelde schatten in enkelvoudige diagramtaken, zoals we eerder hadden gevonden in onze pilotstudie (Boels et al., 2018): een casusstaafdiagram- en een

histogramstrategie. Dit vermoeden werd bevestigd. Een histogramstrategie houdt in dat de leerling het diagram interpreteert als een histogram (figuur 2). Een verticaal kijkpatroon weerspiegelde deze histogramstrategie. Een horizontaal kijkpatroon hing samen met een casusstaafdiagramstrategie. Daarnaast werd een derde nieuwe tel-en-berekenstrategie gevonden die alleen correct was voor casusstaafdiagrammen. Voor het vergelijken van casusstaafdiagrammen en histogrammen—dus voor dubbele diagramtaken— werden nog twee strategieën gevonden: een verdelingsgeïnformeerde histogramstrategie en een verdelingsgeïnformeerde casusstaafdiagramstrategie (figuur 3). Verdelingsgeïnformeerd betekent dat specifieke kenmerken van het diagram werden gebruikt zoals: beide diagrammen hebben dezelfde symmetrie en posities van de staven dus is het gemiddelde hetzelfde, of de diagrammen hebben een vergelijkbare vorm maar de staven zijn naar rechts verschoven dus het gemiddelde is hoger. Ze gebruikten dus "vorm" en "verschuiving" (cf. Frischemeier & Biehler, 2016).

Het percentage correcte strategieën varieerde tussen gemiddeld 43% voor enkelvoudige histogrammen en 100% voor casusstaafdiagrammen; dit laatste percentage was verdeeld over een casusstaafdiagramstrategie (71%) en een tel-en-berekenstrategie (29%). Deze resultaten waren vergelijkbaar met resultaten uit een pilotstudie met studenten (Boels et al., 2018) en een studie met docenten (Boels et al., 2019b). Voor dubbele diagramtaken varieerde het percentage correcte strategieën tussen gemiddeld 50% voor dubbele-histogramtaken en 90% voor dubbele-casusstaafdiagrammen; de laatste was vooral een casusstaafdiagramstrategie (87% punten). Tot onze verrassing gebruikten leerlingen in gemiddeld 9% van de taken met dubbele casusstaafdiagrammen een verdelingsgeïnformeerde histogramstrategie die resulteerde in een onjuist antwoord, bijvoorbeeld door gebruik te maken van symmetrie van de diagrammen. Verder negeerden sommige leerlingen staven met frequentie of meetwaarde nul, ondanks dat ze er wel naar keken (cf. delMas & Liu, 2005).

Figuur 3 Een correcte verdelingsgeïnformeerde strategie voor het vergelijken van de gemiddelden van twee histogrammen in taak09 met behulp van vergelijkbare vorm, verschoven naar rechts en voor het vergelijken van de gemiddelden van twee casusstaafdiagrammen in Taak07 met behulp van vorm en aantal staven
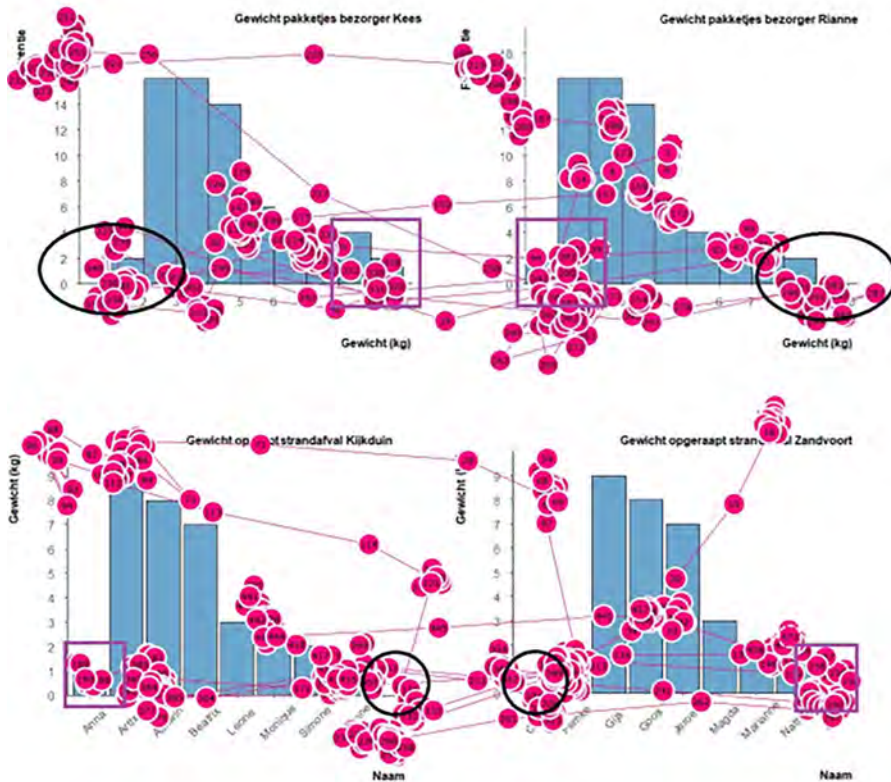
Opmerking. Leerlingen vergeleken specifiek de positie van de 'nul'-staven (zwarte ovalen) en andere staven op vergelijkbare posities (bijvoorbeeld paarse vierkanten). Correcte antwoorden op de vraag waar het gemiddelde gewicht hoger is: boven is dat Kees; onder is op beide stranden gemiddeld hetzelfde gewicht geraapt.

Automatische, op oogbewegingen gebaseerde identificatie van leerlingstrategieën

De patronen die we vonden in de oogbewegingsdata van leerlingen voor afzonderlijke histogrammen stimuleerden ons om te onderzoeken of automatische herkenning van strategieën van leerlingen mogelijk zou kunnen zijn door middel van een analyse met machine learning (hoofdstuk 4). Identificatie van leerlingenstrategieën is een voorwaarde voor gerichte intelligente feedback, bijvoorbeeld bij online leren. Het was echter nog

onduidelijk hoe live identificatie van taakspecifieke oplossingsstrategieën op basis van de oogbewegingen van studenten in het statistiekonderwijs geautomatiseerd kon worden. De studie in hoofdstuk 4 is een eerste stap in dit automatiseringsproces.

> *V3: Hoe kunnen oogbewegingsdata worden gebruikt om automatisch taakspecifieke strategieën van leerlingen te identificeren op enkelvoudige histogrammen?*

Wij gebruikten een gesuperviseerd machine learning algoritme (MLA; random forest) geïmplementeerd in een softwaretool (Mathematica Classify Function) met de oogbewegingsdata als input. Dit MLA kon vaststellen of leerlingen een histogram-interpretatiestrategie gebruikten of een andere strategie—meestal een strategie die correct zou zijn geweest als het diagram een casusstaafdiagram was—bij het schatten van het gemiddelde uit een enkel histogram. Het MLA had acceptabele prestaties (figuur 4) en de nauwkeurigheid varieerde tussen kansniveau (38%) en ruim daarboven (88%), afhankelijk van de gebruikte validatie procedure. Waarden boven 70% worden als goed beschouwd, aangezien deze ruim boven kansniveau liggen. Een nadeel van dit MLA is echter dat het niet uitlegt hoe het voor een individuele student tot een beslissing is gekomen en daarom beschouwen wij het als een black box. De resultaten van het MLA vormden voor ons een ijkpunt voor het transparante, interpreteerbare wiskundige model (IWM) dat wij construeerden. Dit IWM was theoretisch zinvol en had goede prestaties met nauwkeurigheden tussen 62% en 84%, acceptabele sensitiviteit en erg hoge specificiteit (Figuur 4). Wij slaagden er bovendien in om het MLA te trainen met de oogbewegingsdata van leerlingen op één taak om daarmee strategieën te identificeren op alle andere taken. Deze resultaten wijzen erop dat de strategieën van de leerlingen kunnen worden afgeleid uit de oogbewegingsdata.

In de toekomst zouden de resultaten van een dergelijke geautomatiseerde strategie-identificatie in aan docenten ter beschikking kunnen worden gesteld. Onze methode maakt het mogelijk om directe, gepersonaliseerde feedback te ontwerpen tijdens online leren, huiswerk of *massive open online courses* (MOOC's), door het meten van oogbewegingen met bijvoorbeeld een webcam.

Figuur 4 Resultaten van het MLA en het IWM
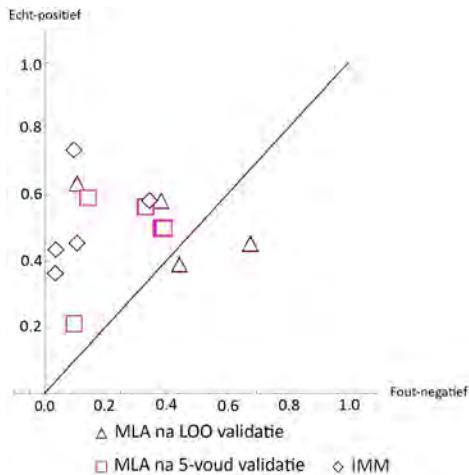
Opmerking. Idealiter zijn de punten geconcentreerd in de linkerbovenhoek van het diagram en liggen ze voor alle taken dicht bij elkaar. Hoewel het IWM goed werkte, laat het diagram zien dat er ruimte voor verbetering is.

Vergelijken van histograminterpretaties van leerlingen voor en na het oplossen van stippendiagramtaken

De voorgaande studies makende details van de oplossingsstrategieën van leerlingen bij histogramtaken zichtbaar. Een lokale instructietheorie in het statistiekonderwijs suggereert dat het oplossen van stippendiagramtaken het interpreteren van histogrammen kan bevorderen, (bv. Bakker & Gravemeijer, 2004; Garfield, 2002; Garfield & Ben-Zvi, 2008b) omdat stippendiagrammen de aandacht van leerlingen kunnen vestigen op de variabele die in beide diagrammen langs de horizontale as wordt gepresenteerd. Wij vroegen ons af of het interpreteren van stippendiagrammen de strategieën van leerlingen bij histogramtaken zouden beïnvloeden. Daarom hebben we tijdens het verzamelen van de oogbewegingsdata zes stippendiagramtaken opgenomen onmiddellijk na de eerste twaalf taken met histogrammen, gevolgd door opnieuw drie histogramtaken. In de studie in hoofdstuk 5 onderzochten we:

> *V4: Op welke manier veranderen de histograminterpretaties van leerlingen in 4–6 vwo na het oplossen van stippendiagramtaken?*

De kijkgegevens van leerlingen op vier histogramtaken werden gebruikt als input voor een MLA (*random forest*). Onze MLA kan vrij nauwkeurig classificeren of de oogbewegingsdata van leerlingen behoren tot een taak die is opgelost vóór of na de stippendiagramtaken. De resultaten wijzen erop dat

er een verandering is in het kijkpatroon van de leerlingen. Bovendien vonden wij dat de richting (bv. bijna verticaal) en de lengte van de saccades (snelle overgangen tussen posities waar leerlingen keken) verschillend waren bij de taken voor en na het oplossen van de stippendiagramtaken. Er waren bijvoorbeeld meer verticale en minder horizontale saccades in de oogbewegingen op de histogramtaken na het oplossen van de stippendiagramtaken. Deze veranderingen zouden kunnen wijzen op een verandering in strategieën.

