



Marianne van Dijke-Droogers

Introducing Statistical Inference: Design and Evaluation of a Learning Trajectory

Introducing Statistical Inference: Design and Evaluation of a Learning Trajectory

M.J.S. van Dijke-Droogers

Review committee:

Prof. dr. R. Biehler

Dr. K. Dajani

Prof. dr. M.J. Goedhart

Prof. dr. W.R. van Joolingen

Prof. dr. J.W.F. van Tartwijk

M.J.S. van Dijke-Droogers

Introducing Statistical Inference: Design and Evaluation of a Learning Trajectory/ M.J.S. van Dijke-Droogers – Utrecht: Freudenthal Institute, Faculty of Science, Utrecht University / FI Scientific Library (formerly published as CD-β Scientific Library), no.109, 2021.

Dissertation Utrecht University. With references. Met een samenvatting in het Nederlands.

ISBN: 978-90-70786-49-6

Cover design: Vormgeving Faculteit Bètawetenschappen

Printed by: Xerox, Utrecht

© 2021 M.J.S. van Dijke-Droogers, Utrecht, the Netherlands

Introducing Statistical Inference: Design and Evaluation of a Learning Trajectory

**Het introduceren van statistische inferentie:
Ontwerp en evaluatie van een leertraject**

(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 30 juni 2021 des middags te 4.15 uur

door

Maria Johanna Sophia van Dijke-Droogers

geboren op 18 april 1975

te Tholen

Promotor:

Prof. dr. P.H.M. Drijvers

Copromotor:

Dr. A. Bakker

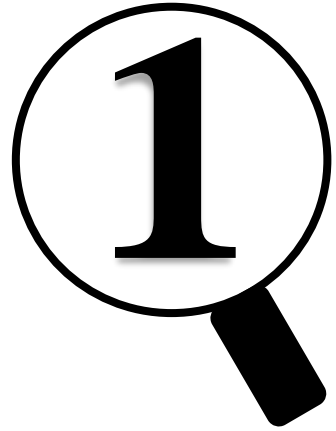
Acknowledgement

This research received funding from the Ministry of Education, Culture and Science under the Dudoc program.

Permission to publish the photos in this thesis has been granted by the persons shown.

Table of Contents

Chapter 1	Introduction	7
Chapter 2	Repeated sampling with a black box to make informal statistical inference accessible	15
	Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2020). <i>Mathematical Thinking and Learning</i> , 22(2), 116–138.	
Chapter 3	Statistical modeling processes through the lens of instrumental genesis	53
	Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2021). <i>Educational Studies in Mathematics</i> .	
Chapter 4	Introducing statistical inference: Design of a theoretically and empirically based learning trajectory	93
	Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted).	
Chapter 5	Effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy	125
	Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted).	
Chapter 6	General Discussion	155
	References	175
	Supplementary Materials	185
	Samenvatting (summary in Dutch)	219
	Dankwoord (acknowledgements in Dutch)	241
	Curriculum Vitae	245
	Publications related to this thesis	246
	Presentations related to this thesis	247
	FI Scientific Library	249



Introduction

*This is a time for looking at the data and saying let's do what makes the
most sense*
(New York Times, 24 April 2020)¹

Introduction

Overwhelming amounts of data, statistics and predictions about the current COVID19 situation were presented to society last year. This illustrates the growing importance of teaching statistics and probability in the classroom, to help students develop the statistical literacy needed to understand claims provided in our data-based society (Watson & Callingham, 2020).

Research Topic

Statistical literacy is considered one of the 21st-century skills that students should acquire. Gal (2002) defines statistical literacy as the ability to interpret, critically evaluate and reason with statistical information. Statistical inference is at the heart of statistics as “it provides a means to make substantive evidence-based claims under uncertainty when only partial data are available” (Makar & Rubin, 2018, p. 262).

Learning inferences is difficult for students, and therefore in most countries, including the Netherlands, not taught until Grade 10 or higher. Many difficulties of students are caused by a limited understanding of key statistical concepts required for inferences (Castro Soto et al., 2007; Konold & Pollatsek, 2002). An emphasis on complex formal procedures in Grades 10 to 12 and higher education, exacerbates students' conceptual problems. To help students overcome these difficulties, *informal* approaches have been sought in recent decades. Engaging in activities that involve informal inferences in the early years might facilitate learning about more complex inferential statistics later on (Zieffler, Garfield, delMas, & Reading, 2008). Makar and Rubin (2009) define informal statistical inference in terms of three main principles: generalization beyond data, data as evidence for these generalizations, and probabilistic reasoning about the generalization. In an informal approach, familiar experiences are incorporated into inferential processes to facilitate the understanding of statistical concepts required. Recently developed digital tools provide opportunities to deepen students' conceptual understanding.

¹ Quote by Dr. Peter Collignon, a physician and professor of microbiology at the Australian National University who has worked for the World Health Organization (Cave, 2020). Vanquish the Virus? Australia and New Zealand Aim to Show the Way - The New York Times (nytimes.com)

The increasing use of digital technology in today's society requires an educational shift towards learning from and with digital tools. This is particularly urgent for statistics education, where digital technology is indispensable for interpreting statistical information (Gal, 2002; Thijs, Fisser, & Van der Hoeven, 2014). Insight into underlying statistical models is fundamental for such interpretations (Manor & Ben-Zvi, 2017), and in particular for making inferences. Digital environments, such as VUstat and TinkerPlots, offer an informal approach to deepen students' understanding of statistical modeling and models (Biehler, Frischemeier, & Podworny, 2017). Within these environments, students can build a model of a given situation for simulating samples, which enables them to informally investigate the behavior of the model. By visualizing sample and sampling distributions—at varying sample sizes and at varying numbers of repeated samples—students can explore sampling variability, (un)likely sample results, and uncertainty involved in inferences. During these modeling activities, key concepts for inferences are visualized, explored and deepened. As such, modeling with digital tools seems promising for introducing statistical inference. Figure 1.1 shows an example of statistical modeling in TinkerPlots, in the context of a black box. A black box filled with 1,000 marbles, 750 yellow and 250 orange, is modeled in the bar graph top left. A simulated sample size 40, and the sampling distribution for repeated samples are visualized on the right.

Given the importance of and difficulties in teaching statistical inference, knowledge about efficient learning trajectories for secondary school students is needed. Embedding informal inferential activities in earlier years seems promising, in particular when combined with learning from and with digital tools. However, little is known about how statistics curricula with a descriptive focus can be transformed to a more inferential focus, to anticipate subsequent steps in students' statistics education. More knowledge is needed about well-substantiated learning trajectories. This is especially important for students in the pre-university stream (VWO is the Dutch abbreviation), the 15% best achieving students of our educational system, for whom statistical knowledge is essential in preparing for higher education. The aim of this research project is to gain knowledge about a theoretically and empirically based learning trajectory to introduce 9th-grade students to statistical inference. The guiding research question is:

How can a theoretically and empirically based learning trajectory introduce 9th-grade students to statistical inference?

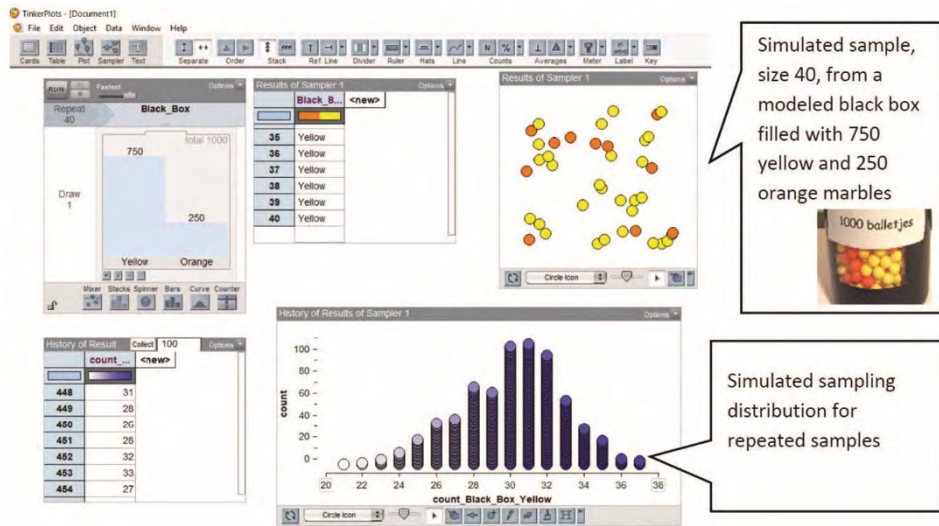


Figure 1.1. Illustration of the digital environment of TinkerPlots

Research Methods

As educational resources and teaching materials in which 9th-grade students are introduced to statistical inferences hardly exist, the formulated research question involves a dual question. Answering the question requires both the design and the evaluation of the learning trajectory. A design-based research method (Bakker, 2018) seems to address this duality. According to Euler (2017), a design-based research begins with the following question: How can an intended, initially vaguely stated, goal be achieved with a yet-to-be-developed design? As the research process progresses, interventions are conducted and evaluated. Design-based research is characterized by a cyclical process in which educational materials for learning environments are designed, implemented, and evaluated, for following cycle(s) of (re)design and testing (McKenney & Reeves, 2012). In this research project, three cycles were completed, starting from a one-class teaching experiment, through an intervention in three classes, to implementing the learning trajectory in thirteen classes at different schools. Furthermore, between cycles 2 and 3, a case study was conducted into learning from and with technology. In particular, this domain-specific case study focused on the intertwined development of learning techniques for using a digital tool and conceptual understanding. Figure 1.2 provides an overview of the cycles and studies in this research project, and the chapters of the thesis.

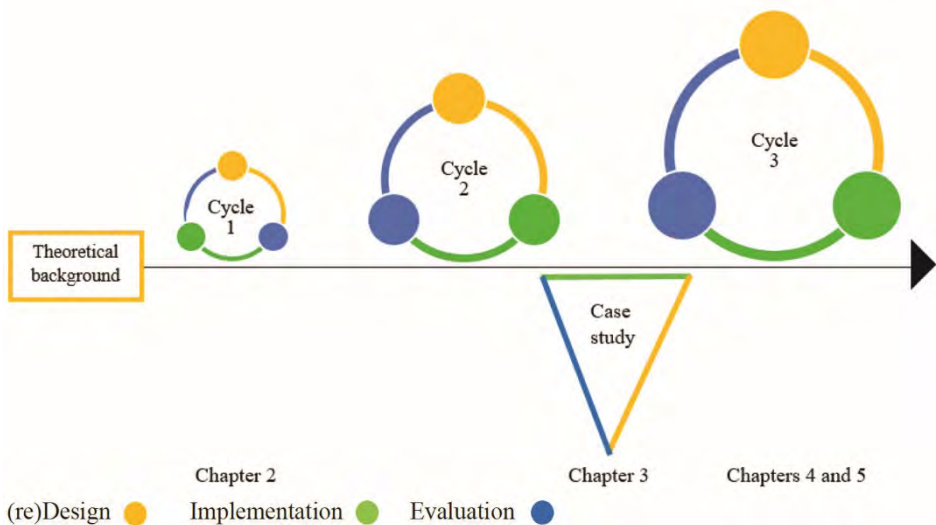


Figure 1.2. Overview of the phases in the research project and the chapters of the thesis.

As a design and research instrument to structure and connect the elements involved in a learning trajectory, we used a Hypothetical Learning Trajectory (HLT). According to Simon (1995), who introduced this notion, and Simon and Tzur (2004), an HLT consists of a learning goal for students, a description of promoting activities that will be used to achieve these goals, and hypotheses about the students' learning processes. Based on a literature study, an HLT was developed and implemented during the first research cycle. In following cycles, the initial HLT was (re)designed, implemented and tested, to develop an efficient trajectory.

Research Overview

Chapters 2 to 5 present the results from research cycles 1 and 3, and the findings from the case study. Results obtained from cycle 2 are not elaborated in this thesis, to reduce overlap between chapters. Insights from cycle 2 were used for (re)design in cycle 3. We report in Chapter 2 on the first three steps of the trajectory and in Chapter 4 on the whole trajectory, respectively. The case study is presented in Chapter 3, and Chapter 5 reports on a quantitative evaluation of the whole designed trajectory. We now elaborate on how these four chapters align with the aim of this research project: the design of a theoretically and empirically based learning trajectory for introducing statistical inference.

Chapter 2 presents the results of the first cycle that focused on the design, implementation and evaluation of the first part of a learning trajectory for introducing 9th-grade students to statistical inference. Twenty Grade-9 students (14–15 years old) took part in the learning trajectory. In the first three steps of the trajectory, ideas of repeated sampling with a black box and statistical modeling were embedded, to introduce students to key concepts of inferences. In particular, this study addressed the following research question:

RQ1: How can repeated sampling with a black box introduce 9th-grade students to the concepts of sample, frequency distribution, and simulated sampling distribution?

Chapter 3 presents the results of the domain-specific case study. This study examined 9th-grade students' intertwined development of techniques for using TinkerPlots and conceptual understanding of statistical modeling, by using the theoretical perspective of instrumental genesis. In this study, we addressed the following question:

RQ2: Which instrumentation schemes do 9th-grade students develop through statistical modeling processes with TinkerPlots and how do emerging techniques and conceptual understanding intertwine in these schemes?

Chapter 4 reports on the results of the third research cycle on the design, implementation and evaluation of the whole 8-step learning trajectory. In this study, the designed learning trajectory was empirically substantiated by analyzing students' progression during a large-scale intervention. The aim was to evaluate how the eight steps of the trajectory fostered students' learning processes and proficiency in statistical inference. We addressed the following research questions:

RQ3.1: What are the specific effects of the designed Learning Trajectory (LT) on students' understanding of statistical inference, in terms of the intended LT-step related learning goals?

RQ3.2: How do the designed steps of the learning trajectory foster students' learning processes?

Chapter 5 presents the results of a quantitative study on the effects of the learning trajectory on students' proficiency in the domains of statistical literacy, and inferences in particular. Although the designed learning trajectory

concentrated on statistical inference—the SI domain within statistical literacy—we conjectured that a focus on more complex learning activities for statistical inference would also have a positive effect on students’ understanding of other domains of statistical literacy. In this study, we addressed the following research question:

RQ4: What are the effects of a learning trajectory for statistical inference on 9th-grade students’ statistical literacy?

Chapter 6 presents the general conclusions. Here, the main findings of the four studies are summarized, aggregated, and discussed. The contribution of the research project is elaborated, including implications for future research and educational design.



Repeated Sampling with a Black Box to Make Informal Statistical Inference Accessible

This chapter is based on

Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2020).
Repeated sampling with a black box to make informal statistical inference
accessible. *Mathematical Thinking and Learning*, 22(2), 116–138.

Abstract

While various studies suggest that informal statistical inference (ISI) can be developed by young students, more research is needed to translate this claim into a well-founded learning trajectory (LT). As a contribution, this paper presents the results of a cycle of design research that focuses on the design, implementation and evaluation of the first part of a LT for ISI, in which 9th-grade students ($N = 20$) are introduced to the key concepts of sample, frequency distribution and simulated sampling distribution. The results show that a LT starting from repeated sampling with a black box may support the accessibility of these concepts, as these students were able to make inferences with the frequency distribution from repeated samples as well as with corresponding simulated sampling distributions. This suggests a promising way to make ISI more accessible for students.

Keywords

design research, informal statistical inference, learning trajectory, repeated sampling, statistics education.

Introduction

Drawing inferences about an unknown population is at the heart of statistics, and therefore important to learn. Sample data are commonly used to reason about a larger whole. For informed citizenship in a society in which data play an increasingly important role, reasoning with statistical information is essential (Gal, 2002). As such, statistical reasoning is considered as one of the 21st century skills that students should acquire (Thijs, Fisser, & Van der Hoeven, 2014).

However, learning and applying statistical inference is difficult for students. The emphasis on formal and procedural knowledge results in the inability of students to interpret the results (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007) or to understand statistical concepts such as sampling, variation and uncertainty (Konold & Pollatsek, 2002).

Recent research suggests that studying informal statistical inference (ISI) at an early age may facilitate the later transition to formal procedures (Zieffler, Garfield, delMas, & Reading, 2008). In general, ISI focuses on ways in which students without knowledge of formal statistical techniques, such as hypothesis testing, use their statistical knowledge to support their inferences about an unknown population based on observed samples. Although ISI has several definitions, the commonly used framework from Makar and Rubin (2009) identifies three key principles: generalization beyond data; data as evidence for these generalizations; and probabilistic reasoning about the generalization. At an informal level, familiar experiences can be used for making such inferences. By turning common predictions and expectations into inference processes, interpreting and understanding statistical concepts become more accessible (Paparistodemou & Meletiou-Mavrotheris, 2008).

Although various studies have shown that ISI can be developed in young students (Ben-Zvi, 2006; Doerr, delMas, & Makar, 2017; Makar, 2016; Meletiou-Mavrotheris & Paparistodemou, 2015) more research is needed to translate these promising results into compact theoretically underpinned learning trajectories in which students are introduced to sampling and the associated probability component in a short period of time. In particular, it is important to investigate *how* students can learn to draw informal inferences and what learning steps are needed to develop this ability among young students, as well as *which* learning activities may foster these learning steps. In most countries, school curricula for grades 7–9 focus on descriptive statistics (Ben-Zvi, Bakker, & Makar, 2015) and, as a result, pay little attention to informal

statistical inference. This also holds for the Dutch curriculum, in which statistics education progresses from descriptive statistics in the early years, to preparing for a more formal approach to inferential statistics from grade 10 and in higher education (Van Streun & Van de Giessen, 2007). As time in educational practice is limited, both in the Netherlands and abroad, we aim for a concise approach.

This research focuses on the question of how to provide students with opportunities to learn to draw conclusions about a population based on samples. To provide an answer to such a how-question we look for an idea of how such a learning goal can be achieved (design), to implement this idea, and to find evidence that the learning goal was indeed achieved. Hence, the aim of the research reported here is to design, implement, and evaluate the first part of a learning trajectory (LT) for 9th-grade students, that focuses on informal inferential reasoning and three statistical key concepts of sample, frequency distribution, and simulated sampling distribution.

Theoretical Background

To set up the study's theoretical background, we now elaborate the role of informal inferential reasoning and the key statistical concepts to enhance ISI.