Wij vonden drie aanwijzingen dat histograminterpretaties van de leerlingen veranderden na het oplossen van de stippendiagramtaken: een verandering in de kijkpatronen (afgeleid uit het MLA-resultaat), een verbetering in de schattingen van de rekenkundige gemiddelden voor taken met enkelvoudige histogrammen en een verschuiving in de gerapporteerde strategie voor het oplossen van histogramtaken. Het aantal correcte antwoorden veranderde echter niet significant. Voor enkelvoudige histogrammen was dit aantal gevoelig voor de keuze van de onderzoekers binnen welk bereik antwoorden correct werden gerekend. Daarnaast is het bewijs zwak voor de veronderstelling dat de verandering in het kijkgedrag duidt op leren, dat op zijn beurt uitsluitend zou kunnen worden toegeschreven aan het oplossen van de histogramtaken. We beschouwen de meest waarschijnlijke verklaring voor de gemengde resultaten dat de actie van het oplossen van stippendiagramtaken rijpheid voor leren creëert (Church & Goldin-Meadow, 1986). Reflectie op hun eigen strategie—opgeroepen door de gestimuleerde herinnering aan de gebruikte strategie door het interview met een volwassene—deed de leerlingen vervolgens beseffen dat zij bijvoorbeeld de frequenties verkeerd interpreteerden als zijnde de gemeten waarden. Deze studie suggereert dat activiteiten met stippendiagrammen leerlingen mogelijk kunnen ondersteunen bij het begrijpen van histogrammen. Bovendien merkte Konold (2007) al op dat stippendiagrammen het begrip van histogrammen door leerlingen kunnen ondersteunen door acties als "scheiden, ordenen, [...] stapelen" en "versmelten" van de stippen (p. 282). Het samensmelten van stippen resulteert in een staaf die al deze stippen bevat. Samengevat zouden de resultaten kunnen wijzen op een leereffect van het oplossen van de stippendiagramtaken—afhankelijk van hoe leren wordt gedefinieerd.

Figuur 5 Saccades van lengte 200 pixels of meer van alle deelnemers op taak11 en taak21 (dubbele histogrammen, boven) en taak02 en taak20 (enkel histogram, onder)
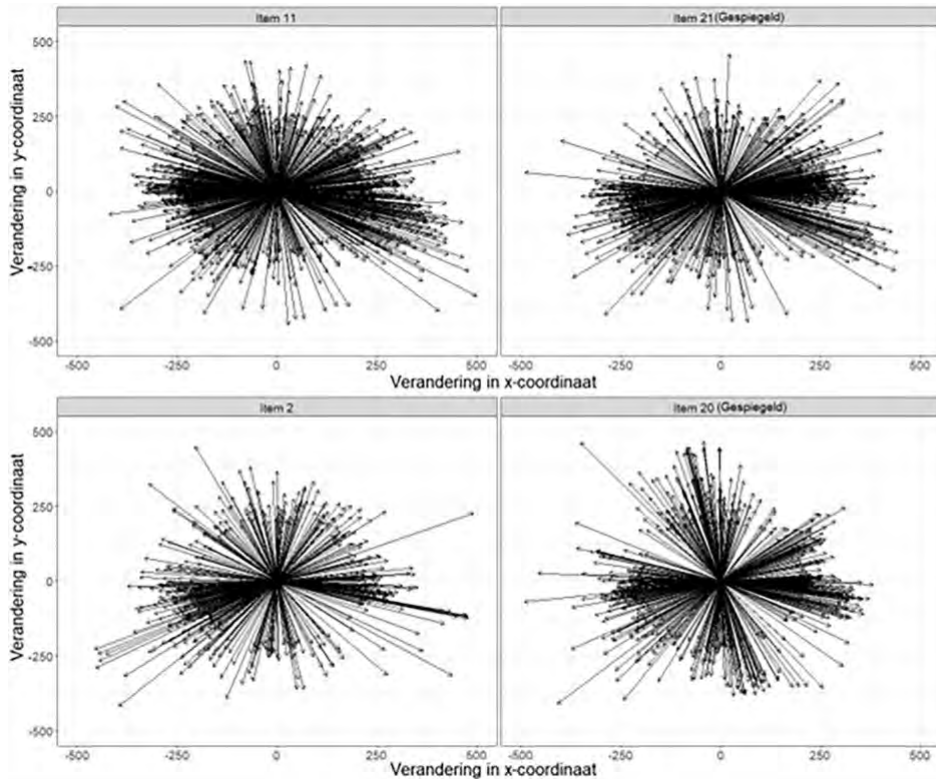
Opmerking. Let op het verschil in witheid en zwartheid van de richtingen van de saccades van studenten tussen de taken *ervoor* (linker kolom) en *erna* (rechter kolom). Verschillen tussen de bovenste en onderste rij zijn zeer waarschijnlijk veroorzaakt door verschillen in typen taken en daarom waarschijnlijk irrelevant voor onze onderzoeksvraag.

Ontwerp van een belichaamd leertraject

Het literatuuronderzoek (hoofdstuk 2) maakte ook duidelijk dat bestaande interventies onvoldoende succesvol waren om leerlingen te leren histogrammen correct te interpreteren. Uit de oplossingsstrategieën (hoofdstuk 3) bleek dat veel van deze Nederlandse leerlingen niet goed begrepen hoe en waar gegevens in histogrammen worden weergegeven. Het interpreteren van stippendiagrammen kan leerlingen helpen bij het begrijpen van histogrammen (hoofdstuk 5), omdat ze de aandacht vestigen op de as waarlangs de gegevens in zowel histogrammen als stippendiagrammen worden weergegeven (namelijk de horizontale as). Het was echter nog

onduidelijk hoe een interventie kon worden ontworpen die het leren van statistische kernconcepten door leerlingen zou ondersteunen door het leren interpreteren van stippendiagrammen en histogrammen.

Gezien de hardnekkige problemen van leerlingen met het interpreteren van histogrammen veronderstelden wij dat het in het onderwijs wellicht had ontbroken aan een belichaamde basis van hoe histogrammen worden geconstrueerd. Daarnaast vermoedden wij dat er onvoldoende aandacht was geweest voor hoe histogrammen instrumenten worden bij statistisch redeneren. Bij belichaamde ontwerpen spelen de acties van leerlingen een belangrijke rol, zoals de in de vorige alinea beschreven acties (Konold, 2007). Daarom hebben we, met een belichaamde-instrumentatieaanpak (Drijvers, 2019) als theoretische lens, een leertraject ontworpen (hoofdstuk 6) waarbij we gebruikmaakten van bevindingen en inzichten uit eerdere studies Deze ontwerpstudie is een eerste cyclus van een ontwerponderzoek naar hoe enkele van de belangrijkste aspecten van de kernconcepten van statistiek kunnen worden onderwezen via onderwijs in histogrammen. De onderzoeksvraag voor deze studie was:

> *V5: Welke opeenvolging van taken ontworpen vanuit een belichaamde instrumentatie perspectief kan het begrip van studenten van histogrammen en de onderliggende kernconcepten ondersteunen?*

Ons veronderstelde leertraject bestaat uit vijf fasen: (1) Uitlokken van het leren—ervaren van onbegrip, (2) heruitvinden van de rol van de horizontale schaal in univariate diagrammen, (3) heruitvinden van de rol van de verticale schaal in histogrammen, (4) heruitvinden van rekenkundige gemiddelden in histogrammen, en (5) bevestigen van het leren—overdracht naar andere contexten en omgevingen.

Onze meervoudige gevalstudie met vijf leerlingen (4–6 vwo) suggereerde dat de meeste vermoedens over het leertraject werden bevestigd, maar dat transfer kan worden verbeterd. Als bijdrage aan de verdere theorievorming over belichaamde instrumentatie bespraken we heuristieken voor het ontwerpproces. Bovendien lieten we zien hoe meer ingewikkelde artefacten (bv. histogrammen) kunnen worden heruitgevonden uit acties met eenvoudiger artefacten (bv. het plaatsen van stippen op een schaal, stippendiagrammen).
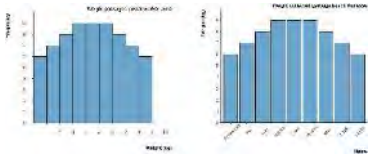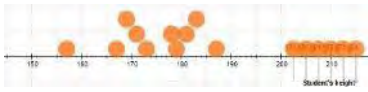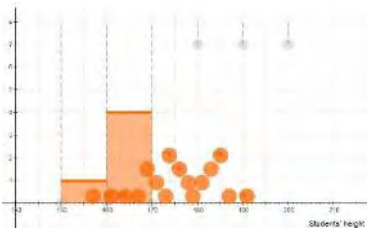
Figuur 6 Voorbeeld van pogingen van leerlingen bij het heruitvinden dat de hoogte van de staven in histogrammen gelijk is aan het aantal gemeten waarden in elke staaf, onjuist (links) en juist (rechts)
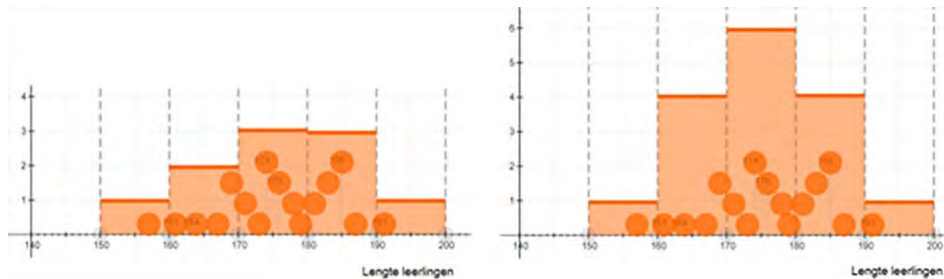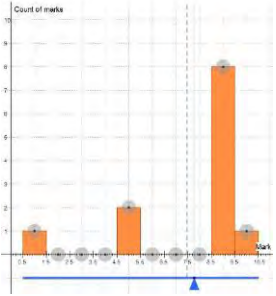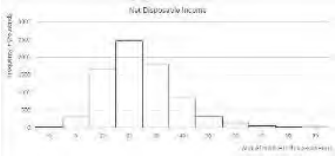

Ons hypothetisch leertraject (HLT, tabel 1) werd gebruikt om de leerlingen te helpen om enkele van de belangrijkste aspecten van de concepten van data en verdeling in histogrammen beter te begrijpen. Voor data waren deze aspecten: waar en hoe de data in een histogram worden weergegeven, inclusief dat de verticale as geen gemeten waarden weergeeft. Voor verdeling waren deze aspecten: hoe het gemiddelde wordt beïnvloed door de spreiding in en de vorm van het histogram, als voorloper voor het begrijpen van variabiliteit in een histogram. In het voortgezet onderwijs ligt de nadruk veelal op het berekenen van centrummaten en het tekenen van histogrammen (bv. Burrill, 2020). In ons ontwerp lag de nadruk op de belangrijkste concepten die we de leerlingen wilden laten begrijpen, in plaats van hen alleen te leren hoe ze een histogram moeten construeren.

Nieuw in onze HLT is dat we een belichaamde-instrumentatieaanpak gebruikten. Omdat kernconcepten niet tastbaar zijn, onderscheidden we (semiotische) artefacten—of tekens zoals Bakker en Hoffmann (2015) ze noemden—waardoor deze geleerd kunnen worden. Artefacten zijn bijvoorbeeld stippendiagrammen, intervallen, een stapel. Vervolgens hebben we deze artefacten gedeconstrueerd naar acties die tot het ontstaan van deze artefacten hadden kunnen leiden. We ontwierpen taken waarin leerlingen acties uitvoerden met deze artefacten om het zo zelf 'nieuwe' artefacten te laten heruitvinden en creëren. Door leerlingen bijvoorbeeld een staaf te laten optrekken—een actie—werden leerlingen begeleid om opnieuw uit te vinden dat de hoogte van de staven in een histogram staat voor het aantal metingen in de staaf (figuur 6). Leerlingen gebruikten artefacten waarmee ze al vertrouwd zijn, zoals bijvoorbeeld een verticale schaal. De acties van de leerlingen brachten hen ertoe aandacht te besteden aan enkele moeilijke aspecten van de weergave van data, en hun verdeling, in histogrammen.

Tabel 1 Overzicht van het hypothetische leertraject

Stap	Voorbeeldtaak	Activiteiten	Voorbeeld van vermoeden
1 Leerinitiatie: Onbegrip ervaren		Vergelijken gemiddelde en variatie in een histogram (links) en casusstaafdiagram (rechts).	H1b: Doordat leerlingen initieel verwarring en onbegrip ervaren wordt intentionaliteit en motivatie voor komende taken gecreëerd.
2 Heruitvinden van rol horizontale schaal in univariate diagrammen		Schuiven stippen naar hun correcte positie op een schaal.	H2b: Door stippen horizontaal naar hun juiste plek op een horizontale schaal te schuiven, merken leerlingen dat de *positie* van een stip de gemeten waarde weergeeft.
3 Heruitvinden rol verticale schaal in histogrammen		Over een stippendiagram bouwen leerlingen een histogram.	H3b: Door de (oranje) schuiven omhoog te slepen, merken de leerlingen dat de hoogte van de staven samenhangt met het aantal metingen in een staaf—bij gelijke klassenbreedten.

Stap	Voorbeeldtaak	Activiteiten	Voorbeeld van vermoeden
4 Rekenkundige gemiddelde in histogrammen heruitvinden		Versterken relatie tussen data en gemiddelde; ontdekken invloed van uitschieters, gaten en verdeling op gemiddelde.	H5a: Door het evenwichtspunt van het diagram te vinden, ervaren de leerlingen dat het gemiddelde kan worden gezien als het punt waar het diagram "in evenwicht" is.
5 Leren bevestigen: transfer naar andere situaties		Construeren en interpreteren van histogrammen op papier. Sorteren van histogrammen en evenbeelden.	H18b: Door het op papier tekenen van een histogram bij een frequentietabel wordt transfer naar een andere omgeving (papier) tot stand gebracht.

Nieuw in het ontwerp is ook dat we taken hebben toegevoegd die geïnstrumenteerde acties vereisen (Shvarts et al., 2021). Geïnstrumenteerde acties kunnen worden opgevat als acties die worden beïnvloed door digitale technologie, dus door de specifieke manier waarop de digitale taken zijn ontworpen. Een voorbeeld hiervan is dat de schaal voor één eenheid bij de hoogte van de staven in histogrammen niet gelijk is aan de hoogte van één stip, waardoor leerlingen door een productieve worsteling (bv. Kapur, 2014; Roth, 2019) begrepen dat deze hoogte gelijk is aan het aantal metingen (in histogrammen met constante klassenbreedte).

Als we de prestaties van leerlingen vergelijken met de vermoedens uit het HLT, suggereert onze gevalstudie dat de meeste vermoedens werden bevestigd: leerlingen ondervonden onbegrip bij de eerste stap, hadden geen moeite om de rol van de horizontale schaal opnieuw uit te vinden, vonden de rol van de verticale schaal in histogrammen na enige moeite opnieuw uit, leken het schatten van het evenwichtspunt van een histogram gemakkelijk te vinden en gaven aan dat dit het rekenkundig gemiddelde is. Uit de laatste opgaven

bleek dat de leerlingen meestal in staat waren de verworven kennis ook op papier, dus in een andere omgeving, toe te passen. Hun gebaren wezen op het gebruik van acties uit eerdere taken om de vervolgtaken op te lossen. Alles bij elkaar suggereren de resultaten dat belichaamde ervaringen gevolgd door reflectie hebben bijgedragen aan het overwinnen van enkele bekende misinterpretaties. Enkele verbeteringen voor toekomstig ontwerp zijn ook voorgesteld, zoals het toevoegen van transfertaken na de horizontale en verticale acties in het HLT. Om de noties van verdeling en variabiliteit bij leerlingen verder te ontwikkelen, is het misschien nodig om in een toekomstig ontwerp meer aandacht te besteden aan de begrippen oppervlakte en interval.

Conclusies en discussie

Een belangrijke component van statistische gecijferdheid is grafische gecijferdheid. Een histogram kan worden beschouwd als een spin in een web van kennis. Begrip van histogrammen is bijvoorbeeld een goede voorbereiding op belangrijke concepten zoals kansverdeling en kansdichtheid in de kansrekening. Het doel van dit onderzoek was bij te dragen aan een empirisch onderbouwde theorie over hoe histogrammen kunnen worden onderwezen als middel om bij te dragen aan de statistische gecijferdheid van leerlingen. Wij beantwoordden de vraag hoe leerlingen in 4–6 vwo ondersteund kunnen worden bij het begrijpen van histogrammen. Het belangrijkste antwoord is een hypothetisch leertraject (HLT) (Simon, 2020) dat erop gericht is de noties van leerlingen over enkele belangrijke aspecten van eerst data en vervolgens verdelingen in diagrammen van univariate data te ontwikkelen. Dit HLT is gebaseerd op een uitgebreide literatuurstudie en methodologisch innovatieve oogbewegingsstudies, en is bovendien ontworpen vanuit het perspectief van belichaamde instrumentatie. Het is een stap naar een domeinspecifiek instructiekader over hoe leerlingen te leren diagrammen van univariate data correct te interpreteren, waaronder histogrammen, stippen-, steel-blad- en hoeddiagrammen (Konold, 2007), boxplots, frequentiepolygonen en histodots (hoofdstuk 2). Voor toekomstige ontwerpen zou kunnen worden onderzocht of een intelligent tutoring systeem kan worden gemaakt voor automatische feedback, gebaseerd op kijkpatronen op alleen het diagramgedeelte van histogrammen. Een dergelijk systeem zou webcams vereisen die oogbewegingen kunnen meten.

Een wetenschappelijke bijdrage van ons werk is dat we hebben laten zien hoe theoretische (hoofdstuk 2) en empirische (hoofdstukken 3, 5) inzichten over de moeilijkheden van studenten met statistische concepten kunnen worden verwerkt in een opeenvolging van taken die zijn ontworpen

vanuit een belichaamd instrumentatieperspectief (hoofdstuk 6). Dit is het eerste ontwerp in het statistiekonderwijs dat gebruikmaakt van een belichaamde-instrumentatiebenadering. Daarnaast hebben we richtlijnen voor een belichaamd instrumentatieontwerp ontwikkeld, getest en geëvalueerd.

Een methodologische bijdrage van ons werk is dat we enkele nieuwe onderzoeksinstrumenten in het statistisch onderwijsonderzoek hebben geïntroduceerd en toegepast: oogbewegingsmetingen, machine learning algorithmen (MLA's) en een interpreteerbaar wiskundig model (IWM; hoofdstuk 4). Deze gereedschappen kunnen worden gebruikt voor het onderzoeken van de strategieën van leerlingen en het informeren van ontwerpen (oogbewegingsmetingen, MLA) en voor het ontwerpen van intelligente tutoring systemen die feedback geven (MLA en IWM). Bovendien toonden wij aan dat de vormen van scanpaden op het diagramgebied van statistische diagrammen de strategieën van leerlingen bij het vergelijken en schatten van gemiddelden uit deze diagrammen kunnen onthullen.

Een methodologische beperking van ons werk is de geografische selectiebias die lijkt te bestaan in de reviewstudie (hoofdstuk 2) en het aantal studenten in de oogbewegingsmetingen studie (N = 50, hoofdstukken 3–5) en meervoudige gevalstudie (N = 5; hoofdstuk 6). Toch is de aanpak geschikt voor verdere opschaling en lijken de resultaten onafhankelijk van deze specifieke omstandigheden.

Een implicatie voor onderzoek is dat oogbewegingsmetingen mogelijk nieuw licht kunnen werpen op hardnekkige didactische problemen in het wiskundeonderwijs, aangezien de scanpaden van leerlingen correcte redeneringen kunnen onthullen, zelfs wanneer antwoorden onjuist zijn. Daarnaast zouden oogbewegingsdata in combinatie met een MLA en IWM een krachtig instrument kunnen zijn voor het valideren van kwalitatieve onderzoeksbevindingen.