The Role of Inferential Reasoning to Enhance ISI

This research focuses on inferential reasoning underpinning interpretations of sample data. In contrast to descriptive statistics, which concerns describing the data under investigation, inferential reasoning includes handling sampling variation and uncertainty. An inferential statement is fairly meaningless without the reasoning in which it must be embedded (Makar, Bakker, & Ben-Zvi, 2011). Therefore, an inference should be accompanied by reasoning based on the data. Following Zieffler et al. (2008), we consider *informal inferential reasoning* as making inferences about unknown populations based on observed samples without using formal techniques such as hypothesis testing using probability distributions. Informal inferential reasoning is about drawing on, utilizing, and integrating knowledge from meaningful experiences, as decisions are commonly made on the basis of predictions and estimates. These experiences can be used to make statistical concepts accessible. Informal inferential reasoning may include foundational statistical concepts, such as the notion that a sample may be surprising given a particular claim and the use of statistical language.

Key Statistical Concepts for ISI

Informal inferential reasoning, and the use of statistical concepts to seek evidence for interpretations of data, can be developed through various experiences with data over time (Makar et al., 2011). The question is which statistical concepts are important for 9th-grade students who are inexperienced with sampling. From the literature, three concepts appear to be central: (1) sample (including ideas of sampling variation, sample size, and repeated samples), (2) frequency distribution of data obtained from repeated sampling, and (3) simulated sampling distribution. Figure 2.1 provides an overview of the three central aspects and the build-up in handling variation and uncertainty: from the introduction to variation in observed samples, by visualizing variation within a frequency distribution, towards interpreting variation and uncertainty of samples with the simulated sampling distribution, which we elaborate on below.

First, inferential reasoning involves understanding the concept of *sample*. However, students are often reported to have conceptual problems with samples. On the one hand, students may assume every sample to be different and are therefore hesitant to draw conclusions about a population (Ben-Zvi, Aridor, Makar, & Bakker, 2012). On the other hand, students may consider a sample as a mini-population with the same characteristics as the underlying population and, as a consequence, students expect a small sample size to be a good reflection of the underlying population (Tversky & Kahneman, 1971).

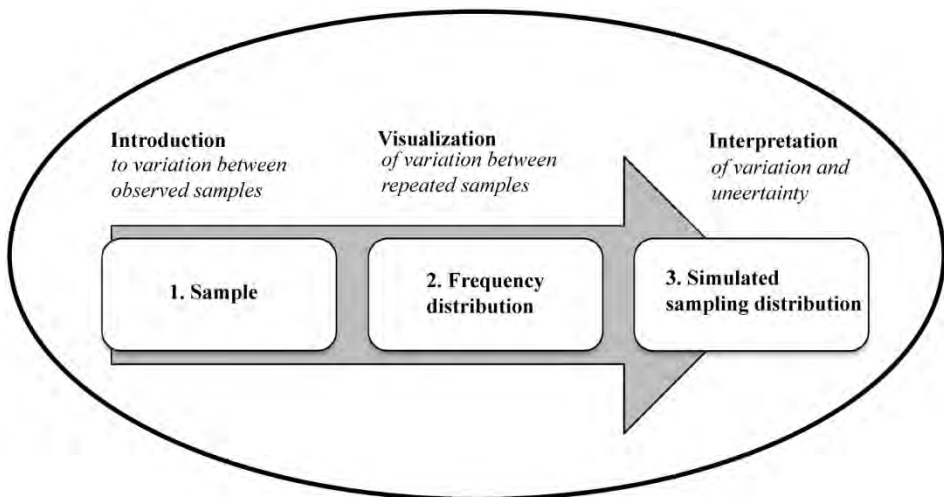


Figure 2.1. Overview of key concepts for ISI and the connection with handling variation and uncertainty

This is confirmed by Innabi and El Sheikh (2007), who showed that students in grade 11 did not take the sample size into account when interpreting sample results. Students often make grand statements based on small samples and are insufficiently aware of the role of sample size. Experimenting with repeated sampling enables students to become aware of the variation and uncertainty of a sample, but also of the “representativeness” of a sample (Saldanha & Thompson, 2002). Through repeated sampling, students are confronted with both variation and similarities, which allows them to gain insight into the variation versus stability of particular characteristics. Wild, Pfannkuch, Regan and Horton (2011) invited students to experiment with various sample results from a given population, with variation in sample size and sample repetitions, to raise awareness and understanding of sampling variation. In this respect, Wild and Pfannkuch (1999) emphasized the importance of the exchange and comparison of sample results. Although repeated samples vary, some sample results are more likely than others. Thinking about the question: “What happens if a sample is repeated?” contributes to getting a grip on variation and uncertainty. This “what if” question is paramount in understanding statistical inference (Rossman, 2008).

With respect to the second key concept, a graph of the *frequency distribution from repeated samples* allows for visualization of obtained results and gives an overview of variation and stability among samples. A sample leads to data, a dataset has particular characteristics (such as proportion), and these characteristics are compiled in the frequency distribution of results from repeated sampling. The horizontal axis of a graph of such a distribution contains the values of the dataset characteristic. The vertical axis shows the number of samples for which each value occurred. A graph of the frequency distribution from repeated samples can be made manually or by using a computer and gives insight into (un)likely sample results. As such, the graph of the frequency distribution functions as a model of obtained results from repeated sampling and can be used to display sampling variation and, as a next step, to further investigate variation and uncertainty.

As a third key concept, the *simulated sampling distribution* can be utilized as a model for making statements about variation and uncertainty. If used as intended, the simulated sampling distribution extends the concept of frequency distribution from step 2, that functioned as a model (or visualization) of obtained sample results. Experts in statistics may think of these as conceptually the same (cf. Sfard & Lavie, 2005), but from a learning perspective there may still be a developmental transition from a model of into a

model *for* (Gravemeijer, 1999). The switch from a visualization of specific datasets to a more abstract simulation model for interpreting variation and uncertainty, we assume, enables emergent statistical reasoning. Such a sampling distribution, based on a large number of simulations, can easily be made with computer software (such as TinkerPlots). This simulated sampling distribution can be used to determine informally the probability of certain sample results (Rossman, 2008; Watson & Chance, 2012) and can assist in determining whether a sample result is likely (Garfield, Ben-Zvi, Le, & Zieffler, 2015; Manor & Ben-Zvi, 2015; Pfannkuch, Ben-Zvi, & Budgett, 2018; Watson & Chance, 2012). Reasoning with the sampling distribution from repeated sampling is a meaningful preparation for the more formal reasoning with the theoretical sampling distribution in higher education (Garfield et al., 2015; Watson & Chance, 2012).

The black box activity in research of Van Dijke-Droogers, Drijvers and Bakker (2018), seemed a promising way to introduce students to the key statistical concepts of ISI. Here, students investigated the content of a black box filled with marbles by gathering, exchanging and comparing results from physical and later simulated samples with different sizes and different number of repetitions.

Research Question

Given the importance of informal inferential reasoning and the corresponding key concepts in enhancing ISI, the main question of this research is:

How can repeated sampling with a black box introduce 9th-grade students to the concepts of sample, frequency distribution, and simulated sampling distribution?

Methods

Over the past ten years, research has increasingly focused on informal statistical inference and has developed various educational materials for young students (Doerr et al., 2017; Meletiou-Mavrotheris & Paparistodemou, 2015). However, educational resources and teaching materials in which 9th-grade students are introduced to concepts of sample, frequency distribution and simulated sampling distribution in a short period of time, hardly exist. Therefore, this research required a design research approach. Design research is characterized by a cyclical process in which educational materials for learning environments are designed, implemented, and evaluated, for following cycle(s) of (re)design and testing (McKenney & Reeves, 2012). The research reported here comprised

a first cycle of design, implementation, and evaluation of a LT for ISI. We focused on the first steps, as part of a longer LT. We designed a hypothetical learning trajectory (HLT) of eight steps to map out and structure all elements involved in the learning and teaching approach, and to make explicit the expectations about how these elements function in interaction to promote learning. The LT was implemented and tested during a classroom intervention. Here, we report on the design, implementation, and evaluation of the first three steps and indicate how the results of these steps were used for revision.

HLT as a Design Research Instrument

As a design and research instrument to structure and connect the elements involved in an LT, we used a hypothetical learning trajectory (HLT). According to Simon (1995), who introduced this notion, and Simon and Tzur (2004), an HLT consists of a learning goal for students, a description of promoting activities that will be used to achieve these goals, and hypotheses about the students' learning process. It includes the simultaneous consideration of mathematical goals, student thinking models, teacher and researcher models of students' thinking, sequences of teaching tasks, and their interaction at a detailed level of analysis of processes (Clements & Sarama, 2004). Research by Gravemeijer, Bowers, and Stephan (2003) showed how an HLT can be used to bridge the gap between students' ideas and solutions on the one hand and the teachers' mathematical goal on the other. In this way, an HLT can give guidance to anticipate the collective practices in which students get involved and the ways in which they reason with the various artifacts and activities. An HLT provides insight into how students learn and aims for a well-founded theory of the learning process. According to Sandoval (2014), the hypotheses (or conjectures, as he calls them) in educational design research are typically about how tools and materials, task structures, participant structures, and discursive practices lead to required mediating process and intended outcomes.

In this research, the HLT is used as a design and research instrument to empirically and theoretically connect all elements of the LT, including theoretical background, learning steps, teaching approach, lesson activities with tools and materials, practical guidelines for implementation, expected student behavior, and data collection, involved in the implementation of the LT. This report focuses on the role of the first three steps. Because our HLT was extensive, we restrict ourselves in this report to a concise description of theoretical background, activities designed, hypotheses and corresponding indicators of students' learning behavior, data collection, and implementation characteristics.

Educational Guidelines to Frame the HLT

Educational guidelines to promote inferential reasoning and key concepts, extracted from literature, formed the starting point of the HLT design. To promote inferential reasoning, an inquiry-based approach with meaningful contexts is recommended (Ainley, Pratt & Hansen, 2006; Ben-Zvi et al., 2012; Van Dijke-Droogers, Drijvers, & Tolboom, 2017; Franklin et al., 2007; Makar & Rubin, 2009; Pfannkuch, 2011). In particular, a holistic approach, in which a concrete investigation question is answered by going through all steps of statistical investigation from collecting to interpreting data, is expected to stimulate reasoning about generalizations, variation, and uncertainty. This approach was also addressed by Lehrer and English (2017), who recommended systematic and cohesive involvement of students in practices of inquiring, visualizing, and measuring variation instead of a piecewise approach. Rossman (2008) advised starting with categorical data, so that students can focus on the inferential process and only switch to more complex data later. Categorical data can be captured by means of a sample proportion, while summarizing numerical data requires the determination of measures of center and spread. In addition, the distribution of one sample with numerical data may lead to confusion with the sampling distribution.

To promote students' concepts of sample and sampling variation, Saldanha and Thompson (2002) advocated investigating repeated samples. Wild et al. (2011) advised an approach in which students experiment with sample size and repeated samples from a given population. In this respect, Wild and Pfannkuch (1999) suggested that students should exchange and compare their sample results. The use of growing samples can help students understand the effect of sample size and the relation between sample and population (Bakker, 2004). With the growing samples task design, students are introduced to increasing sample sizes that are taken from the same population. For each sample, they draw informal inferences based on their data. Subsequently, they predict what might change in a following larger sample. Students are required to search for and reason with variable processes and are encouraged to think about how certain they are about their inferences. This inquiry-based growing samples approach can help students enhance their inferential reasoning (Ben-Zvi et al., 2012).

As a next step, letting students think about the question: "What happens if a sample is repeated?" contributes to understanding of variation and uncertainty (Rossman, 2008). Additionally, making predictions, which are then tested, stimulates students' involvement and statistical reasoning (Bakker, 2004).

When working with a computer model for simulations, a strong connection with a meaningful experiment is preferred (Chance, Ben-Zvi, Garfield, & Medina 2007; Konold & Kazak, 2008; Manor & Ben-Zvi, 2015).

An Outline of the HLT

To design the HLT, the above educational guidelines were translated into hypotheses about students' learning. This three-step HLT—addressing the concepts of sample, frequency distribution, and simulated sampling distribution—is summarized in Table 2.1. The central column presents the hypothesis about how to promote students' understanding of each concept; the last column shows the connection with educational guidelines. The connection between the designed learning activities and the hypothesized students' learning processes is shown in Table 2.2. The upper part of this Table displays the features of each step. For each HLT step, Row 1 provides a brief description of the designed activity, Row 2 contains the key concepts, Row 3 indicates the type of expected inferential reasoning, and Row 4 describes the student activity.

For each step, a concise description is given of the designed activities, the corresponding hypothesis and indicators of students' learning behaviour that would support the hypothesis.

The first HLT step: How many yellow balls does the black box contain?

The first HLT step is carried out during Lesson 1 of the intervention. At the start of Lesson 1, the first task is to investigate the number of yellow balls in a black box filled with a mix of 1,000 yellow and orange balls, by looking through a small viewing window. The students shake up the box to mix the objects and estimate the content within the given time according to their own approach. Students note their findings on a student worksheet. Next, the sample results are exchanged and discussed in a whole-class discussion. Students repeat the experiment with a larger viewing window. Again, they note their findings on a worksheet. At the end of Lesson 1, they compare their estimates from a small and a large sample and make an inference about the effect of sample size.

Table 2.1. Overview of HLT Steps 1–3, including the Connection with Corresponding Educational Guidelines

HLT step	Hypothesis	Educational guidelines
1	<p>The first hypothesis concerning step 1 is that students will become aware of the concept of a sample by confrontation with sampling variation in a meaningful context with categorical data and will use repeated samples to estimate the population proportion.</p> <p>By exchanging and comparing sample results, students will understand that this estimate may not necessarily be exactly the same as the actual proportion of the underlying population.</p> <p>As a follow up, students will be introduced to the effect of sample size by using growing samples and will understand that usually a larger sample corresponds better to the underlying population.</p>	<ul style="list-style-type: none">• Use meaningful contexts (Ainley, et al., 2006; Ben-Zvi, et al., 2015; Dijke-Droogers, et al., 2017; Franklin et al., 2007; Makar & Rubin, 2009; Pfannkuch, 2011)• Start with categorical data (Rossman, 2008)• Offer repeated samples (Saldanha & Thompson, 2002)• Let students exchange and compare sample results (Wild & Pfannkuch, 1999)• Use the growing sample task design (Bakker, 2004; Ben-Zvi et al., 2012)• Ask what happens if this sample is repeated (Rossman, 2008) and let students make predictions (Bakker, 2004)• Reason with the frequency distribution on repeated sampling (Rossman, 2008; Watson & Chance, 2012)
2	<p>The second hypothesis concerning step 2 is that students make the conceptual switch from frequency distribution as a model of obtained results to a model for investigating variation and uncertainty, by imagining, visualizing and reasoning with the frequency distribution on repeated sampling, based on the question: ‘What happens if this sample is repeated?’.</p>	

In addition, students understand that most sample results will be close to the population proportion and that strong deviations are unlikely.


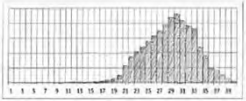
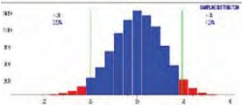
Furthermore, students will use the frequency distribution to determine the probability of certain sample results.

The third hypothesis in step 3 is that students will simulate, investigate and reason with the simulated sampling distribution from repeated sampling in order to interpret the variation and uncertainty involved.

In addition, that they understand that repeated sampling with a larger sample size reduces the variation in the accompanying estimates of the population and hence to a more certain inference, and that sampling with a larger number of repetitions leads to less variation in the mean and hence to a better estimate of the population.

- Work with simulations of many repeated samples to determine whether a sample result is likely (Garfield, et al., 2015; Manor & Ben-Zvi, 2015; Watson & Chance, 2012)
- Let students experiment with various sample results from a given population, with variation in sample size and sample repetitions (Wild, et al., 2011)
- Work with a computer model for simulations that has a strong connection with a concrete experiment (Chance, et al., 2007; Konold & Kazak, 2008; Manor & Ben-Zvi, 2015)

Table 2.2. Overview of Designed Teaching Activities for each HLT Step

	Step 1	Step 2	Step 3
Teaching activity	Conduct physical black box experiment (with small and large window)	Imagine the frequency-distribution at more than 100,000 repetitions of physical black box experiment	Simulate sampling distribution of physical black box experiment to interpret variation and uncertainty (with ICT)
Concept	Sample Sampling variation Repeated sampling Sample size	Frequency distribution on data from repeated sampling	Simulated sampling distribution from repeated sampling
Inferential Reasoning	In words with argumentation on: sampling variation, repeated sampling and sample size	In words with argumentation from the frequency distribution on (un)likely sample results	In words with argumentation from the simulated sampling distribution on variation and uncertainty
Student activity	Estimate the content of the black box 	Sketch the expected frequency distribution of data from >100,000 repetitions. Determine (un)likely sample results 	Inferential reasoning with the simulated sampling distribution 

The hypothesis in the first step, concerning the concept of sample, is that, in conducting the designed activity, students become aware of sampling variation and that they investigate the effect of repeated sampling and sample size. The following indicators are considered as supporting the hypothesis:

- 1a) Students note that sample results vary;
- 1b) Students choose a repeated sampling approach, with calculating the average, to estimate the number or proportion of yellow balls;
- 1c) Students note that it is possible to estimate the content based on samples;
- 1d) Students note that it is impossible to determine the exact content based on samples;
- 1e) Students note that the larger the sample size, the more confident they are about their estimate;
- 1f) Students note that working with a larger sample (usually) leads to a better estimate of the content.

The second HLT step: What happens if this experiment is repeated?

In Lesson 2, students are asked to think about the question “What happens if this experiment is repeated many times?” During a whole-class discussion, the students share their expectations for the number of yellow balls in a sample of 40 from a box consisting of 750 yellow and 250 non-yellow balls and discuss the boundaries of possible sample results. Subsequently, students are asked to sketch on their worksheet the expected frequency distribution if the experiment was repeated 100,000 times. The students are given a coordinate system with the values 0 to 40 along the horizontal axis and no values vertically. As a follow-up, students are asked to estimate the probability of ranges of particular sample results, based on their sketch of the frequency distribution, and to note this on their worksheet.