Een implicatie voor de onderwijspraktijk is dat histogrammen een centrale rol kunnen spelen bij het leren van statistische kernbegrippen zoals data, verdeling, variabiliteit, en centrale tendentie, en dat meer aandacht nodig is voor het kernconcept data. Bovendien wordt het aanbevolen om meer nadruk te leggen op het interpreteren van histogrammen en minder op technische vaardigheden zoals het tekenen ervan en het berekenen van gemiddelden van in diagrammen gepresenteerde data. Voor het initiële leren lijkt een belichaamde instrumentatiebenadering een vruchtbare route voor het ontwikkelen van grafische gecijferdheid van leerlingen als onderdeel van statistische gecijferdheid. Met dit in gedachten roepen wij ontwerpers op om onze richtlijnen voor belichaamde-instrumentatieontwerpen te gebruiken bij

hardnekkige didactische knelpunten in het wiskundeonderwijs en om zich bij elk aspect van de artefacten (assen, schaal, oppervlakte) af te vragen of de wiskundige acties en 'denken' door de software of de leerling moet worden gedaan.

Curriculum Vitae

Lonneke Boels was born in 1966 in Heerlen, the Netherlands. She completed her pre-university track secondary education (VWO) and received her master's degree in electrical engineering (electricity supply) from Delft University of Technology in 1991. Parallel to her studies, she held a one-year replacement position at a vocational education (MBO). She worked as an electrical engineer for about ten years, mostly at CE Delft, a consulting firm. In 2003, she received her master's degree in mathematics education from Delft University of Technology. In the meantime, she became the mother of three children and two foster children. She worked as a secondary school mathematics teacher for almost twenty years, first at the Alfrink College in Zoetermeer and later at the Christelijk Lyceum Delft (CLD). For several years, she also owned a company that was involved in remedial mathematics teaching from primary to university education. In the same period, she held replacement positions at a primary school and at three applied universities in the primary education teacher training program. She also participated in several European projects. At the CLD, in addition to teaching mathematics, she was a coordinator for the Technasium (a Technasium is a variant of pre-college and pre-university track secondary education that includes research and technical design). In 2016, she was awarded a personal grant for a PhD trajectory for teachers by the Dutch Research Council (NWO, grant number 023.007.023). This grant made it possible to combine working at the CLD with conducting doctoral research.

Dankwoord

Het doen van promotieonderzoek is als een reis. Als ervaren docent was het wennen om ineens een onervaren onderzoeker in het sociale domein te zijn. Zo'n reis kan soms eenzaam zijn. Gelukkig ben ik tijdens mijn onderzoek door heel veel mensen gesteund. Ik wil iedereen heel hartelijk bedanken daarvoor. Speciale dank gaat uit naar de volgende personen.

Allereerst diepe dankbaarheid voor mijn copromotor Arthur. Vanaf de eerste kennismaking tijdens mijn zoektocht naar een onderwerp en begeleider was er een klik. Jouw begeleiding was precies was ik nodig had: intensief als ik daarom vroeg, op afstand als ik het minder nodig had. Je reageerde altijd zeer vlot op mijn e-mails en vragen. Met jouw rake vragen en kritiekpunten kon ik soms dagen bezig zijn. Je hebt een enorm inzicht in de literatuur en kunt goed schakelen tussen theorie en praktijk. Onze gesprekken gingen niet alleen over de inhoud van het onderzoek, maar bijvoorbeeld ook over hoe docent-onderzoekers het best konden worden begeleid en hoe ik werk, privé, een eigen bedrijf en onderzoek combineerde. Ik heb dat zeer gewaardeerd. Zeer veel dank voor je tijd en inspiratie.

Ook mijn promotor Paul wil ik heel hartelijk danken. Zeker in het begin was je begeleiding, zoals ook afgesproken, meer op afstand. Onze band werd nauwer toen je tijdelijk de begeleiding overnam omdat Arthur aan een boek werkte. Ik ben je verzoeken om mijn artikelen kort te houden steeds meer gaan waarderen. Je hebt een scherp oog voor details en kan goed aangeven wanneer ze de redeneerlijn onderbreken. Je hebt ook zorg voor de mens achter het onderzoek. Je kritiek is opbouwend en stimulerend. Zeer veel dank voor je energie en inspiratie.

Verder wil ik mijn tweede promotor Wim Van Dooren heel hartelijk danken. Jouw expertise op het gebied van oogbewegingsonderzoek was zeer waardevol. Je hebt net als Arthur de gave om met een paar kritiekpunten mij dagen aan het denken te zetten. Ik waardeer het ook enorm dat je juist tijdens het schrijven van mijn laatste artikel, op het moment dat Arthur niet beschikbaar was, snel en adequaat bent ingesprongen om mij op tijd van waardevolle feedback te voorzien. Ik zie ernaar uit om onze samenwerking in een project rondom oogbewegingen voort te zetten.

Voorts bedank ik Nathalie Kuijpers hartelijk voor het controleren van het document op APA-stijl en Engels, het verbeteren van Engelse zinnen, voor het maken van één lay-out van alle afzonderlijke artikelen en hoofdstukken, voor het doorvoeren van alle wijzigingen en de enorme klus om van alle

afzonderlijke referentielijsten één lijst te maken. Ciera Lamb dank ik voor het controleren van vrijwel alle teksten in dit proefschrift op correct Amerikaans Engels.

Specifiek voor het literatuuronderzoek bedank ik Marianne van Dijke-Droogers voor de tweede codering van de data, Floris Kooij en Rian Ligthart voor hun bijdrage aan het digitaliseren van de classificatie van de kernconcepten, het screenen van een aantal dissertaties, het toevoegen van aantallen leerlingen, studenten en landen aan de gegevens en het zoeken van literatuur om na te gaan of geen belangrijke studies waren gemist (via de procedure 'backward snowballing'). Verder bedank ik ICT ontwikkelaar Kees van Eijden voor het schrijven van sommige R-code die gebruikt is voor het maken van grafieken, in het bijzonder de ongestapelde dotplot en histodot.

Specifiek voor het oogbewegingsonderzoek bedank ik medeauteur Rutmer Ebbes voor zijn bijdrage aan de pilotstudie voor de oogbewegingen, Aline Boels voor het programmeren van de html-bestanden met de items voor de oogbewegingsstudie, Juri Boels voor de meeste transcripties van de mondelinge data met behulp van geautomatiseerde transcripties als start (Oral history software, Yilmaz & Gompel, n. d.), Iljo Boels voor het exporteren van de gaze plots en heatmaps, Alex Lyford voor het berekenen van nauwkeurigheids- en precisiematen van de oogbewegingsdata, Gerben van der Hoek voor de tweede codering van de oogbewegingsdata en Hidde Leplaa voor hulp bij het opzetten van een NVivo structuur. Verder bedank ik alle mensen die betrokken waren bij de eye-tracking seminars aan de UU en die met me mee hebben gedacht, feedback hebben gegeven tijdens presentaties en me behoed hebben voor een aantal beginnersfouten (zie de bijlage van hoofdstuk 3). In het bijzonder bedank ik Ellen Kok, Margot van Wermeskerken, Roy Hessels, Ignace Hooge en Jos Jaspers. Ignace heeft bovendien waardevolle feedback gegeven op het onderzoeksvoorstel dat aan dit promotieonderzoek ten grondslag lag. Dank ook aan de Faculteit Sociale Wetenschappen voor het lenen van de laptop en Tobii-XII-60 eye-tracker voor dit onderzoek. Tot slot ook grote dank aan alle leerlingen en docenten die vrijwillig aan mijn onderzoeken deelnamen.

Ik bedank mijn medeauteurs van de artikelen voor het meedenken over het onderzoek, de uitvoering en hun waardevolle feedback. Enrique, bedankt voor alle energie die je belangeloos hebt gestoken in de analyses van de oogbewegingsdata met modellen en algoritmen. Ik weet dat de herziening van ons artikel soms erg slecht voor je uitkwam. Dank dat je dan toch doorging! Alex, bedankt voor je geweldige vaardigheden in R waarmee je de oogbewegingsdata hebt geanalyseerd om inzicht te krijgen of dotplottaken invloed kunnen hebben op hoe leerlingen histogrammen interpreteren. Het is

fantastisch om met jou samen te werken en ik zie uit naar ons volgende artikel over dubbele histogramtaken. Dank aan jullie beiden dat ik heb mogen leren van jullie enorme ervaring op het gebied van machine learning algoritmen. De ideeën die we gezamenlijk hebben voor vervolgonderzoek vind ik enorm inspirerend en ik zie uit naar onze toekomstige projecten. Anna, bedankt voor het assisteren tijdens de laatste dag van de dataverzameling bij het oogbewegingsonderzoek, het programmeren in de digitale leeromgeving van de belichaamde taken, het meedenken over de ontwerpen ervan en de vele sessies over belichaamde ontwerpen en instrumentatie op het FI. Je hebt een geweldig theoretisch inzicht, enorm geduld en goede feeling voor wat leerlingen kunnen. Ik hoop dat we onze discussies over hoe de theorie praktisch kan worden gemaakt nog jaren zullen voortzetten in mooie projecten.

Ik bedank ook mijn kamergenoten op het FI, in het bijzonder Marianne van Dijke-Droogers, Rosa Alberto, Nathalie van der Wal, Winnifred Wijnkers en Annemiek van Leendert. Met jullie voerde ik inspirerende discussies over allerlei onderwerpen, ook buiten het promoveren om. Dank voor jullie steun en waardevolle gesprekken. Verder dank ik alle collega-promovendi en andere collega's van het FI met wie ik heb mogen samenwerken of discussies voeren. Ik dank Mariozee Wintermans en haar collega's voor de ondersteuning vanuit het secretariaat en het Freudenthal Instituut voor het mogen gebruiken van vele faciliteiten, waaronder het Teaching en Learning Lab.

Ik ben heel veel collega's op het Christelijk Lyceum Delft dank verschuldigd voor hun bijdragen aan mijn onderzoek en onderzoeksvoorstel. Dank aan de directie voor het meedenken en mogelijk maken van dit onderzoek. Speciale dank aan collega's die al tijdens het schrijven van mijn onderzoeksvoorstel meedachten: Suzanne van der Waal, Inge Verhoev, Remko Schoot Uiterkamp, Roel de Rijk, Josje Schokkenbroek, Thalie Beudeker en Simon Belder. Dank ook aan de wiskundesectie, het vwo bovenbouwteam en alle anderen die regelmatig informeerden hoe het met mijn onderzoek ging en die meedachten of werk overnamen omdat ik minder beschikbaar was voor school. Ik bedank ook Jos Tolboom voor zijn support bij het schrijven van het onderzoeksvoorstel.

Daarnaast bedank ik de mensen van de SIG27 Earli gemeenschap en de statistiekonderwijs-onderzoekers die mij hebben geïnspireerd tijdens conferenties. Op het gevaar af mensen tekort te doen of te vergeten, bedank ik specifiek Halszka Jarodzka voor haar inspiratie en aanmoediging bij het doen van oogbewegingsonderzoek, Dani Ben-Zvi en Katie Makar voor hun waardevolle feedback tijdens vele conferenties inclusief de door Marianne en

mij georganiseerde pre-CERME conferentie in Utrecht en de SRTL12-conferentie die uiteindelijk online werd gehouden. Ik bedank ook de anonieme reviewers van alle artikelen voor hun behulpzame suggesties en feedback.

Grote dank gaat ook uit naar mijn beide ouders, Jan en Nolda. Jullie hebben mij altijd gesteund en mijn keuzes enerzijds als volstrekt normaal beschouwd (studie Elektrotechniek, ook al was ik een van de weinige vrouwen) en anderzijds als heel bijzonder gevonden (een universitaire studie, een tweede master en vervolgens dit promotieonderzoek). Ik weet dat jullie heel trots op me zijn. Het was zwaar om tijdens mijn promotieonderzoek, in de coronatijd, afscheid te moeten nemen van mijn vader. Hij had de plechtigheid graag bijgewoond maar ik weet dat hij er in gedachten bij is. Het is fijn dat ik hem nog uitgebreid heb kunnen bedanken voor alles wat hij voor mij gedaan heeft. Nolda, dankjewel dat je me hebt geleerd dat een slimme meid op haar toekomst voorbereid is. Hoe jij werk en moederschap combineerde was voor mij een inspirerend voorbeeld. Ik denk bovendien dat ik mijn schrijfvaardigheid van jou heb. Dank voor al je liefde, steun, aanmoedigingen en de vele hulp in huis wanneer dat nodig was.

Verder bedank ik de liefde van mijn leven. Willem, je hebt me al die tijd met raad en daad bijgestaan op alle mooie momenten maar ook op alle momenten dat het leven een tegenslag voor mij in petto had. Je hebt me gesteund in deze reis ook al wist je bij voorbaat dat het ook van jou offers zou vragen. Je nam het grootste deel van het huishouden op je, naast je drukke baan. Je hebt vele vakanties in het buitenland alleen gewandeld omdat ik overdag aan een artikel voor mijn proefschrift wilde werken. Alleen de avonden waren dan voor ons samen. Zelfs dit dankwoord schrijf ik tijdens onze vakantie. Dank voor al je liefde, je steun, je opofferingen en je geduld.

Dank ook aan mijn zussen Daniëlle en Eveline die mij steunden, aanmoedigden en regelmatig vertelden hoe trots ze zijn op mijn werk. Daniëlle, heel erg bedankt voor het mooie ontwerp van de voorkant van mijn proefschrift en de door data (stippen) omgeven nummering bij de hoofdstuktitels.

Daarnaast bedank ik mijn kinderen Juri, Aline en Iljo. Ook jullie hebben mij regelmatig moeten missen op belangrijke momenten als ik weer eens niet beschikbaar was omdat ik een ingewikkelde Engelse tekst aan het schrijven was, aan het lesgeven was of naar een conferentie was. Ditzelfde geldt voor mijn pleegkinderen Boris en Adriaan en mijn pleegkleindochter Sofia. Zeker in de coronatijd en daarna heb ik de contacten met jullie veel te lang

verwaarloosd. Ik hoop dat jullie me dat vergeven en dat ik de draad weer met jullie kan oppakken.

Verder bedank ik Barbara Meinert. Je houdt al jaren ons huis schoon en strijkt al onze kleren. Dankzij jouw hulp kon ik me concentreren op mijn werk en promotieonderzoek. Ook bedank ik Inge Siebring die mijn bedrijf draaiende hield toen ik er steeds minder tijd voor had. Je was zakelijk en persoonlijk mijn steun en toeverlaat. Ik bedank ook Gert de Kleuver die mij op een cruciaal moment vroeg waarom ik niet nu ging promoveren, in plaats van ooit.

Tot slot bedank ik al mijn vriendinnen en vrienden. Ik heb jullie flink verwaarloosd. Ik hoop dat ik de tijd krijg de komende jaren om dat goed te maken. Bedankt voor al jullie steun, aanmoedigingen en liefde. In het bijzonder bedank ik Marjanne Klom voor haar inspirerende gesprekken over onderwijs, onderzoek en het leven en Afke Posthuma voor haar meedenken over academische vereisten. Ik bedank Paul Haima voor de wiskundige grappen die hij me steeds stuurde die me hielpen om het luchtig te houden. Ik bedank Marianne van Dijke-Droogers voor haar collegiale steun, de fijne conferenties samen en haar bijdrage in de rol van paranimf. Heel veel dank ook aan Mariet Lohman die al jaren mijn vriendin is. Je hebt een flinke klus gehad aan het mede-organiseren van het promotiefeest en het uitnodigen van alle gasten.

Met de verdediging van dit proefschrift eindigt een intensieve en heel leerzame periode. Het is ook het begin van een nieuwe reis in onderwijsonderzoek. Ik zie reikhalzend uit naar dat vervolg.

Publications and presentations related to this dissertation

Chapter 1

Boels, L. (2018). Kleintje didactiek. Statistiek: Gemiddelde en verdeling [Little didactics. Statistics: Mean and distribution]. *Euclides, 93*(7), 20–21. https://archief.vakbladeuclides.nl/bestanden/093_2017-18_07.pdf

Boels, L. (2018). Kleintje didactiek. Statistiek kan levens redden [Little didactics. Statistics can save lives]. *Euclides, 94*(3), 10. https://archief.vakbladeuclides.nl/bestanden/094_2018-19_03.pdf

Boels, L. (2019). FIzier. Wat elke docent zou moeten weten over histogrammen [What every teacher needs to know about histograms, translation author]. *Euclides, 94*(4), 10–13.

Boels, L. (2020). Kleintje didactiek. Datasaurus [Little didactics. Datasaurus]. *Euclides 95*(5), 17. https://archief.vakbladeuclides.nl/bestanden/095_2019-20_05.pdf

Chapter 2

Boels, L. (2017). Kleintje didactiek. Meetniveaus [Little didactics. Measurement levels]. *Euclides, 93*(4), 28–29. https://archief.vakbladeuclides.nl/bestanden/093_2017-18_04.pdf

Boels, L. (2017). Histogrammen: Lastiger dan gedacht [Histograms: More challenging than thought]. Presentation at the *studiedag Nederlandse Vereniging voor Wiskundeleraren [Study day Dutch Association for Teachers of Mathematics].*

Boels, L., Bakker, A., Drijvers, P., & Van Dooren, W. (2017). Conceptual difficulties with histograms: A review. In B. Kaur, W. K. Ho, T. L. Toh, & B. H. Choy (Eds.), *Proceedings of the 41st PME Conference,* Vol. 1 (p. 172). Singapore: PME.

Boels, L., Bakker, A., Drijvers, P., & Van Dooren, W. (2017). Conceptuele problemen bij het gebruik van histogrammen: Een reviewstudie [Conceptual difficulties during the use of histograms]. Paper presented at the *Onderwijs Research Dagen [Education Research Days].*

Boels, L., Bakker, A., Drijvers, P., & Van Dooren, W. (2016). Students' interpretations of histograms: A review. Poster presented at *the 13th International Congress on Mathematical Education*.

Boels, L., Bakker, A., Drijvers, P., & Van Dooren, W. (2016). Students' interpretations of histograms: A review. Poster presented at the *Onderwijs meets onderzoek conferentie [Education meets research conference].*

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019). Conceptual difficulties when interpreting histograms: A review. *Educational Research Review, 28*, Article 100291, 26 p. https://doi.org/10.1016/j.edurev.2019.100291

Two short videos (5 mins) related to the ERR article:

https://youtu.be/zpRHhixoYmg   and https://youtu.be/5od2uB908PI

Chapter 3

Boels, L. (2018). Histogrammen: Lastiger dan gedacht [Histograms: More challenging than thought]. Presentation at the *wiskundedialoog [math dialogue]*.

Boels, L. (2020). Waarom leerlingen weinig van histogrammen begrijpen [Why students understand little about histograms]. Presentation at the *Nationale Wiskunde Dagen [National Mathematics Days]*.

Boels, L. (2022). Kleintje didactiek. Interpreteren van histogrammen en stippengrafieken [Tiny didactics. Interpreting histograms and dotplots]. *Euclides, 97*(5), 20. https://archief.vakbladeuclides.nl/jaargang_097.html

Boels, L., Bakker, A., & Drijvers, P. (2019). Eye tracking secondary school students' strategies when interpreting statistical graphs. In M. Graven, H. Venkat, A.A. Essien, & P. Vale (Eds.), *Proceedings of the Forty-third Conference of the International Group for the Psychology of Mathematics Education, 2,* (pp. 113–120). PME. http://www.igpme.org/publications/

Boels, L., Bakker, A., & Drijvers, P. (2019). Unravelling teachers' strategies when interpreting histograms: An eye-tracking study. In U.T. Jankvist, M. Van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education*, (pp. 888–895). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. http://www.mathematik.uni-dortmund.de/~prediger/ERME/CERME11_ Proceedings_2019.pdf (or: https://hal.archives-ouvertes.fr/hal-02411575/document)

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2022). *Secondary school students' strategies when interpreting histograms and case-value plots: an eye-tracking study.* [Manuscript submitted for publication]. Freudenthal Institute, Utrecht University.

Boels, L., Ebbes, R., Bakker, A., Van Dooren, W., & Drijvers, P. (2018). Revealing conceptual difficulties when interpreting histograms: An eye-tracking study. Invited paper, refereed. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics,* pp. 1–4, International Statistical Institute. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8E2.pdf

Boels, L., Ebbes, R., Bakker, A., Van Dooren, W., & Drijvers, P. (2018). *Students' strategies when solving problems about statistical graphs.* Poster presented at the 2nd Earli SIG27 Conference. Warsaw, Poland.