The hypothesis in the second step is that engaging in the designed activity prepares students to make the conceptual switch from using the frequency distribution as a visualization of (model of) results obtained from repeated sampling to using it as a model for interpreting variation and uncertainty. As such, it was expected that students would understand that most sample results will be close to the population proportion and strong deviations are unlikely, and that the frequency distribution can be used to determine the probability of ranges of particular sample results. The following indicators are considered as supporting the hypothesis:

- 2a) Students note that sample results corresponding to the population proportion will often occur;
- 2b) Students note that strongly deviating sample results are unlikely to appear;
- 2c) Students sketch a graph of the frequency distribution with a peak at the population proportion (in this case 30);
- 2d) Students sketch a graph of the frequency distribution in which the extreme

- values (in this case 0–10 or 35–40) hardly occur;
- 2e) Students estimate the probability of ranges of particular sample results on the basis of their sketched frequency distribution.

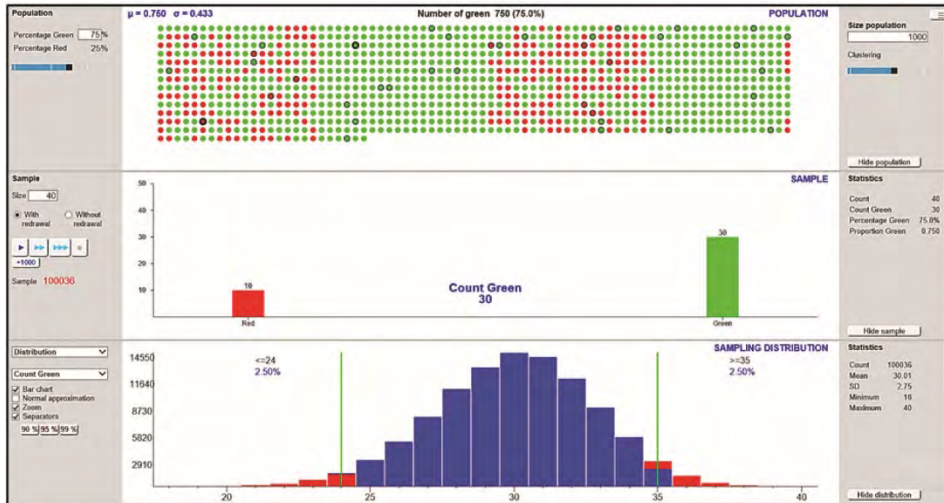


Figure 2.2. Screenshot of the VU Stat sampling distribution app (<https://www.vustat.eu/apps/yesno/index.html>)

The third HLT step: How can computer simulation help?

In Lesson 3, students are asked to simulate the experiment from Lesson 1 with a computer to investigate variation and uncertainty. To this end, they use a sampling distribution app from VU Stat. In this app, the population is displayed using colored balls, which creates a strong connection with the black box activity. In our view, the app seems user-friendly, with easy input of the population size, population proportion and sample size. As shown in Figure 2.2, the software provides a clear overview of the population in the upper screen, of each individual sample result in the middle screen, and of the sampling distribution for many repetitions in the lower screen. Students simulate the sampling distribution with a large number of repetitions and use this distribution as a model for investigating most common sample results with accompanying estimates of the population. Subsequently, we expect students to use the simulated sampling distributions from a given population at varying sample sizes and at varying numbers of repetitions as a model for investigating the effect of sample size and repeated samples on the accompanying estimates of the population. Students note their findings on a student worksheet.

The hypothesis in the third step, concerning the simulated sampling distribution, is that engaging in the designed activity makes students aware that the simulated sampling distribution can be used as a model *for* interpreting variation and uncertainty, and more particularly that repeated sampling with a larger sample size reduces the variation in the accompanying estimates of the population and hence leads to a more certain inference; and that sampling with a larger number of repetitions leads to less variation in the mean and hence to a better estimate of the population. The following indicators are considered as supporting the hypothesis:

- 3a) Students compare the simulated sampling distributions at varying sample sizes and note that repeated sampling with a larger sample size leads to less variation in the accompanying estimate of the population;
- 3b) Students compare the simulated sampling distributions at varying sample sizes and note that repeated sampling with a larger sample size leads to a better estimate of the population;
- 3c) Students compare the simulated sampling distributions from varying number of repetitions and note that from repeated sampling with a larger number of repetitions the mean of these samples is less variable;
- 3d) Students compare the simulated sampling distributions from varying number of repetitions and note that repeated sampling with a larger number of repetitions leads to a better estimate of the population;
- 3e) Students describe how the simulated sampling distribution from repeated sampling can be used to determine most common sample results.

Each HLT step focuses on one key concept, in which entailing aspects—for example sample size, repeated sampling, sampling variation, (un)certainty of an estimate, probability of samples, visualizations—are addressed from an exploratory perspective from the concrete black box in step 1 to a more abstract perspective in step 3 by simulating the sampling distribution by repeated sampling. An overview of the build-up in complexity by these aspects of ISI is displayed in Table 2.3.

Data Collection

With respect to the first step, data included individual student worksheets filled in by students who worked in pairs with the black box and video-recordings from a 10-minute whole-class discussion. In the second step, we collected data from student worksheets that were individually filled in by students and video-recordings from a 12-minute whole-class discussion. For the third step, we collected data from student worksheets that were individually filled in by

students who worked in pairs on a computer and eight video-recordings of 2-minute interactions between teacher and students.

Table 2.3. Overview of Indicators expected in Different Data Sources, by Entailing Aspects of ISI

	HLT step 1		HLT step 2		HLT step 3	
	Sample		Frequency distribution		Simulated sampling distribution	
Build-up in complexity	Introduction to variation and uncertainty		Visualization of variation		Interpretation of variation and uncertainty	
Data source	WD1 ¹	SW1 ²	WD2 ¹	SW2 ²	TSI3 ³	SW3 ²
Sample size	1e, 1f	1e, 1f				3a, 3b
Repeated sampling	1b	1b	2a, 2b	2c, 2d	3c, 3d	3c, 3d
Sampling variation	1a		2a, 2b	2c, 2d	3a, 3b, 3c, 3d	3a, 3b, 3c, 3d
(Un)certainty of estimate of population	1c, 1d	1c, 1d, 1e				3b, 3d
Probability of particular sample results			2a, 2b	2c, 2d, 2e	3a, 3b, 3c, 3d	3a, 3b, 3c, 3d
Determination of (un) likely sample results			2a, 2b	2d	3a, 3b, 3c, 3d	3a, 3b, 3c, 3d
Use of visualization				2c, 2d, 2e	3a, 3b, 3c, 3d	3a, 3b, 3c, 3d

¹ Whole-class Discussion (WD), ² Student Worksheet (SW), ³ Teacher-Student Interaction (TSI)

The video recordings were made by a research assistant and were preceded by detailed instruction on specific recording details. The worksheets of the students were distributed at the start of each lesson and collected at the end, to prevent information being added or lost.

Implementation Characteristics

To empirically verify whether the three hypotheses could be confirmed, we implemented this LT in one class at a secondary school in the Netherlands. In this report we focus on three 45-minute lessons, the first three of a more extensive series of ten lessons. We do so because this is where the students are introduced to the three key concepts by repeated sampling with the black box. The subsequent five steps of the LT, concerning seven more lessons, consist of applying these concepts in new situations with build-up in complexity of data.

Participants

The participants consisted of twenty 15-year-old students in Grade 9 of the pre-university level, who are among the 20% best performing students in the Dutch education system. The twenty students formed one class with both talented and less gifted mathematics students. The students were inexperienced with sampling. They had some basic knowledge of descriptive statistics: center and distribution measures, such as mean, quartiles, class division, absolute and relative frequencies, and boxplot. The lessons were conducted during the regular mathematics lessons over a period of one week.

The teacher was the first author. In this design research, it was an advantage that the teacher-researcher was so familiar with the designed materials. This allowed all attention to be focused on the design without deviations from the designer's intentions. Although there is added value in investigating field-based trials of an activity to see how teachers tend to implement the materials, at this stage, it is sensible to tackle challenges one by one (Tessmer, 1993).

Data Analysis

To answer the research question, we analyzed the data with respect to the indicators that would support the hypotheses as formulated in the HLT. The main data sources were the student worksheets and the video recordings of both the whole-class discussions and the teacher-student interactions. Table 2.3 displays the distribution of the expected occurrence of indicators in the different data sources by sub-area of ISI.

Answering the how-question of our research includes design, implementation, and evaluation. To make the connection between these phases explicit, we work out indicator 1f as an illustrative example. The other indicators were elaborated in a similar way. Indicator 1f states: “Students note that working with a larger sample (usually) leads to a better estimate of the content.” In the design phase we designed a student activity in which students collect data, analyze their data, and formulate an inference on the basis of a given investigation question. The activity concerns the context of a black box experiment and is built up according to the educational ideas of repeated and growing samples. The full HLT incorporated a detailed description of the designed activity, including all implementation issues involved. During the implementation phase in lesson 1, students worked on the designed activity in pairs and noted their data collection and data analysis, as well as their inference (estimate of the content) along with an indication of their (un)certainty, on their worksheet as an answer to tasks 1–3. In the following whole-class discussion, the results were exchanged and discussed. Subsequently the teacher posed the question: “What happens if we enlarge the viewing window?” Different options were exchanged and discussed, with attention for the uncertainty involved. After the whole-class discussion, students doubled the sample size (larger window) of their black box and again collected data, analyzed these and made a new inference with an indication of their (un)certainty, and noted the results on their worksheet as an answer to tasks 4–6. After that, students were asked in task 7 on the student worksheet to compare their answers for tasks 1–6 and draw a conclusion about the effect of sample size on the estimate of the content. During the analyses we used data from tasks 1–7 on the student worksheets and video data of the whole-class discussion. The data analysis of these sources is elaborated on below.

All video data, both whole-class discussions and teacher-student interactions, were transcribed and coded. The code book consisted of the indicators in Table 2.3. The unit of analysis during the discussions and interactions was a central question brought forward by the teacher to check out the indicators, and the corresponding reactions by the students. For example, a central question in the discussion of step 1 was “What do you know *for sure* about the number of yellow balls in the black box?” which refers to indicators 1c and 1f. To distinguish clear instances and less clear instances, the evidence was coded as strong, weak or no evidence. Strong evidence refers to indicators that were explicitly present during the class discussion or interaction, for example conclusions that were expressed literally or assumptions that were used

and repeated more than once. Weak evidence refers to indicators that were partly observed, for example incomplete conclusions or assumptions discussed indirectly. No evidence refers to indicators that were attended during the discussion but were not confirmed or contradicted.

Student worksheets were coded according to the same code book. The worksheets consisted of structured and open tasks. For the analysis, only open tasks were used in which students were explicitly asked to clearly motivate their answer. The worksheets contained specific tasks that were directly related to the indicators. For example, task 6 (“Are you sure of your estimate from a larger sample?”) and 7 (“What did you learn from a larger sample?”) on Worksheet 1 refer to indicators 1d and 1e, respectively. The frequency with which each indicator was coded was noted.

Indicators that were not attended to during the whole-class discussion or on the worksheet were indicated as “non-applicable.” A second coder was used to analyze the video data of the whole-class discussions and the teacher-student interactions, as well as the answers to open tasks on the worksheets. A random sample of 25% of the data was checked by the second coder. Cohen’s kappa was .83, indicating a good interrater reliability.

Results

For each hypothesis, this section describes whether the supporting indicators were observed.

First Step: Sample

The hypothesis in the first step, introducing the concept of a sample, was confirmed as the indicators 1a to 1f were coded in the data collected. Table 2.4 displays the observed indicators.

Table 2.4. Overview of Results for LT Step 1

Indicators	Student worksheet (N = 20) (observed number of students)	Video (strong, weak, no evidence)
1a. Students note that their sample results vary	Non-applicable	Strong
1b. Students choose a repeated sampling approach, with calculating the average, to estimate the number or proportion of yellow balls	n = 20	Strong

1c. Students note that it is possible to estimate the content based on samples	n = 20	Strong
1d. Students note that it is impossible to determine the exact content based on samples	n = 20	Strong
1e. Students note that the larger the sample size, the more confident they are on their estimate	n = 17	Strong
1f. Students note that working with a larger sample (usually) leads to a better estimate of the content	n = 20	Strong

The strategies for investigating the number of yellow balls in a black box with a small viewing window that students showed on their worksheet, corresponded to indicator 1a to 1d. Table 2.5 gives an overview of these results on Worksheet 1 on small samples from the black box.

As a first strategy, after shaking up the box to mix the objects, most students (14 out of 20) counted the visible yellow balls in the viewing window and extrapolated this number into the total content. They then repeated this shaking and counting five to ten times and used the average of these counts to estimate the content. This strategy was also expressed during the whole-class discussion, in which Ruben (all names are pseudonyms) added the following:

Teacher: How did you get the estimate?

Joerie: We counted the number of yellow balls and counted the total number of balls in the window. Then we converted the numbers into the total content. We repeated this about ten times and then calculated the average.

Teacher: Are there students with a different approach?

Ruben: Well, about the same thing, but we counted the orange balls, there are less of them, and then converted to the total. We repeated this about seven times and calculated the average.

Table 2.5. Answers on Worksheet 1 on Small Samples (size 20) from the Black Box (N = 20, number of students)

Task	Students' answers	Examples from written work
1. Estimate the number of yellow balls in the black box.	682, 750, 750, 625, 700, 730, 700, 735, 700, 725, 750, 675, 675, 730, 750, 725, 682, 733, 700, 625 (<i>estimate of each student</i>) (n = 20).	
2. Explain your estimate.	Approach 1: count balls, calculate average, convert to the contents of the entire box (n = 14). Approach 2: estimate the ratio of yellow balls after shaking a few times and convert to the entire box (n = 6).	We took ten samples with twenty balls, calculated the average and multiplied this by 50. Always around 15 – 16 yellow and the remaining orange.
3. How confident are you about your estimate?	Not confident (n=2). Quite confident (n=16). Most confident (n=2).	Not sure, just guessing. We don't know exactly, but it's about this number. Most confident, but not 100% sure, because we calculated the averages and extrapolated this number to the content.

As a second strategy, after shaking up the box, some students (6 out of 20) based their estimate not on counts, but on ratios. For example, one of these students indicated: “We have shaken the black box several times and there are always about 15–16 yellow balls and the remaining ones are orange.” All students decided to shake and measure several times, which showed that the students were confronted with sampling variation when estimating the content and opted for repeated sampling to get a better estimate. The students' estimates ranged from 625 to 750. Most students were quite confident about their estimate. Only two students indicated that it was a guess and two students were most confident—although not 100% sure—because their estimate was based on a calculation with the average from multiple counts.

As a follow-up activity, students worked with a larger viewing window. Most students were more confident about their estimate based on a larger window and all students noted that a larger window led to a better estimate of the content, which corresponded to indicators 1e and 1f. Table 2.6 provides an overview of these results with students' answers on Worksheet 1 with larger samples from the black box.

Table 2.6. Answers on Worksheet 1 with Large Samples (size 40) from the Black Box ($N = 20$)

Task	Students' answers	Examples from written work
4. Estimate the number of yellow balls in the black box.	744, 750, 714, 720, 720, 730, 720, 728, 725, 725, 750, 731, 731, 730, 714, 725, 744, 728, 725, 720 (<i>estimate of each student</i>) ($n = 20$).	
5. Explain your estimate.	Approach 1: count balls, calculate average, convert to the contents of the entire box ($n = 20$).	
6. Are you confident about your estimate from a larger sample?	More confident than before, with the small window ($n = 17$).	More sure because the estimates are now less variable.
	Still not confident ($n = 3$).	More sure because you have more information. Not sure yet because the results still vary.
7. What did you learn from a larger sample?	A larger sample size gives a better estimate ($n = 20$).	A larger sample gives more information about the content.

With this larger window, all students used the first strategy of counting the number of yellow balls several times and converting the average to the entire content. This time students' estimates showed less variation, as they ranged from 714 to 750. Most students (17 out of 20) wrote that they were more confident of their estimate based on this larger sample. Some students mentioned in this respect: "We are more confident because the estimates are now less variable" and others quoted: "More sure because you have more information." Three students wrote that they were not confident because the

sample results still varied. However, these three did indicate in the next task that the best way to estimate the content was to use a larger sample.

Although the students initially gave a numerical value as an estimate of the total number of balls, they later switched to an interval, which revealed indicators 1c and 1d. This transition became clearly visible during the discussion when the teacher explicitly asked: “What do you know *for sure* about the number of yellow balls in the black box?”

Daphne: Well, that three-quarters of the balls are yellow, and one-quarter are orange.

Teacher: Are you sure about the three-quarters?

Daphne: Yes, a bit more or less, because.... Yes, there are more yellow balls than orange balls.

Bas: I think the number of yellow balls is around, uhm, 700. It may be little less. In any case, it is between the 625 and 750.

Jesse: Yes, it is in any case between 600 and 800.

Here, Bas took the extreme values of the observed samples as limits for the possible number of yellow balls. Jesse took a broader interval. Both showed that they understood that sample results vary, but can be used to estimate the population. Jesse’s reply indicated that he understood that these extreme values were global estimators that might vary due to chance.

Second Step: Frequency Distribution

Regarding the second step, introducing the concept of frequency distribution, the hypothesis was confirmed. The results showed that indicators 2a to 2e were observed. Table 2.7 displays the observed indicators.

The whole-class discussion focused on the question “What happens if this experiment is repeated?” Students mentioned that results that resembled the population proportion were most likely to appear and that strong deviations were unlikely but possible, which confirmed the expected students’ behavior as described in 2a and 2b. However, it seemed that some students overestimated the possibility of strongly deviating results, as they suspected that with a large number of repetitions there would certainly be outliers. At the same time, students seemed to become aware of the difference between possibility and chance, which followed from the next interview fragment.

Teacher: What sample result is unlikely?
 Iris: Eh, that all the balls are orange.
 Bas: That is possible, though there are more than 40 balls in the box.
 Iris: Yes, but little chance that this will happen.

Table 2.7. Overview of Results for LT Step 2

Indicators	Student worksheet (N = 20) (observed number of students)	Video (strong, weak, no evidence)
2a. Students note that sample results corresponding to the population proportion will often occur;	Non-applicable	Strong
2b. Students note that strongly deviating sample results are unlikely to appear;	Non-applicable	Strong
2c. Students sketch a graph of the frequency distribution with a top at the population proportion (in this case 30);	n = 20	Non-applicable
2d. Students sketch a graph of the frequency distribution in which the extreme values (in this case 0–10 and 35–40) hardly occur;	n = 20	Non-applicable
2e. Students estimate the probability of ranges of particular sample results on the basis of their sketched frequency distribution ^{1,2,3} ;	n = 12	Non-applicable
¹ Supplement: Students estimate the probability of ranges of particular sample results roughly	n = 20	Non-applicable
² Supplement: Students estimates did not correspond to their frequency distribution sketched	n = 7	Non-applicable
³ Supplement: Students overestimated the probability of strongly deviating results	n = 6	Non-applicable

All students were able to understand the frequency distribution of data from repeated sampling, as they made a good sketch of a visualization (or a model) of their expectations in a distribution with a peak at 30 and falling to (almost) zero at the extremes, which corresponded to indicators 2c and 2d. In doing so, all students demonstrated that they were aware that samples vary, but that a sample result that resembled the population proportion (75%) would occur most frequently in more than 100,000 repetitions. The drawings could be divided into the four types shown in Figure 2.3. Eleven students correctly sketched the frequency distribution in the shape of a bar diagram with a peak at 30 and a negative skew (Type 1). Five students indicated that for so many repeated samples, the sample results would not increase/decrease monotonously, but local peaks might occur (Type 2). These five students also correctly sketched the global features of the frequency distribution, although local peaks are unlikely to occur in such a large number of repetitions. These students probably thought that coincidence played a role in this, and they did not (yet) realize that the distribution of samples will stabilize after so many repeated samples (known as the law of large numbers, which we do not expect students to understand here). Two students sketched an almost linear course (Type 3) and two students outlined a smooth curve (Type 4). However, the latter might be caused by the word *sketch* rather than *draw* in the task.

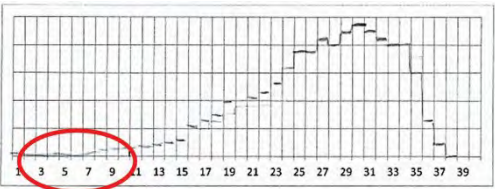
Table 2.8. Students' Estimate of the Probability of a Range of Particular Sample Result on Worksheet 2 (N = 20)

Task	Probability	Examples of written work
1. How do you estimate the probability of a sample result (number of yellow balls) of less than 10 at a sample size of 40? (<i>population proportion 75%</i>)	(Almost) 0% (n = 6)	Very small, but it is possible though.
	1% (n = 6)	A very small probability actually, almost 1%, because there are simply many more yellow than orange balls.
	5% (n = 3)	... because there will always be a chance, only it gets less because the larger majority has that color.
	10% (n = 3)	75 out of 100 balls are yellow.
	Empty (n = 2)	(<i>due to time limitations</i>)

Students were able to estimate the probability of certain sample results roughly, but their estimates did not always correspond to their sketched frequency distribution and some students overestimated the probability of strongly deviating results, and, as a consequence, indicator 2e was only partially

observed. In several tasks, students were asked to estimate the probability of ranges of particular sample results. As an example, Table 2.8 shows the results of one task, in which students were asked to estimate the probability of a sample result of less than 10 in a sample of 40 from a population (size 1,000 and proportion 75%).

Table 2.9. Illustrative Example of the Working Method on Worksheet 2

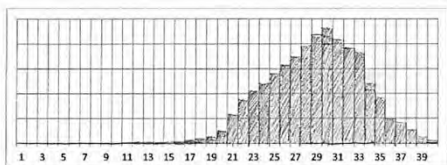
Task	Illustrative example from one student's work
Sketch the expected frequency distribution on the number of yellow balls in a sample of 40, if the first-step physical experiment were repeated 100,000 times. (population proportion 75%)	
How do you estimate the probability of a sample result (number of yellow balls) of less than 10 at a sample size of 40? (population proportion 75%)	I estimate the probability at 10%, because most balls are yellow.

Although it was expected that students would describe their estimate of the probability in words, students apparently felt the need to quantify it (probably because this activity was part of the mathematics lesson) and chose to use percentages. All students estimated the probability of a sample result under 10 less than or equal to 10%, with only six out of twenty students estimating this probability close to zero. These answers demonstrated that students understood that the probability of a strongly deviating result was small. However, this probability was overestimated as six of them indicated that it would be 5% or higher.

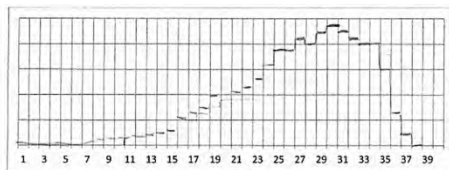
A remarkable result in determining the probability of ranges of particular sample results was that the frequency distribution sketched by the students did not always correspond to their answers (7 out of 20). Table 2.9 shows an example. Although this student wrote down a numerical value, suggesting that he made a calculation or at least made a specific estimate from his sketch, the value did not match his sketched frequency distribution. It seemed that the estimate was based on his intuitive idea of probability rather than being calculated or estimated using his frequency distribution. However, two other students explicitly mentioned that they did calculate the probability, "I estimate

the probability at 0.01% which is about 10 out of 100,000 times.” In this case, the calculation shown was the basis for the correct reasoning.

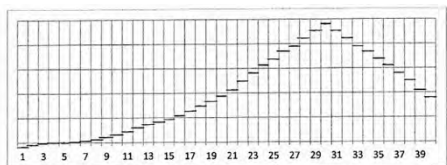
Type 1: Correctly sketched frequency distribution in the form of a bar graph with a peak at 30 and a negative skew (n=11).



Type 2: Correctly sketched global form of the frequency distribution in a bar graph with a peak at 30, but with (unlikely) local peaks (n=5).



Type 3: Correctly sketched global form of the frequency distribution in a bar graph with a peak at 30, but with almost linear progression (n=2).



Type 4: Correctly sketched global form of the frequency distribution with a peak at 30, but with an unrealistic smooth line (continuous distribution) (n=2).

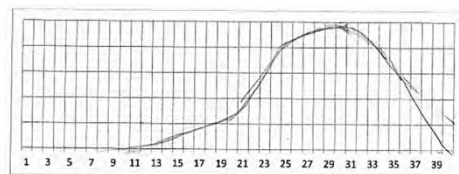


Figure 2.3. Four types of students' sketches (N = 20) of the expected results of repeated sampling (100,000 repetitions) with sample size 40 in a frequency distribution

Third Step: Simulated Sampling Distribution

With regard to the third step, introducing the concept of simulated sampling distribution, the hypothesis was confirmed as the data analysis revealed indicators 3a to 3e. Table 2.10 displays the observed indicators.

The students were able to simulate the sampling distributions at varying sample sizes and from varying number of repetitions and to use these distributions as a model for interpreting the variation and uncertainty involved. By comparing these distributions, they noted on their worksheet that a larger sample size led to less variation in the accompanying estimate of the population and hence to a better inference, and in addition they noted on their worksheet that a larger number of repetitions lead to less variation in the mean of the

samples and hence to a better estimate of the population, which confirmed indicators 3a to 3d.

Populatieproportion <u>75</u> %. Sample size <u>40</u>				
	Number of repetitions <u>10</u>		Sample results converted to the accompanying estimate of the population	
	Most common sample results	Average sample result	Most common estimates of population	Estimate of the population
Simulation 1	26-32	29,5	650-800	738
Simulation 2	26-36	30,1	650-900	753
Simulation 3	25-34	28,4	625-800	710

Populatieproportion <u>75</u> %. Sample size <u>200</u>				
	Number of repetitions <u>10</u>		Sample results converted to the accompanying estimate of the population	
	Most common sample results	Average sample result	Most common estimates of population	Estimate of the population
Simulation 1	136-157	148,2	680-785	741
Simulation 2	143-165	152	715-825	760
Simulation 3	141-156	148,6	705-780	743

Figure 2.4. Example of filled-in table on Worksheet 3

The students could simulate the sampling distributions from repeated sampling easily and independently. They determined the most common sample results for samples with different sizes and different number of repetitions by checking the boundaries of the 95% area with the computer tool. This tool also made it easy to determine the average sample result. Students were asked to examine the effect of sample size and number of repetitions on the sampling distributions. In order to investigate this, they could use the tables on their worksheet. Figure 2.4 displays one filled-in table of a student.

The students were free to decide which population proportion, sample size, and number of repetitions they wanted to examine and compare. Based on

the simulated sampling distributions, they filled in Columns 2 and 3 and subsequently converted these values to the accompanying estimates of the population (size 1,000) in Columns 4 and 5.

By comparing sampling distributions with different sample sizes, students noted that there was less variation in the corresponding estimates of the population and concluded that a larger sample size led to a more accurate outcome. Table 2.11 gives an overview of students reasoning about the effect of sample size on the estimate of the population. In the same way, students compared simulated sample distributions for a varying number of repetitions and found that the average sample result for a large number of repetitions remained almost the same, and from this they concluded that more repeated samples provided a better estimate of the population.

Table 2.10. Overview of Results for LT Step 3

Indicators	Student worksheet (N = 20) (observed number of students)	Video (strong, weak, no evidence)
3a. Students compare the simulated sampling distributions at varying sample sizes and note that repeated sampling with a larger sample size leads to less variation in the accompanying estimate of the population;	n = 20	Strong
3b. Students compare the simulated sampling distributions at varying sample sizes and note that repeated sampling with a larger sample size leads to a better estimate of the population;	n = 20	Strong
3c. Students compare the simulated sampling distributions from varying number of repetitions and note that from repeated sampling with a larger number of repetitions, the mean of these samples is less variable;	n = 20	Strong

3d. Students compare the simulated sampling distributions from varying number of repetitions and note that repeated sampling with a larger number of repetitions, leads to a better estimate of the population;	n = 20	Strong
3e. Students describe how the simulated sampling distribution from repeated sampling can be used to determine most common sample results;	Non-applicable	Weak

Table 2.11. Students' Estimate of the Probability of a Certain Sample Result on Worksheet 2 (N = 20)

Task	Examples from written work
1. What do you notice when you compare the population estimates from a small sample size with those from a larger one?	<p>The estimates from a larger sample size are closer together.</p> <p>With a larger sample size, the average sample result, and the population estimate, are closer together.</p> <p>With a larger sample size, there is less variation in the population estimates.</p>
2. Based on the simulated sample distributions, draw a conclusion on the effect of sample size on the estimate of the population. Complete the following sentence: A larger sample size ...	<p>... leads to a more accurate conclusion.</p> <p>... gives a more precise picture of the number in the population.</p> <p>... reduces the spread of the estimates, which in turn makes your estimate of the population more precise.</p>

Since most students used the boundaries of the 95% to compare the sampling distributions, the teacher asked several individual students what these boundaries meant and how one could use them. These video-taped interactions between teacher and student (TSI) showed that the students' overall idea of the 95% area was correct. For example, based on the screen with the 95% area of the sampling distribution one student explained:

The 95% area consist of two borders, a limit at 2.5% and a limit at 97.5% of the sample results. This is so that you can clearly see what the most common sample results are. The sampling variation is sometimes very large because you carry out many samples. Through the 95% area you can see clearly what the samples usually have as a result.

Not all students were surveyed, and the open nature of the question made it hard to confirm their understanding of the 95% area; as a consequence, we considered indicator 3e as partly observed, even though all students were able to describe how the sampling distribution could be used as a model to interpret variation and uncertainty and to determine most common sample results.

Conclusion and Discussion

In this research we looked for opportunities to make ISI accessible to 9th-grade students. Educational guidelines were extracted from literature and translated into hypotheses about a learning trajectory for students. We addressed the question of how the first part of a learning trajectory that focuses on repeated sampling with a black box introduces students to the concepts of sample, frequency distribution, and simulated sampling distribution. This article reports on the design, implementation and evaluation of the first three steps of a LT for ISI.

The first step of the LT focused on the introduction of sampling. The hypothesis was that students would become aware of sampling variation with categorical data and investigate the effect of repeated sampling and sample size on estimating the population, by conducting the designed activity with the black box. The results show that the indicators associated with the hypothesis were observed. The LT enabled students, inexperienced with sampling, to reason with sample data in a short period of time, including the handling of variation and uncertainty. To estimate the population—the content of the black box—students chose a repeated sampling approach to reduce errors caused by sampling variation. In this specific black-box context, students viewed their sample as “a subset of the population” and not as “a small-scale version of the population” which supports reasoning about variation (Saldanha & Thompson, 2002). Students did not know how to interpret the variation in data, as they noted that they were not entirely confident about their estimates due to the variation in outcomes. This result is in line with studies by Tversky and Kahneman (1971) and Ben-Zvi et al. (2012). The use of whole-class discussions where students exchange and compare their results from repeated sampling

(Wild & Pfannkuch, 1999), along with the growing-samples principle (Bakker, 2004; Ben-Zvi et al., 2012), in which students discuss and test their expectations about increasing sample sizes, was found transportable to our LT. Along the lines of this approach, students predicted what would happen in a following larger sample. While drawing larger samples and exchanging the results, the role of sample size on variation became visible for students. Students experienced and noted that a larger sample size (usually) leads to less variation in the estimate of the population proportion and hence to a better inference. The confrontation with diversity in sample data from the black box supported students' inferential reasoning about the boundaries of variation. Estimates of the population proportions were supported by arguments on sampling variation, repeated sampling, and sample size. As such, this physical black-box experiment seemed a meaningful context to introduce students to the concept of sampling.

The second step of the LT focused on the introduction of the concept of frequency distribution from repeated sampling. The hypothesis was that for students the frequency distribution was primarily a visualization of results obtained from repeated samples. Through considering how this distribution might look like with many repeated samples, students were stimulated to make the conceptual switch to using it as a model for interpreting variation and uncertainty. Along this way, it was expected that during this step, through discussing the question "What happens if this experiment is repeated" and by imagining and visualizing the frequency distribution of 1,000 repeated samples from the black box, students would understand that most sample results will be close to the population proportion and that strong deviations are unlikely. In addition, they were expected to understand that this frequency distribution can be used to estimate the probability of ranges of specific sample results (for example a result of less than ten). The results from step 2 show that most of the corresponding indicators were observed. The question of what happens if the experiment is repeated (Rossman, 2008) was found crucial in this LT. It promoted students' inferential reasoning as they considered and discussed possible sample results. Moreover, this question led to discussion about the difference between probability and chance. By having students draw a sketch of their expectations for many repeated samples in a bar chart, the shape of frequency distribution became visible.

This visualization offered a lead to more enhanced reasoning about variation and uncertainty. As all students were able to consider, sketch, and reason about variation and uncertainty with the frequency distribution on

repeated sampling, students were expected to be able to determine the probability of ranges of particular sample results by using their prior knowledge of ratios. However, as a remarkable result, not all students applied their prior knowledge of ratios to their sketched frequency distribution; some determined the probability on other (maybe more intuitive) ideas. Another finding in this respect is that some students overestimated the probability of strong deviations, which is not surprising at this stage, but can be a point for attention in subsequent lessons. From these results, visualizing the expected frequency distributions on many repeated samplings facilitated more enhanced reasoning about variation and uncertainty, where determining the probability of particular sample results is a point for attention.

The third step focused on the introduction of the concept of sampling distribution. The hypothesis was that students would understand that this distribution can be used as a model for investigating variation and uncertainty. More particularly, that students understand that sampling with a larger number of repetitions leads to less variation in the mean and hence to a better population estimate, and that sampling with a larger sample size reduces the variation in the accompanying estimates of the population and hence leads to a more certain inference, by simulating and comparing sampling distributions with varying sample sizes and from varying number of repetitions. The results of step 3 show that the indicators that supported the hypothesis were observed. From students' experience with the frequency distribution of many repeated samples in step 2, the transition to the simulation of the sampling distribution, also called resampling (Garfield et al., 2015; Manor & Ben-Zvi, 2015; Watson & Chance, 2012), was easily made. The students were already familiar with the shape of this distribution. The students were able to determine the most likely sample results by using the digital tool. Here they used the boundaries of the 95% area, which were available in the tool. The students simply adopted these boundaries. Although most students were able to give a correct description of these boundaries, the results do not confirm whether all students understood these boundaries and their application. The comparison of distributions from repeated sampling gave them insight into the effect of repeated sampling and sample size on the estimate of the population. As such, the results show that students were able to use the idea of a simulated sampling distribution as a model for further investigating variation and uncertainty.

This study gave an insight into how a LT that focuses on the concept of sample, frequency distribution, and sampling distribution can enhance 9th-grade students' informal inferential reasoning. The results show how students used

these concepts to underpin their inferences, especially with regard to variation and uncertainty. In addition, the results show what barriers students encountered during their work on the HLT.

From the viewpoint of the researcher as an experienced teacher, the main element in this LT that allowed students to go through the three steps smoothly seemed to be the accessibility of the three successive steps. From their concrete experiences with sampling variation in step 1, through imagining and visualizing the scaling up of this experiment in step 2, the students could easily make the transition to reasoning with the sampling distribution in step 3. From their point of view, the computer took over their manual work. This approach provided them insight into how the sampling distribution arises and how it can be used as *a model for* investigating possible sample results to interpret variation and uncertainty. Known difficulties concerning the three main concepts of sample, frequency distribution, and sampling distribution, hardly occurred and apparently were avoided. As a consequence, this approach seems to help students engage with these concepts, which supports them in using new insights in new situations, making ISI accessible.

In our study, the idea of *model of* to *model for* was primarily used as a design heuristic to promote emergent modeling (Gravemeijer, 1999). In retrospect, we have become intrigued by what happens cognitively when students make the transition from seeing a graph as a representation (*model of*) of a frequency distribution to seeing a dataset as a distribution with particular characteristics that help to make inferences (*model for*). It seems promising to analyze such transitions through the theoretical lens of objectification (reification, reflective abstraction, or hypostatic abstraction). Where it concerns the learning of function (Sfard, 1991), it is known that students initially see functions as processes, and typically not as objects with characteristics. The desired dual understanding of functions or other mathematical objects as both process- and object-like has been referred to as “procept” (Gray & Tall, 1994). In our case, we speculate that the sampling process in which students see (sampling) distributions emerge may be such a process view, which forms a basis for seeing a distribution as an object (cf. Bakker, 2007b). From such a perspective, the question arises whether objectification is indeed the mechanism that enables the cognitive transition between the learning steps, and thus conceptualization.