Chapter 4

Boels, L., Garcia Moreno-Esteva, E., Bakker, A., & Drijvers, P. (accepted). Automatic gaze-based identification of students' strategies in histogram tasks: Machine

learning algorithm and interpretable model. *International Journal for Artificial Intelligence in Education*.

Hannula, M., Garcia Moreno-Esteva, E., & Boels, L. (2023). Histogram recognition: An algorithmic model of eye movement. Presentation at *The 20th Biennial EARLI Conference.*

Chapter 5

Boels, L., & Lyford, A. (2022). Gaze-based machine learning analysis of students' learning during solving graph items. Presentation at the *EARLI SIG27.*

Boels, L., Lyford, A., Bakker, A., & Drijvers, P. (in press). Assessing students' interpretation of histograms before and after interpreting dotplots: A gaze-based machine learning analysis. *Frontline Learning Research*.

Boels, L., & Van Dooren, W. (2023). Secondary school students interpreting and comparing dotplots: An eye-tracking study. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education, 2,* (pp. 123–130). PME.

Lyford, A., & Boels, L. (2022). Using machine learning to understand students' gaze patterns on graphing tasks. Invited Paper - Refereed. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022)* (pp. 1–6)*.* ISI/IASE. https://doi.org/10.52041/iase.icots11.T8D2

Chapter 6

Boels, L. (2019). Kleintje didactiek. Histodot—Een nieuw type grafiek [Little didactics. Histodot—A new type of graph], *Euclides 94*(5), 31. https://archief.vakbladeuclides.nl/bestanden/094_2018-19_05.pdf

Boels, L. (2021). Designing embodied tasks in statistics education for Grade 10–12. Invited extended paper. *The 14th International Congress on Mathematical Education,* Shanghai, China. (8 p.)

Boels, L. (2022). Embodied design van histogramtaken. Presentation at the Onderwijs meets onderzoek conferentie *[Education meets research conference].*

Boels, L., Alberto, R. & Shvarts, A. (2023). Actions behind mathematical concepts: A logical-historical analysis. *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education.*

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2022). *Understanding histograms in upper-secondary school: Embodied design of a learning trajectory.* [Manuscript submitted for publication]. Freudenthal Institute, Utrecht University.

Boels, L., & Shvarts, A. (2023). Introducing density histograms to Grades 10 and 12 students. Design and tryout of an intervention inspired by embodied instrumentation. In G. Burrill, L. de Oliveria Souza & E. Reston (Eds.), *Research on*

students' interactions with data in teaching statistics: international perspectives. Advances in Mathematics Education. Springer Nature.

Video of a presentation at the mathematics seminar (2019):

https://www.youtube.com/watch?v=SYKPbkqVmmo

Chapter 7

Boels, L. (2023). Reflections on gaze data in statistics education. *Teaching Statistics,* 1–12. https://doi.org/10.1111/test.12340

Boels, L., Bakker, A., & Drijvers, P. (2022). Learning from gaze data in statistics education research: the case of drawing inferences from graphs. Presentation at *The Twelfth International Research Forum on Statistical Reasoning, Thinking and Literacy* (SRTL-12).

ICO Dissertation Series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, Radboud University Nijmegen, University of Amsterdam, University of Antwerp, University of Ghent, KU Leuven, Université Catholique de Louvain, University of Groningen, University of Twente, Utrecht University, Vrije Universiteit Amsterdam, and Wageningen University, and formerly Tilburg University (until 2002).

304. Tacoma, S.G. (15-11-2020) *Automated intelligent feedback in university statistics education* Utrecht: Utrecht University

305. Boonk, L.M. (04-12-2020) *Exploring, measuring, and evaluating parental involvement in vocational education and training* Heerlen: Open University of the Netherlands

306. Kickert, R. (04-12-2020) *Raising the bar: Higher education students' sensitivity to the assessment policy* Rotterdam: Erasmus University Rotterdam

307. Van der Wal, N.J. (09-12-2020) *Developing Techno-mathematical Literacies in higher technical professional education* Utrecht: Utrecht University

308. Vaessen, B.E. (08-01-2021) *Students' perceptions of assessment and student learning in higher education courses.* Eindhoven: Eindhoven University of Technology

309. Maureen, I.Y. (15-01-2021) *Story time in early childhood education: designing storytelling activities to enhance (digital) literacy development*. Enschede: University of Twente

310. Van Alten, D.C.D. (19-03-2021) *Flipped learning in secondary education history classrooms: what are the effects and what is the role of self-regulated learning*. Utrecht: Utrecht University

311. Gestsdóttir, S.M. (08-03-2021) *Observing history teaching: historical thinking and reasoning in the upper secondary classroom.* Amsterdam: University of Amsterdam

312. Chim, H.Q. (30-03-2021) *Physical Activity Behavior and Learning in Higher Education.* Maastricht: Maastricht University

313. Krijnen, E. (15-04-2021)*Family literacy in context: Exploring the compatibility of a family literacy program with children's homes and schools.* Rotterdam Erasmus University Rotterdam

314. Stolte, M. (07-05-2021) *(In)attention for creativity: Unraveling the neural and cognitive aspects of (mathematical) creativity in children.* Utrecht: Utrecht University

315. Rathé, S. (12-05-2021) *Focusing on numbers – An investigation of the role of children's spontaneous focusing on Arabic number symbols in early mathematical development*. Leuven: KU Leuven

316. Theelen, H. (12-05-2021) *Looking around in the classroom. Developing preservice teachers' interpersonal competence with classroom simulations.* Wageningen: Wageningen University

317. De Jong, L.A.H. (20-05-2021*) Teacher professional learning and collaboration in secondary schools.* Leiden: Leiden University

318. Sincer, I. (20-05-2021) *Diverse Schools, Diverse Citizens? Teaching and learning citizenship in schools with varying student populations.* Rotterdam: Erasmus University Rotterdam

319. Slijkhuis, E.G.J. (20-05-2021) *Fostering active citizenship in young adulthood.* Groningen: University of Groningen

320. Groothuijsen-Vrancken, S.E.A. (02-06-2021) *Quality and impact of practice-oriented educational research*. Utrecht: Utrecht University

321. Hingstman, M. (07-06-2021) *Supporting struggling students: prevention and early intervention with Success for All*. Groningen: University of Groningen

322. Gerdes, J. (14-06-2021) *All inclusive? Collaboration between teachers, parents and child support workers for inclusive education in prevocational schools.* Amsterdam: Vrije Universiteit Amsterdam

323. Bai, H. (18-06-2021) *Divergent thinking in young children*. Utrecht: Utrecht University

324. Wijnker, W. (23-06-2021) *The Unseen Potential of Film for Learning: Film's Interest Raising Mechanisms Explained in Secondary Science and Math*. Utrecht: Utrecht University

325. Brummer, L. (24-09-2021). *Unrooting the illusion of one-size-fits-all feedback in digital learning environments.* Groningen: University of Groningen

326. Veldman, M.A. (01-07-21) *Better together, social outcomes of cooperative learning in the first grades of primary education*. Groningen: University of Groningen

327. Wang, J. (06-07-2021) *Technology integration in education: Policy plans, teacher practices, and student outcomes.* Leiden: Leiden University

328. Zhang, X. (06-07-2021) *Teachers' teaching and learning motivation in China*. Leiden: Leiden University

329. Poort, I.C. (02-09-2021) *Prepared to engage? Factors that promote university students' engagement in intercultural group work*. Groningen: University of Groningen

330. Guo, P. (07-09-2021) *Online project-based higher education Student collaboration and outcomes.* Leiden: Leiden University

331. Jin, X. (21-09-2021) *Peer feedback in teacher professional development.* Leiden: Leiden University

332.   Atherley, E.N. (27-09-2021) *Beyond the struggles: Using social-developmental lenses on the transition to clinical training.* Maastricht: Maastricht University

333.   Martens, S.E. (15-10-2021) *Building student-staff partnerships in higher education.* Maastricht: Maastricht University

334.   Ovbiagbonhia, R. (08-11-2021) *Learning to innovate: how to foster innovation competence in students of Built Environment at universities of applied science*. Wageningen: Wageningen University

335.   Van den Boom-Muilenburg , S.N. (11-11-2021) *The role of school leadership in schools that sustainably work on school improvement with professional learning communities*. Enschede: University of Twente

336.   Sachishal, M.S.M. (11-11-2021) *Science interest - Conceptualizing the construct and testing its predictive effects on current and future behavior.* Amsterdam: University of Amsterdam

337.   Meeuwissen, S.N.E. (12-11-2021*) Team learning at work: Getting the best out of interdisciplinary teacher teams and leaders*. Maastricht: Maastricht University

338.   Keijzer-Groot, A.F.J.M. (18-11-2021) *Vocational identity of at-risk youth – Tailoring to support career chances*. Leiden: Leiden University

339.   Wolthuis, F. (25-11-2021) *Professional development in practice. Exploring how lesson study unfolds in schools through the lens of organizational routines*. Groningen: University of Groningen

340.   Akkermans-Rutgers, M. (06-12-2021) *Raising readers. Stimulating home-based parental literacy involvement in the first, second, and third grade*. Groningen: University of Groningen

341.   Hui, L. (06-12-2021) *Fostering Self-Regulated Learning: The Role of Perceived Mental Effort*. Maastricht: Maastricht university

342.   Jansen, D. (08-12-2021) *Shadow education in the Netherlands: The position of shadow education in the educational landscape and students' school careers*. Amsterdam: University of Amsterdam

343.   Kamphorst, F. (15-12-2021) *Introducing Special Relativity in Secondary Education*. Utrecht: Utrecht University

344.   Eshuis, E.H. (17-12-2021*) Powering Up Collaboration and Knowledge Monitoring: Reflection-Based Support for 21st-Century Skills in Secondary Vocational Technical Education*. Enschede: University of Twente

345.   Abacioglu, C. S. (18-01-2022) *Antecedents, Implications, and Professional Development of Teachers' Multiculturalism.* Amsterdam: University of Amsterdam

346.   Wong, J. (21-01-2022) *Enhancing Self-Regulated Learning Through Instructional Support and Learning Analytics in Online Higher Education.* Rotterdam: Erasmus University Rotterdam

347.   Schophuizen, M.J.F. (18-02-2022) *Educational innovation towards organizational development: the art of governing open and online*

education in Dutch higher education institutions. Heerlen: Open University of the Netherlands