Objectification involves constructing an object in a representational system, for example the visualization of a sample, experimenting with this

object and then observing the results of experimenting, as a reflection phase. New objects can be created in the reflection phase, when part of an object is seen as an entity in itself (Bakker, 2007a). The central point of this reflection phase is that the new objects formed by this process can be used as a means for further objectification, for new, higher-level processes, and for further steps in improving additional knowledge, and thus conceptualized as an independent entity. According to Sfard and Lavie (2005), the development of objectification begins with participation in offered routines with the object, whereby these routines gradually transform into real explorations with the object as an independent entity. They emphasize that this learning process is a one-way street, which is difficult to reverse, making it hard for adults to recognize. With regard to the transfer from LT step 1 to step 2, objectification may involve the transfer of sampling as the process of creating elements of a dataset, to a sample as an element or new object in the frequency distribution from repeated samples in step 2. In this way, objectification can be viewed as the mechanism underlying the ideas of repeated and growing samples. With regard to the transfer from LT step 2 to 3, objectification may facilitate the transfer of the frequency distribution as a model of results generated from repeated sampling into a model for further investigating variation and uncertainty. Given the importance of such mechanism of objectification we recommend follow-up research in this area.

Our advice for redesign of the HLT focuses on two main points. The first point includes the integration of students' prior knowledge about proportions to determine the probability of ranges of particular sample results with the frequency distribution. This could be achieved by calculating proportions by using (and discussing possible) units on the vertical axis of the (expected) frequency distribution and by applying and discussing this distribution in multiple and more different situations. The second point for redesign is to pay more attention to reasoning with the simulated sampling distribution in various situations and not automatically use the 95% area. Discussion on students' analyses and not only on the results, will support students' development of strong mathematical arguments (McClain, McGatha, & Hodge, 2000).

Considerations for using the LT in other settings are the following. The LT was implemented in the classroom of the teacher-researcher, who was very familiar with the class and the designed materials. We are aware that this favorable condition should be taken into account when readers want to use the ideas presented here in other contexts. Researchers who would like to repeat such activities should also consider that most Dutch students are not used to

whole-class discussions during the mathematics lessons. In our research, it was important to encourage them to reason and discuss with each other from the beginning of the trajectory. Teachers and researchers should also take into account that Dutch students are used to closed assignments from their textbook and are unfamiliar with working with more open and inquiry-based tasks. Another point of consideration is that we worked with pre-university students, the top 20% of our education system. In other situations, students may need more time. As the results in this research are based on a small-scale pilot in the class of the teacher-researcher, these results are not generalizable to a regular classroom without further research.

This research that focused on repeated sampling with the black box as a first part of a LT seems a promising proof of principle how to make ISI—e.g., reasoning about variation and uncertainty—accessible for students along the lines of sample, frequency distribution, and simulated sampling distribution. The results of this study will be used to revise the first part of the LT, and as a next step, in a follow-up study, to (re)design the whole LT, and to improve effectiveness, efficiency, scale up and compare our HLT with alternatives.



Statistical Modeling Processes Through the Lens of Instrumental Genesis

This chapter is based on

Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2021). Statistical modeling processes through the lens of instrumental genesis. *Educational Studies in Mathematics*. <https://doi.org/10.1007/s10649-020-10023-y>

Abstract

Digital technology is indispensable for doing and learning statistics. When technology is used in mathematics education, the learning of concepts and the development of techniques for using a digital tool are known to intertwine. So far, this intertwinement of techniques and conceptual understanding, known as instrumental genesis, has received little attention in research on technology-supported statistics education. This study focuses on instrumental genesis for statistical modeling, investigating students' modeling processes in a digital environment called TinkerPlots. In particular, we analyzed how emerging techniques and conceptual understanding intertwined in the instrumentation schemes that 28 students (aged 14–15) develop. We identified six common instrumentation schemes and observed a two-directional intertwining of emerging techniques and conceptual understanding. Techniques for using TinkerPlots helped students to reveal context-independent patterns that fostered a conceptual shift from a model *of* to a model *for*. Vice versa, students' conceptual understanding led to the exploration of more sophisticated digital techniques. We recommend researchers, educators, designers, and teachers involved in statistics education using digital technology to attentively consider this two-directional intertwined relationship.

Keywords

statistical modeling, instrumental genesis, statistical reasoning, TinkerPlots, simulated sampling distribution

Introduction

The increasing use of digital technology in our society requires an educational move towards learning from and with digital tools. This is particularly urgent for statistics education where digital technology is indispensable for interpreting statistical information, such as real sample data (Gal, 2002; Thijs, Fisser, & Van der Hoeven, 2014). For such interpretations, understanding underlying statistical models is fundamental (Manor & Ben-Zvi, 2017). Current technological developments offer digital tools—for example TinkerPlots, Fathom, and Codap—that provide opportunities to deepen understanding of statistical modeling and models. These digital tools enable students to build statistical models and to use these models to simulate sampling data, and therefore offer means for statistical reasoning with data (Biehler, Frischemeier, & Podworny, 2017). As such, modeling with digital tools is promising for today's and tomorrow's statistics education.

Although statistics education is developing as a domain distinct from mathematics, the use of digital tools is a shared problem space and collaboration within shared spaces can strengthen each domain (Groth, 2015). From other domains in school mathematics, for example algebra, it is well known that as soon as digital tools are used during the learning process, the development of conceptual understanding becomes intertwined with the emergence of techniques to use the digital tool (Artigue, 2002; Drijvers, Godino, Font, & Trouche, 2013). For teachers, researchers, educators, and designers, insight into this intertwined relationship of learning techniques and concepts is a prerequisite for deploying digital tools in such a way that they are productive for the intended conceptual understanding. In the meantime, due to a lack of insight into this intertwining, undesired influence of techniques for using the digital tool on the intended conceptual development can be overlooked. This complex relationship, however, has so far received little attention in research on technology-supported statistical modeling processes.

A useful perspective to grasp the relationship between the learning of digital techniques and conceptual understanding is instrumental genesis (Artigue, 2002). In this theoretical view, learning is seen as the simultaneous development of techniques for using artifacts, such as digital tools, and of domain-specific conceptual understanding, for example statistical models and modeling. The perspective of instrumental genesis seems promising to gain knowledge about learning from and with digital technology. As such, the aim of

this study is to explore the applicability of the instrumental genesis perspective to statistics education, and to statistical modeling processes in particular.

Theoretical Framework

In this section we elaborate on two main elements of this study: statistical modeling and instrumental genesis.

Statistical Modeling: Techniques and Concepts

Digital tools for statistical modeling have the potential to deepen students' conceptual understanding of statistics and probability, and enable them to explore data by deploying techniques for using the tool. They also offer possibilities to visualize concepts that previously could not be seen, such as random behavior (Pfannkuch, Ben-Zvi, & Budgett, 2018). Such educational digital tools, for example TinkerPlots, provide opportunities for statistical reasoning with data, as students build statistical models and use these models to simulate sample data (Biehler et al., 2017).

Modeling processes with a digital tool such as TinkerPlots require the development of digital techniques. Digital TinkerPlots techniques for setting up statistical models and simulating data are helpful to introduce key statistical ideas of distribution and probability (Konold, Harradine, & Kazak, 2007). The research by Garfield, delMas, and Zieffler (2012) suggests that students can learn to think and reason from a probabilistic perspective—or, as the authors call it, “really cook” instead of following recipes—by using TinkerPlots techniques to build a model of a real-life situation and to use this model for simulating repeated samples. This way to understand the probability involved in inferences is also reflected in our previous study (Van Dijke-Droogers, Drijvers, & Bakker, 2020) in which an approach based on repeated sampling from a black box filled with marbles seemed to support students in developing statistical concepts. In this approach, students developed TinkerPlots techniques to investigate what sample results would likely occur by chance. Statistical modeling in the study presented here requires TinkerPlots techniques for building a model by choosing a graphical representation (e.g., a bar or pie chart), entering population characteristics (e.g., population size, attributes, and proportions) and entering the sample size, of a real-life situation from a given context to solve a problem. Next steps include TinkerPlots techniques for simulating repeated samples by running the model and visualizing the results in a sampling distribution, for enabling to reason about probability—taking into account number of repetitions and sample size—and to answer the problem using simulated data.

Statistical modeling processes with TinkerPlots also require, in addition to the development of TinkerPlots techniques, an understanding of the concepts involved. The literature elaborates several viewpoints on statistical modeling. We discuss three viewpoints and indicate how we incorporated them in our study. First, Büscher and Schnell (2017) argue that the notion of emergent modeling (Gravemeijer, 1999)—the conceptual shift from a model of a context-specific situation to a model for—can also be applied to statistical reasoning in a variety of similar and new contexts. Second, statistical modeling involves the interrelationship between the real world and the model world. This relationship is elaborated in Patel and Pfannkuch’s framework (2018) that displays students’ cognitive activities about understanding the problem (real world), seeing and applying structure (real world–model world), modeling (model world–real world), analyzing simulated data (model world), communicating findings (model world–real world). Third, for reasoning with models and modeling, Manor and Ben-Zvi (2017) identify the following dimensions: reasoning with phenomenon simplification, with sample representativeness, and with sampling distribution. Statistical modeling includes the process of abstracting the real world into a model and then using this model for understanding the real world. In short, Büscher and Schnell (2017) emphasize the importance of developing context-independent models for statistical modeling processes, Patel and Pfannkuch (2018) outline the interaction between the real and the model world, and Manor and Ben-Zvi (2017) address the different dimensions when reasoning with models. These viewpoints provide insight into the development of concepts for statistical modeling. In the study presented here, we embodied the viewpoints in the design of students’ worksheets. On these worksheets, students are requested to build and run a model of a real world situation in TinkerPlots and to use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real world situation.

Understanding and reasoning with the simulated sampling distribution from repeated samples is, as mentioned by Manor and Ben-Zvi (2017), essential for statistical modeling. However, the concept of sampling distribution is difficult for students. The study by Garfield, delMas, and Chance (1999) focused on the design of a framework to describe stages of development in students’ statistical reasoning about sampling distributions. Their initial conception of the framework identified five levels that evolve from (1) idiosyncratic reasoning—knowing words and symbols related to sampling distributions, but using them without fully understanding and often incorrectly—through (2) verbal reasoning, (3) transitional reasoning and (4) procedural

reasoning, towards (5) integrated process reasoning—complete understanding of the process of sampling and sampling distributions, in which rules and stochastic behavior are coordinated. In our study, these levels will be used to indicate students' conceptual understanding of statistical modeling. Students' difficulties in reasoning with the sampling distribution are often related to misconceptions about basic statistical concepts such as variability, distribution, sample and sampling, the effect of sample size and confusion of results from one sample with the sampling distribution. According to Chance, delMas, and Garfield (2004), ways to improve students' level of understanding statistical modeling include techniques for exploring samples, comparing how sample behavior mimics population behavior, and for both structured and unstructured explorations with the digital tool. As such, conceptual understanding of statistical modeling involves the building, application and interpretation of context-independent statistical models—in our study the sampling distribution of repeated sampling—to answer real-life problems.

Instrumental Genesis

Using digital tools in a productive way for a specific learning goal requires insight into the intertwined relationship between emerging digital techniques and conceptual understanding. A useful perspective to grasp the intertwining of learning techniques and concepts is *instrumental genesis*. A fundamental claim in this theory is that learning can be seen as the intertwined development, driven by the student activity in a task situation, of techniques for using artefacts—for example a digital tool—and cognitive schemes that have pragmatic and epistemic value (Artigue, 2002; Drijvers et al., 2013). In this perspective, the conception of “instrument” and instrumental genesis are used in the sense described by Artigue (2002):

The instrument is differentiated from the object, material or symbolic, on which it is based and for which is used the term “artefact”. Thus an instrument is a mixed entity, part artefact, part cognitive schemes which make it an instrument. For a given individual, the artefact at the outset does not have an instrumental value. It becomes an instrument through a process, called instrumental genesis, involving the construction of personal schemes or, more generally, the appropriation of social pre-existing schemes. (p. 250)

According to Vergnaud (1996), a scheme is an invariant organization of behavior for a given class of situations. Such a scheme includes patterns of

action for using the tool and conceptual elements that emerge from the activity. In the study presented here, the tasks on students' worksheet intend to construct personal instrumentation schemes consisting of TinkerPlots techniques and conceptual understanding of statistical modeling. The identification of schemes can structure and deepen the observation of students' emerging technical actions and statistical reasoning, and hence provides insight into the intertwined development of techniques and concepts.

As the application of instrumental genesis within the field of statistics education hardly exists, we present an example from a study within the context of algebra. Table 3.1 shows an instrumentation scheme concerning the use of a symbolic calculator for solving parametric equations, from a study by Drijvers et al. (2013). The intertwined relationship can be seen, for example, in scheme D. Here, students were asked to solve the parametric equation with respect to x . On the one hand, in order to use the correct techniques, students must be able to identify the unknown in the parameterized problem situation to enter the correct command "solve with respect to x " into their computer algebra calculator. On the other hand, the available options of the tool invite students to distinguish between the parameter and the unknown. In the study by Drijvers et al., the identification of students' instrumentation schemes provided insight into how the learning of techniques for using a computer algebra system and the conceptual understanding of solving parametric equations emerged in tandem. Furthermore, the identified schemes helped the researchers to reveal several conceptual difficulties students encountered while solving parametric equations with the digital tool.

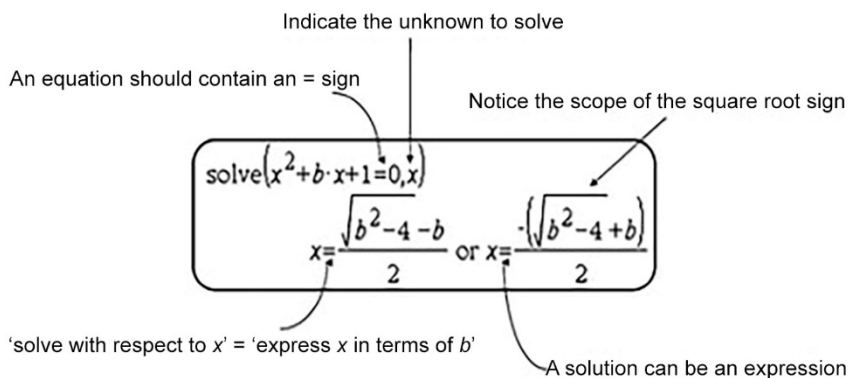
As a second example, we present the findings from one of the scarce studies on instrumental genesis within the field of statistics education, conducted by Podworny and Biehler (2014). In their study, within a course on hypothesis testing and randomization tests with p values, university students used simulations with TinkerPlots. Students noted their own schemes to plan and structure their actions. These schemes drawn up by students proved useful as a personal work plan; however, it was difficult to identify common instrumentation schemes and to unravel how TinkerPlots techniques and conceptual understanding emerged together. Our study differs from theirs, as we identified our students' schemes by observing their actions and reasoning.

In general, instrumental genesis is considered an idiosyncratic process, unique for individual students. Yet, it takes place in the social context of a classroom, and as researchers we are interested in possible patterns. As such,

identifying instrumentation schemes concerns the complexity of unraveling patterns in the diversity of individual schemes that students develop. The study presented here seeks to identify common instrumentation schemes by observing students' actions and reasoning when statistically modeling in TinkerPlots, and then to use these schemes to zoom in on the genesis of the schemes to reveal how emerging TinkerPlots techniques and the conceptual understanding of statistical modeling intertwine.

Table 3.1. Example of an Instrumentation Scheme for Solving Parametric Equations with a Computer Algebra System (Drijvers et al., 2013)

Digital techniques	Conceptual understanding
A. Use the Solve-option of the Graphing Calculator and enter the given function	Knowing that the Solve command can be used to express one of the variables in a parameterized equation in other variables
B. Enter the '=0' sign	Knowing the difference between an expression and an equation
C. Enter the unknown to solve (x)	Realizing that an equation is solved with respect to an unknown
D. Solve the equation with respect to x	Being able to identify the unknown in the parameterized problem situation
E. Give the solution for the parametrized equation	Being able to interpret the result, particularly when it is an expression, and to relate it to graphical representations



Research Aim and Question

To explore the applicability of the instrumental genesis perspective to statistics education, and to statistical modeling in particular, we conducted an explorative case study. This study focuses on 14-to-15-year-old students' intertwined development of learning techniques for using TinkerPlots and conceptual understanding of statistical modeling. We address the following question:

Which instrumentation schemes do 9th-grade students develop through statistical modeling processes with TinkerPlots and how do emerging techniques and conceptual understanding intertwine in these schemes?

Methods

This study is part of a larger design study on statistical inference. Our previous study focused on the design of a learning trajectory in which students were introduced to the key concepts of sample, frequency distribution, and sampling distribution, with the use of digital tools (Van Dijke-Droogers et al., 2020). As a follow up, this study focuses on the specific role of digital techniques on conceptual understanding by examining how 28 9th-grade students work on TinkerPlots worksheets, which were designed to engage in statistical modeling.

Design of Student Worksheets

A suitable stage to investigate students' instrumental genesis—their development of schemes that include TinkerPlots techniques and conceptual understanding of statistical modeling—is after the introduction of the tool and the concepts, when they engage in the emergent modeling process of applying gained knowledge in new real-life situations. Prior to working with the TinkerPlots worksheets, students had a brief introduction to the tool and concepts. These preparatory activities were designed within the specific context of a black box with marbles and involved three 60-min lessons. Two of these lessons concentrated on physical black box experiments and one on simulations. Both the physical and simulation-based preparatory activities introduced students to statistical modeling by addressing concepts such as sample, sampling variation, repeated sampling, sample size, frequency distribution of repeated sampling and (simulated) sampling distribution (Chance et al., 2004). The introduction of TinkerPlots techniques was done in the third 60-min lesson through a classroom demonstration of the tool by the teacher, followed by students practicing themselves using an instruction sheet. On this instruction sheet, the TinkerPlots techniques for making a model, simulating repeated

samples and visualizing the sampling distribution were listed. The brief introduction on techniques and concepts focused on the black box context only.