348. Smeets, L.H. (18-03-2022) *The auditor learning curve: professional development through learning from errors and coaching.* Maastricht: Maastricht University

349. Beuken, J.A. (25-03-2022) *Waves towards harmony – learning to collaborate in healthcare across borders.* Maastricht: Maastricht University

350. Radovic, S. (25-03-2022) *Instructional design according to the mARC model: Guidelines on how to stimulate more experiential learning in higher education.* Heerlen: Open University of the Netherlands

351. Delnoij, L.E.C*.* (01-04-2022) *Self-assessment for informed study decisions in higher education: A design-based validation approach.* Heerlen: Open University of the Netherlands

352. Pieters, J. (01-04-2022*) Let's talk about it: Palliative care education in undergraduate medical curricula.* Maastricht: Maastricht University

353. Vulperhorst, J.P. (08-04-2022) *Students' interest development prior, during and after the transition to a higher education programme.* Utrecht: Utrecht University

354. Aben, J.E.J. (21-04-2022*) Rectifying errors: A reconceptualization of the role of errors in peer-feedback provision and processing.* Groningen: University of Groningen

355. Kooloos, C.C. (19-05-2022) *Eye on Variety: The teacher's work in developing mathematical whole-class discussions.* Nijmegen: Radboud University

356. El Majidi, A. (20-05-2022) *Debate as a Tool for L2 Learning. Investigating the Potential of In-Class Debates for Second Language Learning and Argumentation Skills.* Utrecht: Utrecht University

357. Van Halem, N. (24-05-2022) *Accommodating Agency-Supportive Learning Environments in Formal Education.* Amsterdam: Vrije Universiteit Amsterdam

358. Lee, Y.J. (8-06-2022) *The medical pause in simulation training.* Maastricht: Maastricht University

359. Loopers, J.H. (14-06-2022) *Unravelling the dynamics of intrinsic motivation of students with and without special educational needs.* Groningen: University of Groningen

360. Neroni, J. (17-06-2022) *The Adult Learning Open University Determinants (ALOUD) study: psychosocial factors predicting academic success in adult distance education.* Heerlen: Open University of the Netherlands

361. Schlatter, E. (17-06-2022*) Individual differences in children's scientific reasoning.* Nijmegen: Radboud University

362. Bransen, D. (22-06-2022) *Beyond the self: A network perspective on regulation of workplace learning.* Maastricht: Maastricht University

363.   Valero Haro, A. (27-06-2022) *Fostering Argumentation with Online Learning Systems in Higher Education.* Wageningen: Wageningen University

364.   Schat, E. (05-07-2022) *Integrating intercultural literary competence: An intervention study in foreign language education.* Utrecht University

365.   Soppe, K.F.B. (07-07-2022) *To Match or not to Match? Improving Student-Program Fit in Dutch Higher Education.* Utrecht: Utrecht University

366.   Biwer, F. (08-07-2022) *Supporting students to study smart – a learning sciences perspective.* Maastricht: Maastricht University

367.   Weijers, R.J. (15-09-2022) *Nudging Towards Autonomy: The effect of nudging on autonomous learning behavior in tertiary education.* Rotterdam: Erasmus University Rotterdam

368.   Van der Linden, S. (14-10-2022) *Supporting teacher reflection in video-coaching settings.* Enschede: University of Twente

369.   Wijnen, F. (28-10-2022) *Using new technology and stimulating students' higher-order thinking: a study on Primary school teachers' attitudes.* Enschede: University of Twente

370.   De Jong, W.A. (02-11-2022) *Leading collaborative innovation in schools.* Utrecht: Utrecht University

371.   Bron, R. (04-11-2022) *Collaborative course design in higher education – a team learning perspective.* Enschede: University of Twente

372.   De Vries, J.A. (18-11-2022) *Supporting teachers in formative assessment in the classroom.* Enschede: University of Twente

373.   Le, T.I.N.H. (29-11-22) *Towards a democratic school.* Leiden: Leiden University

374.   Wildeman, E. (30-11-2022) *Vocational teachers' integrated language teaching. On the role of language awareness and related teaching behaviour.* Eindhoven: Eindhoven University of Technology

375.   Wolterinck, C.H.D. (2-12-2022), *Teacher professional development in assessment for learning.* Enschede: University of Twente

376.   Kuang, X. (07-12-2022) *Hypothesis generation, how to put students into motion.* Enschede: University of Twente

377.   Emhardt, S.N. (15-12-2022) *You see? Investigating the effects of different types of guidance in eye movement modeling examples.* Heerlen: Open University of the Netherlands

FI Scientific Library

(formerly published as CD-β Scientific Library)

116. Huang, L. (2022). *Inquiry-based learning in lower-secondary mathematics education in China (Beijing) and the Netherlands*.

115. Jansen, S. (2022). *Fostering students' meta-modelling knowledge regarding biological concept-process models.*

114. Pieters, M.L.M. (2022). *Between written and enacted: Curriculum development as propagation of memes. An ecological-evolutionary perspective on fifty years of curriculum development for upper secondary physics education in the Netherlands*

113. Veldkamp, A. (2022). *No Escape! The rise of escape rooms in secondary science education.*

112. Kamphorst, F. (2021). *Introducing Special Relativity in Secondary Education.*

111. Leendert, A.-M. J. M. van (2021). *Improving Reading and Comprehending Mathematical Expressions in Braille.*

110. Gilissen, M. G. R. (2021). *Fostering Students' System Thinking in Secondary Biology Education.*

109. Dijke-Droogers, M.J.S. van (2021). *Introducing Statistical Inference: Design and Evaluation of a Learning Trajectory.*

108. Wijnker, W. (2021). *The Unseen Potential of Film for Learning. Film's Interest Raising Mechanisms Explained in Secondary Science and Mathematics Education.*

107. Groothuijsen, S. (2021). *Quality and impact of practice-oriented educational research.*

106. Wal, N.J. van der (2020). *Developing Techno-mathematical Literacies in higher technical professional education.*

105. Tacoma, S. (2020). *Automated intelligent feedback in university statistics education*.

104. Zanten, M. van (2020). *Opportunities to learn offered by primary school mathematics textbooks in the Netherlands*

103. Walma, L. (2020). *Between Morpheus and Mary: The Public Debate on Morphine in Dutch Newspapers, 1880-1939*

102. Van der Gronde, A.G.M.P. (2019). *Systematic Review Methodology in Biomedical Evidence Generation.*

101. Klein, W. (2018). *New Drugs for the Dutch Republic. The Commodification of Fever Remedies in the Netherlands (c. 1650-1800).*

100. Flis, I. (2018). *Discipline Through Method - Recent history and philosophy of scientific psychology (1950-2018).*

99. Hoeneveld, F. (2018). *Een vinger in de Amerikaanse pap. Fundamenteel fysisch en defensie onderzoek in Nederland tijdens de vroege Koude Oorlog.*

98.   Stubbé-Albers, H. (2018). *Designing learning opportunities for the hardest to reach: Game-based mathematics learning for out-of-school children in Sudan*.

97.   Dijk, G. van (2018). *Het opleiden van taalbewuste docenten natuurkunde, scheikunde en techniek: Een ontwerpgericht onderzoek*.

96.   Zhao, Xiaoyan (2018). *Classroom assessment in Chinese primary school mathematics education.*

95.   Laan, S. van der (2017). *Een varken voor iedereen. De modernisering van de Nederlandse varkensfokkerij in de twintigste eeuw.*

94.   Vis, C. (2017). *Strengthening local curricular capacity in international development cooperation.*

93.   Benedictus, F. (2017). *Reichenbach: Probability & the A Priori. Has the Baby Been Thrown Out with the Bathwater?*

92.   Ruiter, Peter de (2016). *Het Mijnwezen in Nederlands-Oost-Indië 1850-1950.*

91.   Roersch van der Hoogte, Arjo (2015). *Colonial Agro-Industrialism. Science, industry and the state in the Dutch Golden Alkaloid Age, 1850- 1950.*

90.   Veldhuis, M. (2015). *Improving classroom assessment in primary mathematics education.*

89.   Jupri, Al (2015). *The use of applets to improve Indonesian student performance in algebra.*

88.   Wijaya, A. (2015). *Context-based mathematics tasks in Indonesia: Toward better practice and achievement.*

87.   Klerk, S. (2015). *Galen reconsidered. Studying drug properties and the foundations of medicine in the Dutch Republic ca. 1550-1700.*

86.   Krüger, J. (2014). *Actoren en factoren achter het wiskundecurriculum sinds 1600.*

85.   Lijnse, P.L. (2014). *Omzien in verwondering. Een persoonlijke terugblik op 40 jaar werken in de natuurkundedidactiek.*

84.   Weelie, D. van (2014). *Recontextualiseren van het concept biodiversiteit*.

83.   Bakker, M. (2014). *Using mini-games for learning multiplication and division: a longitudinal effect study.*

82.   Ngô Vũ Thu Hăng (2014). *Design of a social constructivism-based curriculum for primary science education in Confucian heritage culture*.

81.   Sun, L. (2014). *From rhetoric to practice: enhancing environmental literacy of pupils in China*.

80.   Mazereeuw, M. (2013). *The functionality of biological knowledge in the workplace. Integrating school and workplace learning about reproduction.*

79.   Dierdorp, A. (2013). *Learning correlation and regression within authentic contexts.*

78. Dolfing, R. (2013). *Teachers' Professional Development in Context-based Chemistry Education. Strategies to Support Teachers in Developing Domain-specific Expertise.*

77. Mil, M.H.W. van (2013). *Learning and teaching the molecular basis of life*.

76. Antwi, V. (2013). *Interactive teaching of mechanics in a Ghanaian university context*.

75. Smit, J. (2013). *Scaffolding language in multilingual mathematics classrooms*.

74. Stolk, M. J. (2013). *Empowering chemistry teachers for context-based education. Towards a framework for design and evaluation of a teacher professional development programme in curriculum innovations.*

73. Agung, S. (2013). *Facilitating professional development of Madrasah chemistry teachers. Analysis of its establishment in the decentralized educational system of Indonesia.*

72. Wierdsma, M. (2012). *Recontextualising cellular respiration*.

71. Peltenburg, M. (2012). *Mathematical potential of special education students.*

70. Moolenbroek, A. van (2012). *Be aware of behaviour. Learning and teaching behavioural biology in secondary education.*

69. Prins, G. T., Vos, M. A. J., & Pilot, A. (2011). *Leerlingpercepties van onderzoek & ontwerpen in het technasium.*

68. Bokhove, Chr. (2011). *Use of ICT for acquiring, practicing and assessing algebraic expertise.*

67. Boerwinkel, D. J., & Waarlo, A. J. (2011). *Genomics education for decision-making. Proceedings of the second invitational workshop on genomics education, 2-3 December 2010.*

66. Kolovou, A. (2011). *Mathematical problem solving in primary school.*

65. Meijer, M. R. (2011). *Macro-meso-micro thinking with structure-property relations for chemistry. An explorative design-based study*.