For the study reported here, we designed five worksheets. In one 60-min lesson per worksheet, we invited students to apply and expand their emerging knowledge from the preparatory black box activities in new real-life contexts. The design of the worksheets was inspired by studies from Patel and Pfannkuch (2018), Manor and Ben-Zvi (2017), and Chance et al. (2004). In each worksheet, the students were asked to build and run a model of a real world situation in TinkerPlots and to use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real world situation. The structure of these worksheets is shown in Table 3.2. In each Worksheet (W1 to W5) a new context was introduced. We chose contexts with categorical data to minimize the common confusion between the distribution of one sample and sampling distribution (Chance et al., 2004) and to optimize the similarity with the black box context in the preparatory activities. When carrying out the tasks on W1 to W5, students could use the TinkerPlots instruction sheet from the preparatory activities. The aim of the worksheets was to expand students' understanding of statistical modeling—that is, the building, application and interpretation, of context-independent statistical models; in our study, the sampling distribution of repeated sampling—by using TinkerPlots as an instrument.

Participants

We worked with two groups, each consisting of fourteen 9th-grade students. Group 1 consisted of students in school year 2018–2019 and Group 2 of students in school year 2019–2020. All students were in the pre-university stream, and thus belonged to the 15% best performing students in our educational system. The students were inexperienced in sampling and had no prior experience in working with digital tools during mathematics classes.

The students in Group 1 went through the preparatory activities described earlier during the regular math lessons in school. Their teacher had been involved in the research project and had already carried out these lessons several times. All twenty students from the class were invited to participate in the session at Utrecht University's Teaching and Learning Lab (a laboratory classroom) and fourteen of them applied. During the lab session, these students worked on Worksheets 1–3 (W1 to W3), the initial phase of the teaching sequence. For practical reasons—such as missing regular classes and travel time—multiple research sessions with the same students were not possible.

Table 3.2. Structure of Designed Worksheets 1–5

Worksheet component	Student activity
a. Explore and identify important factors from a given real-life problem to build a population model in TinkerPlots	Try to understand the situation and the data collection by defining the problem, making predictions and considering variation. Apply structure by identifying all known real world factors, considering model tools in TinkerPlots to represent real world factors and evaluating whether all relevant factors are included in the model.
b. Build and run the model by simulating sample results and examine the behavior of the model	Use TinkerPlots to examine the behavior of the model by visualizing single sample results and repeated sample results in respectively sample and sampling distributions by, checking variation in simulated data distributions at varying sample sizes and varying number of repeated samples, comparing these data with the contextual knowledge, evaluating model fit by checking how simulated data mimic the model.
c. Examine and interpret the simulated results by using the sampling distribution	Interpret the results of simulated data by identifying and using TinkerPlots tools to answer specific tasks and, in addition, by considering (the probability of a specific) range of outcomes, sampling variability, effect of sample size and number of repeated samples.
d. Answer the problem using the simulated data	Communicate findings by stating background to the problem, making model informed decisions, recognizing effects of underlying randomness and stating limitations of the decisions.

Therefore, one school year later, we performed lab sessions again, but with a different group of students, here called Group 2. These students from the same school and with the same teacher as Group 1, went through the same preparatory lessons and W1 to W3 during their regular math lessons at their school. Again, fourteen students applied to participate in the research sessions at the university. These students in Group 2 were similar to those in Group 1: They performed at a similar level in mathematics, as their overall grades for the school year averaged 6.6 on a scale of 10, which was comparable to 6.9 in

Group 1. In addition, students' performance in the preparatory tasks averaged 8 on a scale of 10 in both groups. The teacher judged the starting level of the two groups to be similar. During the research sessions, the students of Group 2 worked on W4–W5, the more advanced phase of the teaching sequence. An overview of participants can be found in Table 3.3.

Table 3.3. Overview of Participants and Data Collection

Participants		Data collection			
Group	School year	Average math grade	Average grade preparatory tasks	Data from students' work	Total time duration recordings
1 (n=14)	2018 – 2019	6.9	8	W1–W3 Initial phase	28 hrs
2 (n=14)	2019 – 2020	6.6	8	W4–W5 More advanced phase	17 hrs

Data Collection

The data consisted of video and audio recordings from two classroom laboratory sessions. During the first 5-hour session in Utrecht University's Teaching and Learning Lab, the fourteen students of Group 1 worked in teams of two or three on the designed W1–W3. The advantage of this lab setting over a classroom environment was that detailed video recordings could be made of students' actions in TinkerPlots and their accompanying conversations. The students were specifically asked to express their thoughts while solving the problem, the think-aloud method (Van Someren, Barnard, & Sandberg, 1994). The teams worked on a laptop, the screen of which was displayed on an interactive whiteboard. Figure 3.1 shows the setup in the lab. During the second lab session, we collected video and audio recordings from fourteen students of Group 2, while working on W4–W5.

Data Analysis

The data analysis consisted of three phases: (1) identifying common instrumentation schemes, (2) examining the global scheme genesis process during the work, and (3) examining the scheme genesis process in depth (Table 3.4).



Figure 3.1. Students working with TinkerPlots in Utrecht University's Teaching and Learning Lab

Table 3.4. Overview of Data Analysis

Phase and objective	Outline	Data-driven / theory-driven
Phase 1: Qualitative Data Analysis	Step 0 (Prior to data collection): Preformulating instrumentation schemes based on theories on statistical modeling	Theory-driven
Identification of students' instrumentation schemes	Step 1: Observing each student at a certain local segment of the teaching sequence	Data-driven
	Step 2: Categorizing data from step 1, by using preformulated instrumentation schemes	Data- and theory- driven
	Step 3: Identifying global patterns of instrumentation schemes for more students, by using the categorized data of step 2	Data- and theory- driven
Phase 2: Interpretive Content Analysis	Defining technical levels for TinkerPlots techniques and conceptual levels for understanding statistical modeling	Theory-driven
Examination of students' scheme		

development by identifying their TinkerPlots techniques and levels of understanding statistical modeling during the work	Refining and specifying technical and conceptual levels in each instrumentation scheme Assigning levels to student actions when working on W1 and W5	Data- and theory-driven Data- and theory-driven
Phase 3: Case study More detailed examination of students' scheme development	Further examining how TinkerPlots techniques and understanding statistical modeling intertwine in the schemes students develop, that is, how techniques may support conceptual understanding and the other way around, by zooming in on developing personal schemes of students	Data-driven

In phase 1 of the analysis, we used a combined approach of theory-driven (prior to data collection) and bottom-up (based on the data) to identify emerging instrumentation schemes. The final results can be found in Table 3.7. To identify the schemes, we conducted qualitative data analysis as defined by Simon (2019): A process of working with data, so that more can be gleaned from the data than would be available from merely reading, viewing, or listening carefully to the data multiple times (p. 112). In step 0, prior to the data collection, we defined preformulated schemes. These schemes were based on the theories on statistical modeling (Büscher & Schnell, 2017; Chance et al., 2004; Gravemeijer, 1999; Manor & Ben-Zvi, 2017; Patel & Pfannkuch, 2018), and instrumental genesis (Artigue, 2002), and on expertise developed in previous interventions (Van Dijke-Droogers et al., 2020). In these preformulated schemes, specific TinkerPlots techniques were related to students' understanding of statistical modeling. In step 1, we observed each student at a certain local segment of the teaching sequence, for example building a model of the population (W1 Task 5), and analyzed the techniques and concepts that were manifested in students' actions and reasoning at that local segment. In step 2, we categorized the data from step 1 by using the preformulated schemes. To do this categorization, at the same time as preformulated schemes were assigned, we expanded, refined and adjusted them to include the observed data. In step 3, we used the categorized data of step 2 to identify patterns for more students. By systematically and iteratively going

through the categorized data, both within one student over several schemes and across students, we identified global patterns in emerging instrumentation schemes. These global patterns occurred to a certain extent in every student and across students while working on each worksheet. By adapting the preformulated schemes to the global patterns, we identified students' instrumentation schemes.

Table 3.5. Technical Levels for Using TinkerPlots

Level	Description
1. Non-user	Students are not able to carry out the TinkerPlots techniques.
2. Limited user	Students carry out the TinkerPlots techniques by following the instructions stepwise; the techniques are still carried out hesitantly; haphazard trial and error or simply trying something.
3. Developing user	Students carry out correct TinkerPlots techniques by following the instructions most of the time; incorrect techniques are used but later corrected.
4. Experienced user	Students carry out correct TinkerPlots techniques fluently—that is, fast and without mistakes—sometimes augmented by newly explored TinkerPlots techniques.
5. Expert / Discerning user	Students make well considered decisions for correct TinkerPlots techniques.

Concerning phase 2 of the analysis, the data for examining students' instrumental genesis, that is, their scheme development during the work, we used interpretive content analysis (Ahuvia, 2001). This variant of content analysis allowed us to identify both explicitly observed and latent content of students' technical actions and reasoning. To identify possible progress in students' TinkerPlots techniques, we defined five technical levels of proficiency. These levels were based on Davies' (2011) levels of technology literacy and refined by both our experiences from previous research and the collected video data. Davies defined six levels of users (non-user, potential user, tentative user, capable user, expert user and discerning user) each of which

corresponds to ascending levels of use: none, limited, developing, experienced, powerful, and selective. For our study (Table 3.5) we merged the last two levels of technology literacy, as our students were unable to reach the highest level in the short period of time working on W1 to W5. Based on the observed data, we specified the five technical levels for using TinkerPlots, for each instrumentation scheme. The specified technical levels were used to analyze students' scheme development during the work on W1 and W5, respectively. To identify possible progress in students' understanding of statistical modeling, we defined five conceptual levels. The conceptual levels are displayed in Table 3.6. The conceptual levels were merely based on the previously described levels by Garfield et al. (1999). These conceptual levels for understanding statistical modeling were further specified for each instrumentation scheme on the basis of the observed video data and prior experiences.

Table 3.6. Conceptual Levels of Understanding Statistical Modeling

Level	Description
1. Incorrect reasoning	Wrong statements and/or incorrect using words and symbols related to the specific item of conceptual understanding.
2. Idiosyncratic reasoning	Knowing words and symbols related to the specific item of conceptual understanding but using them without fully comprehending and often incorrectly.
3. Verbal reasoning	Verbal understanding of the item but unable to apply it to the actual behavior. For example, the student can reproduce that results from a larger number of repeated samples lead to a better estimate of the population but does not understand how key concepts such as variability and range are integrated.
4. Transitional reasoning	Correctly identifying one or two features of the item without fully integrating these features. For example, identifying and relating just one or two of the features (1–3) in understanding that results from more repeated samples lead to a better estimate of the population: more repeated samples lead to (1) a smoother sampling distribution, without local peaks, with (2) a peak at the

population proportion and (3) to an average that resembles the population.

5. Integrated process reasoning	Understanding of the process of sampling and sampling distributions. For example, understanding the effect of simulating a larger number of repeated samples to the shape, peak and average of the sampling distribution and that, as a consequence, more repeated samples lead to a better estimate of the population.
---------------------------------	---

The specification of both the technical and conceptual levels for coding the data was discussed in-depth with experts in this domain. Although the students worked in teams of two or three, we analyzed their proficiency levels individually. We did so as we noticed considerable differences in individual proficiency within one team and also because there was cooperation and consultation between teams. To check the reliability of the first coder's analysis, a second coder analyzed the video data of students' activities with W1 and W5, for both the coding of technical and conceptual levels. A random sample of 5% of the data (30 out of 600 fragments) was independently rated by the second coder. The second coder agreed on 85% of the codes. Deviating codes, which were limited to two levels difference at most, were discussed until agreement was reached.

In phase 3, to further examine the intertwined relationship between developing TinkerPlots techniques and understanding statistical modeling, that is, how techniques may support conceptual understanding and vice versa, we used case studies to investigate students' instrumental genesis. In these case studies, we zoomed in on developing personal schemes of students in both the initial and more advanced phase.

Results

In this section, we first present the six instrumentation schemes we identified, each including TinkerPlots techniques and conceptual understanding of statistical modeling (Table 3.7). Second, we describe students' global scheme development during the work, by presenting the levels at which the students used the techniques and concepts while working on Worksheets 1 and 5. Third, we describe two students' cases to reveal in more detail the intertwining of

emerging techniques and conceptual understanding in the personal instrumentation schemes that students developed.

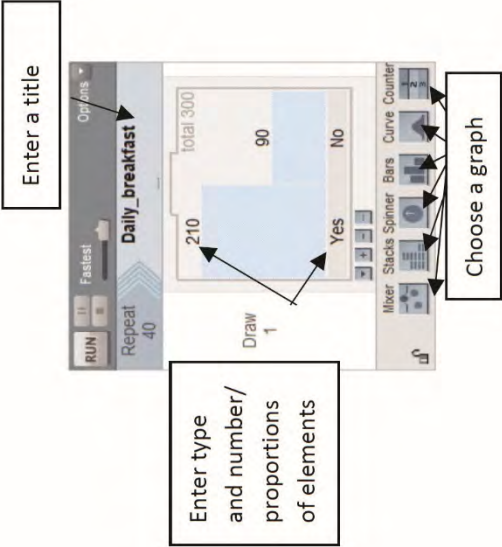
Identified Instrumentation Schemes

In observing students' work, we identified six instrumentation schemes, called (A) Building a model, (B) Running a model, (C) Visualizing repeated samples, (D) Exploring repeated samples, (E) Exploring sample size, (F) Interpreting sampling distribution (see Table 3.7). Column 1 provides a description of each scheme, column 2 displays a screenshot of students' TinkerPlots techniques for the scheme at stake, and column 3 shows students' understanding of statistical modeling that we could distil from their reasoning during these actions. Each instrumentation scheme incorporates a specific modeling process, ranging from building a population model by exploring and identifying important information in a given real-life problem in scheme A, to answering a given problem by interpreting the simulated sampling distribution from repeated sampling in scheme F. As such, the identified instrumentation schemes display how specific TinkerPlots techniques occurred simultaneously with particular elements in students' understanding of statistical modeling.

Students' Global Scheme Development

We now describe students' instrumental genesis by presenting the observed levels at which the students used the TinkerPlots techniques and demonstrated their understanding of statistical modeling in their reasoning, throughout the teaching sequence. In each Worksheet (W1 to W5), instrumentation schemes A to F were addressed. Data from group 1 while working on W1 were used to indicate students' level in the initial phase of the teaching sequence, and data from group 2 while working on W5 for the more advanced phase. Students' technical actions with TinkerPlots were coded in technical levels for each scheme and for each student. For example, concerning W1 scheme A, five students out of fourteen in group 1 were unable to build a population model. They encountered difficulties in finding the input options for the parameters of the population model or for graphical representations of the model and, therefore, we coded their actions for W1 scheme A as technical level 1. As another example, concerning W5 scheme A, six students out of fourteen in group 2 were capable of making well thought out choices from newly explored TinkerPlots options to make a model, for example by using non-instructed options for graphical representations like pie chart or histogram, and we coded their actions technical level 5.

Table 3.7. Instrumentation Schemes for Statistical Modeling through Simulating Repeated Samples with TinkerPlots

<div><div><div>Instrumentation Schemes A: Building a model</div><div>Explore and identify important factors from a given real-life problem to build a population model in TinkerPlots</div></div></div>	
<div><div>TinkerPlots techniques</div><div></div></div>	<div><div>Conceptual Understanding</div><div>A population model can be built by reducing/filtering the given data and using a suitable graphical representation (depending on the description/context given), for example a bar or pie chart, that incorporates the important factors, such as the total number of elements (population size) and the number or proportion of elements with a specific characteristic.</div></div>

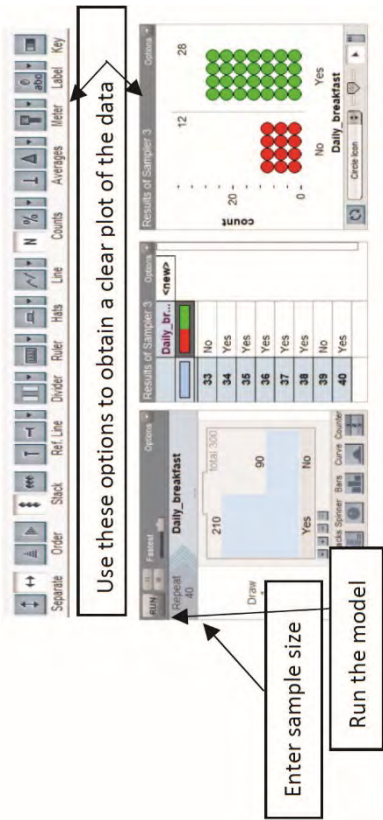
Instrumentation Schemes B: Running a model

Run the model by simulating a sample

TinkerPlots techniques

Conceptual Understanding

Collected results from a sample can be sorted/grouped and summarized in a suitable graph/visualization containing important sample features such as absolute and/or relative numbers for each characteristic.



Instrumentation Schemes C: Visualizing repeated samples

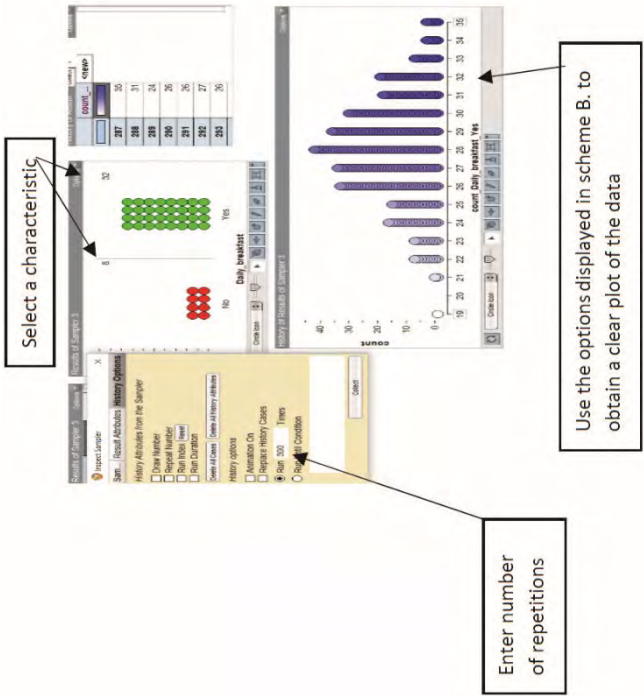
Simulate repeated sample results and visualize the sampling distribution;

Examine the behavior of the model

TinkerPlots techniques

Conceptual Understanding

To answer a real-life problem, the behavior of the model with respect to a specific characteristic (related to the real-life data)—for example the number of students having daily breakfast—can be examined by simulating repeated samples and displaying the results in a sampling distribution. A sampling distribution displays the results of each sample in one bar chart (or stacked dot plot). Along the horizontal axis are the possible sample results displayed and along the vertical axis the frequency with which each result occurs; sample results from the same population model will vary due to chance, with results close to the population model occurring more frequently than strongly deviating sample results.



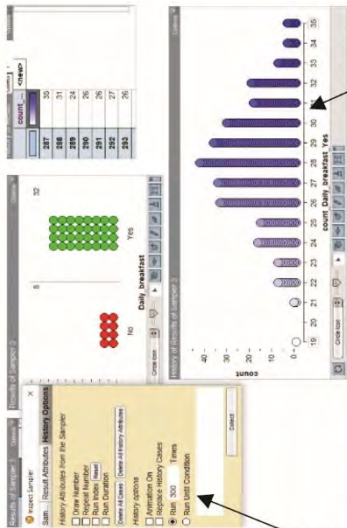
Instrumentation Schemes D: Exploring repeated samples

Simulate more repeated samples and interpret the results using the sampling distribution

TinkerPlots techniques

Conceptual Understanding

Results from a larger number of repeated samples lead to a better estimate of the population, that is, more repeated samples lead to (1) a smoother sampling distribution—without local peaks—(2) with a peak corresponding with the population proportion and (3) to an average that better resembles the population, and these three features ensure that results of more repeated samples lead to a better estimate of the population.



Enter a larger number of repetitions

Use the options displayed in scheme B to obtain a clear plot of the data

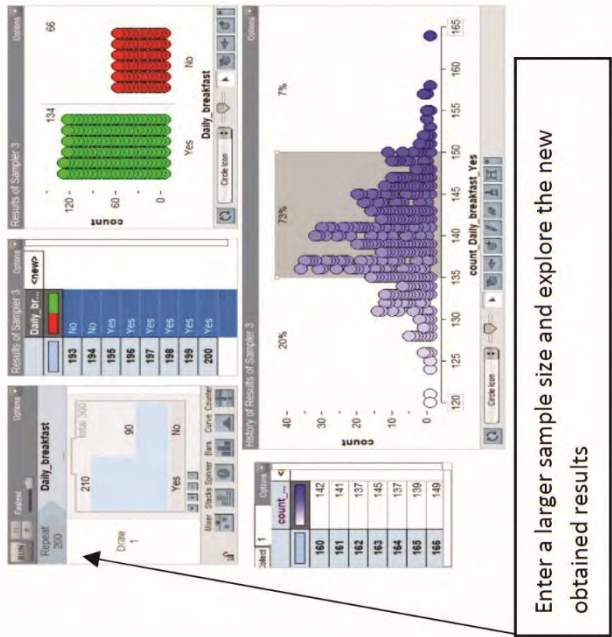
Instrumentation Schemes E: Exploring sample size

Simulate repeated samples at varying sample sizes and interpret the results using the sampling distribution

TinkerPlots techniques

Conceptual Understanding

Results from repeated samples with a larger sample size lead to a better estimate of the population, that is, a larger sample size leads to (1) less variation in the corresponding estimates of the population and (2) to an average that better resembles the population, and these two features ensure that a larger sample size leads to a better estimate of the population.



Instrumentation Schemes F: Interpreting sampling distribution

Interpret the simulated results using the sampling distribution

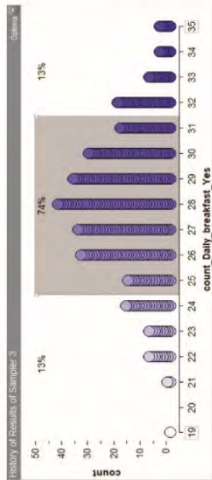
TinkerPlots techniques

Conceptual Understanding



Use the options—for example dividers and counts—to explore particular ranges of sample results

The proportion of sample results belonging to a particular range of results indicates the probability that a random sample result falls into this particular range. In this way, the probability of a given sample result can be interpreted and used to make a statement about a given real-life problem.



Students' average technical levels for each instrumentation scheme on respectively W1 and W5 was calculated to identify students' global development during the work. For example, students' average technical level score on W1 in scheme A was calculated from five students whose actions were coded technical level 1, seven students who scored level 2 and two students on level 3, which resulted in an average technical level score of 1.8. Likewise, we coded students' reasoning and calculated their average conceptual levels for each scheme A to F, while working on W1 and W5. The change in performance on technical and conceptual level from the initial phase in W1 to the more advanced phase in W5 is visualized in Figure 3.2. When comparing students' work on W1 to W5, students showed an improving level of proficiency in their application and control of the tool as well as in their usage and expression of statistical concepts in their accompanying reasoning. As students' development of TinkerPlots techniques and conceptual understanding of statistical modeling was observed simultaneously, the results show a co-development of techniques and concepts.

It is interesting to note that in schemes C and D we observed more progress in students' average conceptual level score than for their technical level score. Both these schemes required more complex TinkerPlots techniques than the other schemes. In the initial phase, concerning these two schemes, most students of group 1 worked carefully according to the TinkerPlots instruction sheet, which enabled them to use the correct techniques. For example, concerning students' TinkerPlots techniques in scheme C during the initial phase with W1, all fourteen students had difficulty using the history option in TinkerPlots to visualize the sampling distribution. They all followed the instruction stepwise. Seven of them made mistakes in their actions—for example, not knowing how to enlarge the history window to enter all required information or not being able to select a useful characteristic for the history option—which made it difficult for them to visualize a correct sampling distribution, and as such, their actions were coded technical level 2. The other seven students made a correct visualization, although they encountered problems with displaying a clear bar chart or entering the correct values, and, as such, their actions were coded technical level 3. Students' reasoning during the initial phase focused on the correct technical actions. For the seven students that were unable to visualize a correct sampling distribution, we also observed incorrect reasoning, that is, wrong statements or incorrectly using words and symbols related to sample, variability, repeated samples and sampling distribution, and we coded their reasoning conceptual level 1. In the data of five

of the seven students that managed to display a correct sampling distribution, we observed superficial but correct reasoning, that is, noticing that the graph looks more or less the same as on the instruction sheet and reading the values on the horizontal axis for common sample result; as such, we coded their reasoning conceptual level 2. The two other students that visualized a correct sampling distribution were in one team. They discussed that the shape of the sampling distribution was not in line with their expectations, as they expected a smooth bell curve. Later in this section we present in detail the work of these two students.

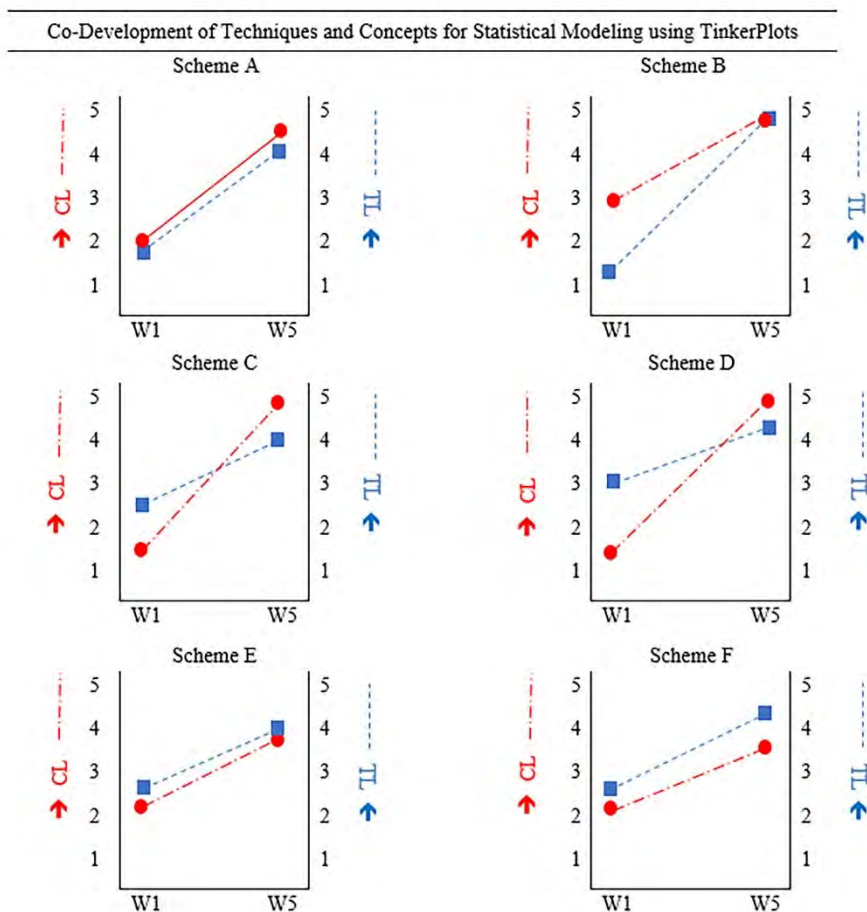


Figure 3.2. Development of students' average Technical Level (TL) and Conceptual Level (CL) for each instrumentation scheme from Worksheet 1 to 5.

In the more advanced phase (W5), with regard to students' technical level in Scheme C, all fourteen students in group 2 displayed a correct sampling distribution and were coded level 3 or higher. Four of them explored a quick start for simulating and adding repeated sample results to a sampling distribution; their actions were coded technical level 5. Concerning students' conceptual level in scheme C with W5, all students' reasoning was coded level 4 or higher, as they correctly stated that more repeated samples led to a smoother shape of the sampling distribution with a peak and average that resembled the modeled population proportion. For example, a student quoted:

This is in line with our expectations. Most of the sample results seem to be in between 43 and 47. This bar at 42 is a bit high [local peak], but yeah, it can happen that within these 100 repeated samples, there are incidentally more with 42...

With regard to the intertwinement of developing techniques and conceptual understanding, based on our findings in scheme C and D, it appeared that for schemes that required complex TinkerPlots techniques, a strong technical focus in the initial phase occurred together with less proficiency on conceptual level, and, additionally, that in the more advanced phase within those schemes, students' statements shifted from discussing techniques to reasoning with concepts, which resulted in more progress for conceptual understanding. In schemes A, B, E and F, we observed a more balanced co-development. Lastly, it is worth mentioning that in the advanced phase most students (10 out of 14) were capable of using the simulated sampling distribution from repeated sampling as a model for determining the probability of a specific range of sample results, and, as such, to interpret the statistical model to solve a given problem.

Two Cases

In this section, we present two cases of students as illustrative examples of how we zoomed in on the observed data to reveal the intertwining of emerging TinkerPlots techniques and conceptual understanding of statistical modeling in the personal schemes students develop. First, we present the case of Elisha and Willie (all student names are pseudonyms) while working on W1, as it illustrates how conceptual understanding influenced TinkerPlots techniques and vice versa in the initial phase. Second, we present the case of William and Brenda while working on W5 in the more advanced phase.

Breakfast worksheet

Introduction: Research at a primary school into the breakfast habits of pupils showed at the beginning of the school year that 210 of the 300 pupils eat breakfast daily. The school management wants to investigate again the number of students having daily breakfast at the end of the school year. However, asking all pupils is a lot of work and therefore they decide to take a sample of 30.

.....

Task 5: Assume that the number of pupils having daily breakfast remained the same during the school year. Use TinkerPlots to simulate sample results (number of pupils having daily breakfast) from the given population and sample size. Fill in the table below, based on the simulated results.

Sample size 30	Simulated sample results in interval notation [...;...]
Most common results	
Exceptionally low results	
Exceptionally high results	

.....

Task 8: In the past school year, a lot of attention has been paid to stimulating 'daily breakfast' at the school. The school management wants to use the results from the sample of 30 to determine whether the breakfast behavior of pupils has improved. Suppose the sample shows that 23 out of 30 pupils have breakfast daily. Can the school management, based on this result, conclude that the pupils' breakfast behavior has improved? Support your answer with the simulated sample results.

Figure 3.3. Tasks 5 and 8 from Worksheet 1

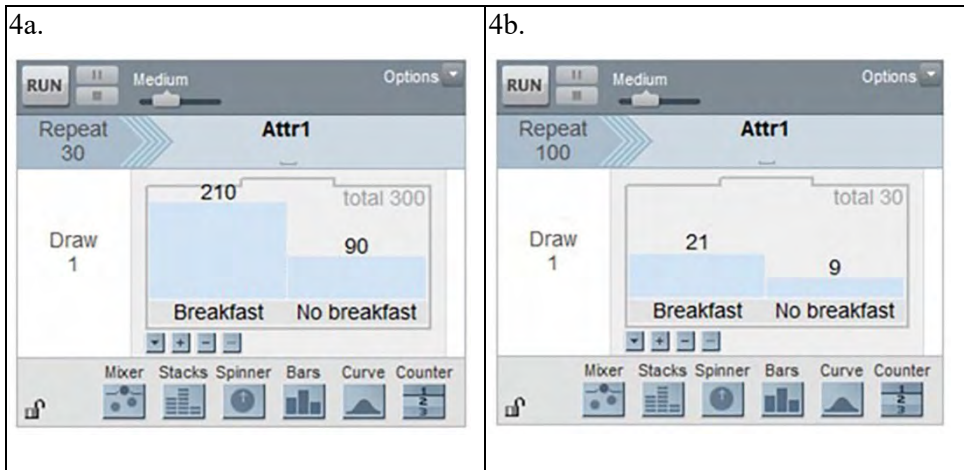


Figure 3.4. Building a population model on Worksheet 1 task 5: the expected model (4a) and Willie and Elisha.s model (4b)

The Case of Elisha and Willie

We focus on Elisha and Willie's work on W1 tasks 5 and 8 (Figure 3.3). We start by highlighting some of their actions and reasoning, followed by an evaluation of how their developing TinkerPlots techniques influenced their conceptual understanding and vice versa. To answer W1 task 5, we expected students in scheme A of the instrumentation scheme to develop TinkerPlots techniques for entering the population characteristics as shown in Figure 3.4a. However, when entering the sample size in scheme B, Elisha and Willie incorrectly entered 100. Later on, when they arrived at scheme F—interpret the results using the sampling distribution—the following discussion in Excerpt 1 took place while the two students were looking at the simulated sampling distribution on their screen (see Figure 3.5a).

[Excerpt 1]

Willie: According to this graph, the sample results vary between 58 and 80 pupils who have breakfast every day [silence]. But... how is this possible? We only have 30 pupils in one sample.....

Elisha: Yes, but we have already filled in 100 [points to the input option 'repeat' on the screen, see Figure 3.4b] and we should have entered 30.

Willie: But why, we do it [simulating repeated samples] 100 times, don't we? We do it 100 times with 30 pupils.

Elisha: Yes, exactly. We repeat it 100 times with 30 pupils. And now, we get for one such thing [points at the visualization of one sample on the screen] a result of 73 pupils who eat breakfast daily and 27 not, that is not correct. So, here [points again to the input option 'repeat' on the screen, see Figure 3.4b], we should have entered the sample size, which is 30, instead of entering 100.

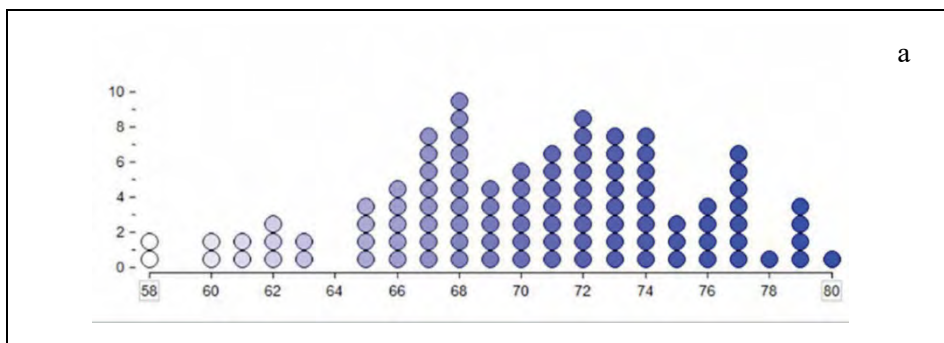
Following this discussion, they deleted their work and started again by entering a population model, but now with a correct sample size of 30. This time, in Scheme C, they entered 100 for the number of repeated samples. This resulted in the simulated sampling distribution of Figure 3.5b. Here, the discussion in excerpt 2 took place.

[Excerpt 2]

Willie: This graph looks weird. What went wrong? Look at all those bumps.

Elisha: Let's do it again [more repeated sampling]. And maybe, we should simulate more than 100 repeated samples. The more, the better, right?

Willie: [After simulating 200 extra repeated samples, their simulated sampling distribution looked like Figure 3.5c]. Yes, that's the way it should look like. Next time we just have to enter more repetitions right away. That's simply the best. So, for now, most of the samples are between 18 and 24.