64. Kortland, J., & Klaassen, C. J. W. M. (2010). *Designing theory-based teaching-learning sequences for science. Proceedings of the symposium in honour of Piet Lijnse at the time of his retirement as professor of Physics Didactics at Utrecht University*.

63. Prins, G. T. (2010). *Teaching and learning of modelling in chemistry education. Authentic practices as contexts for learning*.

62. Boerwinkel, D. J., & Waarlo, A. J. (2010). *Rethinking science curricula in the genomics era. Proceedings of an invitational workshop*.

61. Ormel, B. J. B. (2010). *Het natuurwetenschappelijk modelleren van dynamische systemen. Naar een didactiek voor het voortgezet onderwijs*.

60. Hammann, M., Waarlo, A. J., & Boersma, K. Th. (Eds.) (2010). *The nature of research in biological education: Old and new perspectives on theoretical*

and methodological issues – A selection of papers presented at the VIIth Conference of European Researchers in Didactics of Biology.

59. Van Nes, F. (2009). *Young children's spatial structuring ability and emerging number sense.*

58. Engelbarts, M. (2009). *Op weg naar een didactiek voor natuurkunde-experimenten op afstand. Ontwerp en evaluatie van een via internet uitvoerbaar experiment voor leerlingen uit het voortgezet onderwijs*.

57. Buijs, K. (2008). *Leren vermenigvuldigen met meercijferige getallen*.

56. Westra, R. H. V. (2008). *Learning and teaching ecosystem behaviour in secondary education: Systems thinking and modelling in authentic practices.*

55. Hovinga, D. (2007). *Ont-dekken en toe-dekken: Leren over de veelvormige relatie van mensen met natuur in NME-leertrajecten duurzame ontwikkeling.*

54. Westra, A. S. (2006). *A new approach to teaching and learning mechanics*.

53. Van Berkel, B. (2005). *The structure of school chemistry: A quest for conditions for escape.*

52. Westbroek, H. B. (2005). *Characteristics of meaningful chemistry education: The case of water quality*.

51. Doorman, L. M. (2005). *Modelling motion: from trace graphs to instantaneous change*.

50. Bakker, A. (2004). *Design research in statistics education: on symbolizing and computer tools*.

49. Verhoeff, R. P. (2003). *Towards systems thinking in cell biology education*.

48. Drijvers, P. (2003). *Learning algebra in a computer algebra environment. Design research on the understanding of the concept of parameter*.

47. Van den Boer, C. (2003). *Een zoektocht naar verklaringen voor achterblijvende prestaties van allochtone leerlingen in het wiskundeonderwijs.*

46. Boerwinkel, D. J. (2003). *Het vormfunctieperspectief als leerdoel van natuuronderwijs. Leren kijken door de ontwerpersbril.*

45. Keijzer, R. (2003). *Teaching formal mathematics in primary education. Fraction learning as mathematising process*.

44. Smits, Th. J. M. (2003). *Werken aan kwaliteitsverbetering van leerlingonderzoek: Een studie naar de ontwikkeling en het resultaat van een scholing voor docenten.*

43. Knippels, M. C. P. J. (2002). *Coping with the abstract and complex nature of genetics in biology education – The yo-yo learning and teaching strategy.*

42. Dressler, M. (2002). *Education in Israel on collaborative management of shared water resources.*

41. Van Amerom, B.A. (2002). *Reinvention of early algebra: Developmental research on the transition from arithmetic to algebra*.

40. Van Groenestijn, M. (2002). *A gateway to numeracy. A study of numeracy in adult basic education.*

39.   Menne, J. J. M. (2001). *Met sprongen vooruit: een productief oefenprogramma voor zwakke rekenaars in het getallengebied tot 100 – een onderwijsexperiment*.

38.   De Jong, O., Savelsbergh, E.R., & Alblas, A. (2001). *Teaching for scientific literacy: context, competency, and curriculum.*

37.   Kortland, J. (2001). *A problem-posing approach to teaching decision making about the waste issue.*

36.   Lijmbach, S., Broens, M., & Hovinga, D. (2000). *Duurzaamheid als leergebied; conceptuele analyse en educatieve uitwerking.*

35.   Margadant-van Arcken, M., & Van den Berg, C. (2000). *Natuur in pluralistisch perspectief – Theoretisch kader en voorbeeldlesmateriaal voor het omgaan met een veelheid aan natuurbeelden.*

34.   Janssen, F. J. J. M. (1999). *Ontwerpend leren in het biologieonderwijs. Uitgewerkt en beproefd voor immunologie in het voortgezet onderwijs*.

33.   De Moor, E. W. A. (1999). *Van vormleer naar realistische meetkunde Een historisch-didactisch onderzoek van het meetkundeonderwijs aan kinderen van vier tot veertien jaar in Nederland gedurende de negentiende en twintigste eeuw.*

32.   Van den Heuvel-Panhuizen, M., & Vermeer, H. J. (1999). *Verschillen tussen meisjes en jongens bij het vak rekenen-wiskunde op de basisschool – Eindrapport MOOJ-onderzoek.*

31.   Beeftink, C. (2000). *Met het oog op integratie – Een studie over integratie van leerstof uit de natuurwetenschappelijke vakken in de tweede fase van het voortgezet onderwijs*.

30.   Vollebregt, M. J. (1998). *A problem posing approach to teaching an initial particle model.*

29.   Klein, A. S. (1998). *Flexibilization of mental arithmetics strategies on a different knowledge base – The empty number line in a realistic versus gradual program design.*

28.   Genseberger, R. (1997). *Interessegeoriënteerd natuur- en scheikundeonderwijs – Een studie naar onderwijsontwikkeling op de Open Schoolgemeenschap Bijlmer*.

27.   Kaper, W. H. (1997). *Thermodynamica leren onderwijzen*.

26.   Gravemeijer, K. (1997). *The role of context and models in the development of mathematical strategies and procedures*.

25.   Acampo, J. J. C. (1997). *Teaching electrochemical cells – A study on teachers' conceptions and teaching problems in secondary education*.

24.   Reygel, P. C. F. (1997). *Het thema 'reproductie' in het schoolvak biologie*.

23.   Roebertsen, H. (1996). *Integratie en toepassing van biologische kennis– Ontwikkeling en onderzoek van een curriculum rond het thema 'Lichaamsprocessen en Vergift'.*

22. Lijnse, P. L., & Wubbels, T. (1996). *Over natuurkundedidactiek, curriculumontwikkeling en lerarenopleiding*.

21. Buddingh', J. (1997). *Regulatie en homeostase als onderwijsthema: een biologie-didactisch onderzoek*.

20. Van Hoeve-Brouwer G. M. (1996). *Teaching structures in chemistry – An educational structure for chemical bonding*.

19. Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education.*

18. Klaassen, C. W. J. M. (1995). *A problem-posing approach to teaching the topic of radioactivity.*

17. De Jong, O., Van Roon, P. H., & De Vos, W. (1995). *Perspectives on research in chemical education.*

16. Van Keulen, H. (1995). *Making sense – Simulation-of-research in organic chemistry education.*

15. Doorman, L. M., Drijvers, P. & Kindt, M. (1994). *De grafische rekenmachine in het wiskundeonderwijs.*

14. Gravemeijer, K. (1994). *Realistic mathematics education*.

13. Lijnse, P. L. (Ed.) (1993). *European research in science education*.

12. Zuidema, J., & Van der Gaag, L. (1993). *De volgende opgave van de computer.*

11. Gravemeijer, K., Van den Heuvel-Panhuizen, M., Van Donselaar, G., Ruesink, N., Streefland, L., Vermeulen, W., Te Woerd, E., & Van der Ploeg, D. (1993). *Methoden in het reken-wiskundeonderwijs, een rijke context voor vergelijkend onderzoek.*

10. Van der Valk, A. E. (1992). *Ontwikkeling in Energieonderwijs*.

9. Streefland, L. (Ed.) (1991). *Realistic mathematics education in primary schools.*

8. Van Galen, F., Dolk, M., Feijs, E., & Jonker, V. (1991). *Interactieve video in de nascholing reken-wiskunde.*

7. Elzenga, H. E. (1991). *Kwaliteit van kwantiteit*.

6. Lijnse, P. L., Licht, P., De Vos, W., & Waarlo, A. J. (Eds.) (1990). *Relating macroscopic phenomena to microscopic particles: a central problem in secondary science education.*

5. Van Driel, J. H. (1990). *Betrokken bij evenwicht*.

4. Vogelezang, M. J. (1990). *Een onverdeelbare eenheid*.

3. Wierstra, R. F. A. (1990). *Natuurkunde-onderwijs tussen leefwereld en vakstructuur.*

2. Eijkelhof, H. M. C. (1990). *Radiation and risk in physics education*.

1. Lijnse, P. L., & De Vos, W. (Eds.) (1990). *Didactiek in perspectief*.

Many high school students are unable to draw justified conclusions from statistical data in histograms. A literature review revealed various misinterpretations. Current statistics education often falls short of preventing these. In preparation for new instructional materials, several studies were conducted to better understand where these misinterpretations come from. Five solution strategies were found through qualitative analysis of students' eye movements on histogram and case-value plot tasks. Quantitative analysis of some tasks using a mathematical model and a machine learning model confirmed the results of the qualitative analysis which implied that the strategies could be identified reliably and automatically. Literature suggested that lesson materials with dotplot tasks can support students to correctly interpret histograms. An analysis of students' eye movements on histogram tasks before and after dotplot tasks suggested that students improved their strategies but not their answers. Based on the literature and eye-tracking studies, we conjectured that students most likely lacked embodied experiences with the actions required to construct histograms. Inspired by ideas of embodied instrumentation, we designed and tested instructional materials that provide starting points for scaling up. Together, the studies contribute to theorizing about teaching histograms and the use in statistics education of eye-tracking research, quantitative methods from data science, and instructional materials designed from the perspective of embodied instrumentation.