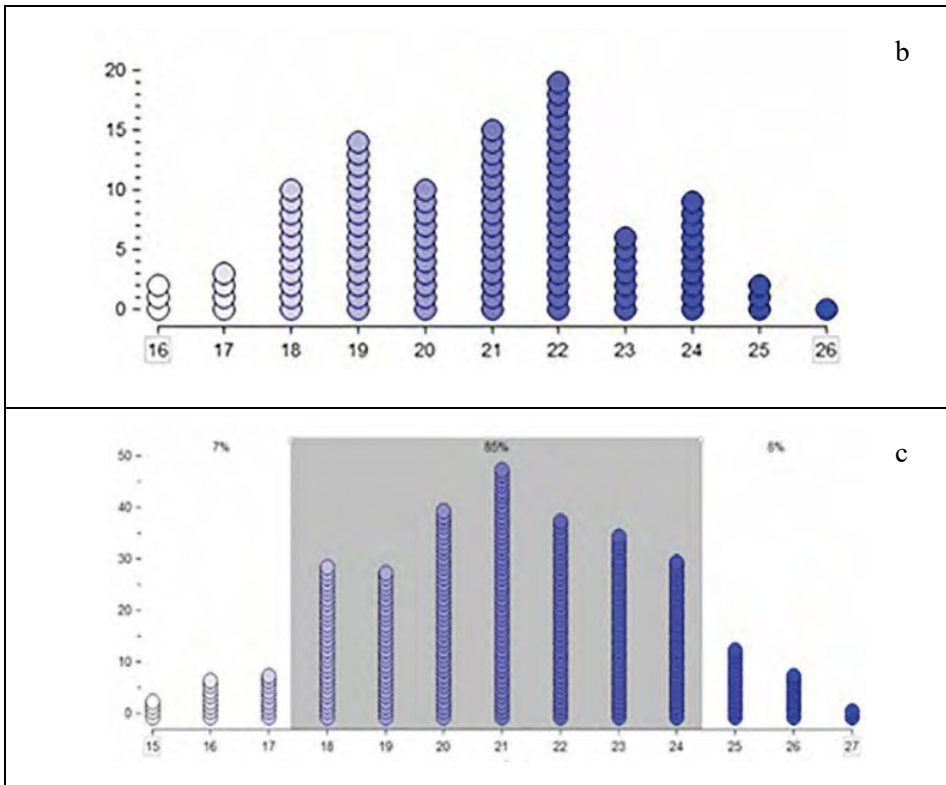


Figure 3.5. Willie and Elisha’s simulated sampling distributions from repeated sampling for Worksheet 1 task 5 (a) Simulated sampling distribution for Worksheet 1 with an incorrect sample size of 100 instead of 30. (b) Simulated sampling distribution for Worksheet 1 with sample size 30 and 100 repeated samples, showing a ‘bumpy’ shape. (c) Simulated sampling distribution for Worksheet 1 with sample size 30 and 300 repeated samples, showing a ‘smooth’ shape.

After the discussion in excerpt 2, they used the simulated sampling distribution (Figure 3.5c) to correctly answer task 5 and 8. For task 5, they stated that most common sample results will vary from 18 to 24 out of 30, they indicated sample results varying from 15 to 17 out of 30 as exceptionally low and results varying from 25 to 27 exceptionally high. They ignored the possibility of sample results below 15 and above 27, probably as these results were not displayed on the x axis of their simulated sampling distribution. For task 8, they stated that 23 out of 30 seemed to be better than 21 out of 30, however, 23 was not exceptionally high and, therefore, the school management could not conclude that the

breakfast habits of pupils have improved. Elisha added that she regarded a sample of 30 as very small in this case.

In summary, the case of Elisha and Willie showed how their conceptual understanding and TinkerPlots techniques co-developed and influenced each other. From excerpt 1, it seems that they mixed up the option in TinkerPlots for entering sample size with entering the number of repeated sampling. When the (incorrect) simulated sampling distribution was displayed on their screen, this sampling distribution did not correspond to their conceptual expectations. The mismatch led them to investigate the options available to see what the problem was, which resulted in applying the correct technical option for entering sample size. Here, their conceptual understanding fostered their technical actions. From excerpt 2, we see how Elisha and Willie used the technical options for repeated sampling to get a better, less “bumpy and smoother” representation of the sampling distribution. The technique of increasing the number of repeated samples helped them understand the effect of more repeated samples by giving them a better picture of the sampling distribution. In this way, the technique of repeated sampling fostered their conceptual understanding of the effect of adding more repeated samples on the sampling distribution in Scheme D.

From excerpt 2, it was difficult to distil the depth of the students’ conceptual understanding about adding more repeated samples in scheme D. Although they stated that they should enter a larger number of repetitions next time, and that a larger number of repetitions would lead to a better graph of the sampling distribution, they did not express clearly how they thought these two were related. However, later on, in W1 task 14, they explicitly mentioned that next time they should simulate a larger number of repeated samples at once in order to reduce the influence of possible outliers and to achieve a well-shaped sampling distribution. Combining students’ statements over several tasks and schemes helped us to identify their understanding of specific concepts.

The Case of William and Brenda

We focus on the work of William and Brenda on W5 tasks 7 and 9 (Figure 3.6). Instead of getting started with TinkerPlots after reading task 7, these two students started a 5-min discussion about possible answers. Excerpt 3 presents a small part.

Worksheet 5: LED lights

Introduction: A do-it-yourself shop is not satisfied with the quality of LED lights they sell. There are too many complaints from customers about defective lights. They are therefore considering a switch to supplier B. This supplier guarantees that at least 90% of the lights will function. The do-it-yourself shop has therefore ordered a batch of 10,000 LED lights. Before selling them in the shop, they use a sample to verify whether the supplier's claim is correct.

... **Task 7:** Suppose there are 42 functioning LEDs in the sample (size 50). What advice would you give the shop about the purchase of the batch of LED lights? Justify your answer.

... **Task 9:** On closer inspection, the shop doubts whether a sample size of 50 would be appropriate. Which sample size would you recommend? Justify your answer.

... **Task 12:** Because the shop has doubts about the quality of the LED lights from supplier B, they ordered a batch of 10,000 from supplier C. They also examine this batch with a sample of 50 to determine whether this batch may be better than the batch of supplier B.

Open the file "LED lights supplier C" with TinkerPlots. Simulate repeated samples from the hidden population model of supplier C. Based on the simulations, estimate the number of functioning LEDs in the total batch of 10,000 LEDs. Justify your answer.

Figure 3.6. Tasks 7, 9, and 12 from Worksheet 5

[Excerpt 3]

William: 42 out of 50, that's not 90%, because then it should be 45, this is not enough. So don't buy it.

Brenda: I agree. 42 is not sufficient. Don't do it.

William: Or... (silence)... it's just a small sample size, only 50. In our earlier social media task with a sample of 50, there was a lot of variation, then 42 is not that unusual.

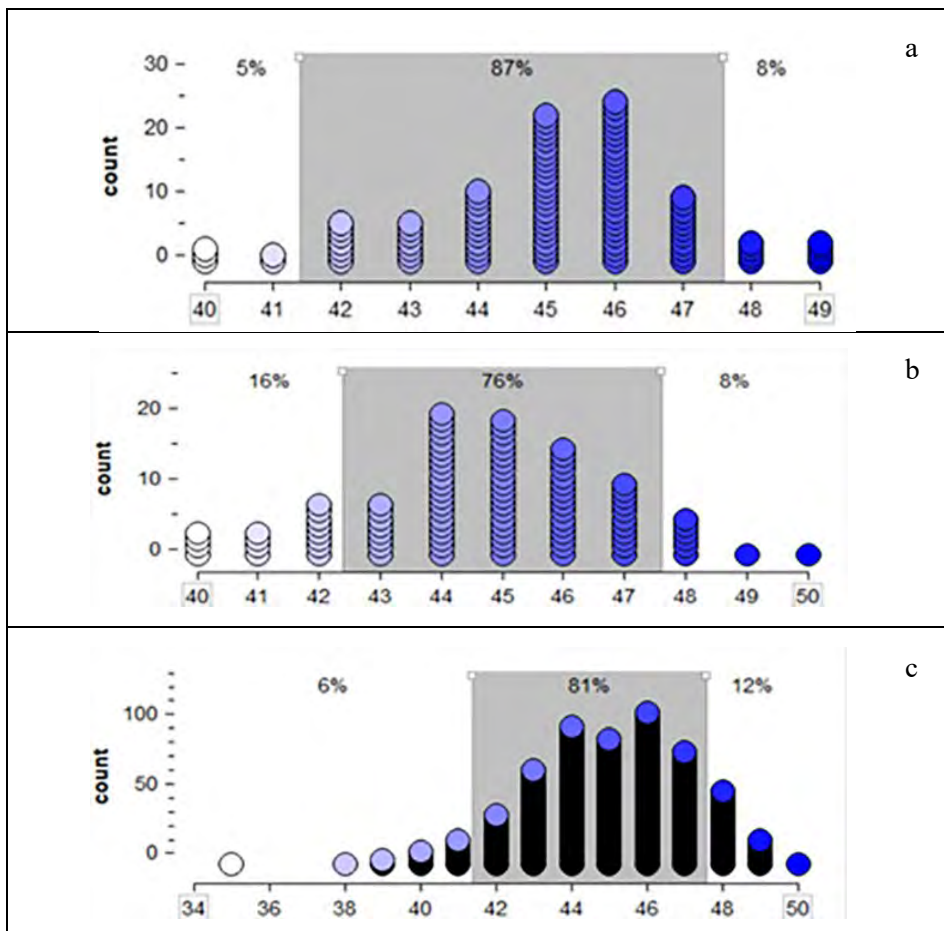


Figure 3.7. William and Brenda’s simulated sampling distributions from repeated sampling for Worksheet 5 task 7. **(a)** Simulated sampling distribution for 100 repeated samples in Worksheet 5 with a left border of the grey area at 5%. **(b)** Simulated sampling distribution for 100 repeated samples in Worksheet 5, second attempt, with a left border of the grey area at 16%. **(c)** Simulated sampling distribution for 500 repeated samples in Worksheet 5 with a left border of the gray area at 5%.

As the discussion progressed, they decided to model the task in TinkerPlots. Without discussing the TinkerPlots techniques, they succeeded within a few minutes and without any hesitations to display the sampling distribution as shown in Figure 3.7a. Their goal was to determine the most common results—in their strategy the middle 80% of the samples—by placing borders for the lowest and highest 10%. After they moved the lower border of the gray area

back and forth a number of times, it turned out to be impossible to get exactly 10% into the left part of the sampling distribution. On that point, the following discussion took place.

[Excerpt 4]

William: This is not a good sampling distribution.

Brenda: How is that?

William: It is not possible to get 10% here [pointing to the left part in the sampling distribution]. It is either 5% or 13%

Brenda: And now what? There's not much we can do with this. Can't we do it again? Then maybe it will be better.

They decided to delete everything and start all over again. This resulted in the sampling distribution of Figure 3.7b. Then the discussion in Excerpt 5 took place.

[Excerpt 5]

William: This isn't much better... now we have 8% or 16%...

Brenda: Let's just do it again.

William: Again? Wait, I think we can do this again faster. We can leave this [points to sub screen 1, 2 and 3] and only have to do the repeats again.

[....]

William: We should have discovered this earlier, that would have saved us a lot of work with the previous worksheets. In fact, we always investigate the same thing, but with a different subject.

Brenda: How do you mean?

William: Well, we investigate possible sample results with a given samples size to answer the questions. It doesn't really matter whether it's breakfast, social media or lights.

This third attempt also resulted in a left area smaller than 10%. At that moment they decided to increase the number of repetitions, as that usually gives a better picture. After William said: "You can probably add samples in a quick way, without starting all over," they explored the techniques and soon found out how to add samples. This resulted in the sampling distribution of Figure 3.7c.

[Excerpt 6]

- William: I don't think there is any point in adding more repetitions, it remains the same. 42 is apparently exactly at the border of common results. And now what?
- Brenda: The sampling distribution hardly changes, so there's no need for more repetition. I think 42 is not much. Most results are higher
- William: Okay, based on these sampling distributions we find 42 to be too few. So our advice is not to buy!

Later on, when they worked on W5 task 9, they fully agreed that the sample size was too small. Brenda stated: “The larger sample the better the results, but very large is not convenient,” at which point William proposed to pick a sample size of 200. As with task 7, they wanted to explore a fast way in which they did not have to remove all the sub screens. To this end, they discussed the views on each sub-screen and finally decided that only sub screen 1 could remain. Here, they discussed concepts such as sample size, difference between sample size and number of repeated samples, and the relationship between the tables and dot plots. When using a fast method for larger sample size, the effect of larger sample size confirmed their conjecture.

In summary, the case of William and Brenda in the more advanced phase showed a focus on conceptual understanding when reading the task, a focus that we saw in almost all students in W5. Excerpt 4 illustrates how the two students, after reading task 7, discussed concepts such as variation, probability and sample size. Moreover, in this excerpt, they related this task to a previous task and context (the context of W3). This also appeared in the second part of excerpt 5, here we saw how the use of similar TinkerPlots techniques in different worksheets and contexts enabled William to discover a context-independent pattern. The TinkerPlots techniques helped him to identify technical patterns in the modeling process and thus to view the concepts involved at a more abstract—context-independent—level. Regarding the intertwined relationship between TinkerPlots techniques and conceptual understanding, excerpt 5 showed how their understanding—in this case their overestimation of variation in many repeated samples—triggered them to explore new techniques. Also, the other way around, how in excerpt 6 the techniques helped them to understand that the sampling distribution of many repeated samples remains stable. Also, their work on W5 task 9 showed, as in

the case of Willie and Elisha, a two-directional relationship between TinkerPlots techniques and conceptual understanding. Their understanding of statistical modeling concerning a general approach and patterns, resulted in a search for more advanced TinkerPlots techniques by using already modeled parts of the process in their sub-screens, and also, the techniques strengthened them in their conjecture about the effect of sample size.

Discussion

The aim of this study was to explore the applicability of the instrumental genesis perspective to statistics education, and to statistical modeling in particular. We identified six instrumentation schemes for statistical modeling processes with TinkerPlots, describing the intertwined development of students' digital techniques and conceptual understanding. We noticed an increase in their mastery of the tool as well as in their statistical reasoning, evidencing students' co-development of techniques and conceptual understanding. We observed a two-directional intertwining of techniques and concepts. The two student cases showed in more detail how students' understanding of concepts informed their TinkerPlots techniques and vice versa. Although we found a two-directional intertwining in all schemes and phases of the teaching sequence, at particular moments we noticed more emphasis in one direction.

In the more advanced phase of the teaching sequence the results show how the use of similar TinkerPlots techniques over different worksheets and contexts enabled students to discover context-independent technical patterns. Students' identification of those technical patterns in the modeling process enabled them to view concepts at a higher, more abstract level. We interpret this as emergent modeling (Gravemeijer, 1999), which involves the conceptual shift from a model of a context-specific situation to a model for statistical reasoning in a variety of similar and new contexts. Although the emphasis here was on technical patterns that informed students' conceptual understanding, we also saw the opposite direction intertwined in this process, as their conceptual understanding concerning a general approach and patterns resulted in a search for more advanced techniques by using already modeled parts of the process on their screen.

In a short period of time, students—who were inexperienced in taking samples and working with digital tools—learned to carry out the modeling processes, including interpreting the simulated sampling distribution. Regarding this promising result, we discuss some possible stimulating factors. As a first factor, it appeared from the identified instrumentation schemes that the required

techniques in the digital environment of TinkerPlots strongly align with key concepts for statistical modeling. This strong alignment probably facilitated students to overcome initial difficulties concerning variability, distribution, sample and sampling, the effect of sample size, difference between the sample and sampling distribution (Chance, delMas, & Garfield, 2004), which we hardly observed in the more advanced phase. For example, concerning the common confusion between the sample and sampling distribution, the distinct visualization of sample and sampling results within the digital environment of TinkerPlots enabled students to distinguish between both distributions. As a second factor, the required TinkerPlots techniques invited students to phenomenon simplification (Manor & Ben-Zvi, 2017). For example, in the initial phase we observed difficulties in distilling sample size and population proportion from the context given for entering the correct model, while these difficulties hardly occurred in the more advanced phase. As a third factor, applying similar statistical modeling processes in TinkerPlots to varying real-life contexts allowed students to distinguish and interact between the model world—using the same digital environment—and the real world using varying contexts (Patel & Pfannkuch, 2018).

The findings presented in this paper should be interpreted in the light of the study's limitations. First, the results of this research are based on students in a classroom laboratory instead of students' regular classroom environment. By conducting the preparatory activities in students' regular classrooms and maintaining the same student teams, lesson design and a familiar teacher, we tried to reduce the influence of the classroom laboratory setting at the university. Second, due to practical reasons we were confined to working with two groups of students, group 1 in the initial phase and group 2 in the more advanced phase of the teaching sequence. Differences between both groups may have affected students' global scheme development. However, the students in both groups performed at a similar level in mathematics and their performances in the preparatory tasks were comparable. The teacher judged the starting level of the two groups to be similar. Third, distilling students' conceptual understanding from their reasoning was challenging. However, by combining the sometimes flawed statements made by the students with their accompanying activities—such as their next action with the tool or their statements later on in their process—we tried to identify their understanding of the concepts. Fourth, we worked with pre-university students, the top 15% achievers in our educational system. Other students may need more time.

Although we focused on statistical modeling processes using TinkerPlots, we consider our findings on the intertwining of emerging digital techniques and conceptual understanding applicable to the broader field of statistics education, and to other educational digital tools as well. Digital tools for other areas in statistics education also structure and guide students' thinking by providing specific options for entering parameters and commands and/or by facilitating explorative options that may strengthen students' conceptual understanding.

Overall, we conclude that the perspective of instrumental genesis in this study proved helpful to gain insight into students' learning from and with a digital tool, and to identify how emerging digital techniques and conceptual understanding intertwine.

Implications

The study's results lead to implications for the design of teaching materials and digital tools, and for future research. In designing teaching materials, it is important to take into account the two-directional relationship between emerging digital techniques and conceptual understanding, both during instruction and during practice. Attention to digital techniques in the initial phase, especially to more complex ones, seems to have a positive effect on learning the associated concepts later on. The development of context-independent techniques and concepts requires sufficient time and practice for students with different contexts and situations. In designing digital tools, the intertwined relationship between digital techniques and conceptual understanding calls for attentive consideration of how the digital techniques are related to the concepts, to deploy the digital tool in a productive way for the intended learning goal.

This also suggests an implication for future research on statistics education using digital tools. Although we focused on statistical modeling using TinkerPlots—that is, solving real-life problems by the building, application and interpretation, of the sampling distribution of repeated samples—we assume our global findings also hold for other statistical processes and digital tools. However, the specific intertwining of emerging digital techniques and conceptual understanding is unique for each digital tool and intended learning goal. To identify the specific intertwinement, we recommend using the perspective of instrumental genesis in analyzing video and conversation data, which can be added by using clinical interviews.

On a final note, this study gave an insight into the applicability of the instrumental genesis perspective in the context of statistics education, and

statistical modeling with digital tools in particular. Instrumental genesis seems a fruitful perspective to design technology-rich activities and to monitor students' learning.



Introducing Statistical Inference: Design of a Theoretically and Empirically Based Learning Trajectory

This chapter is based on

Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted).
*Introducing Statistical Inference: Design of a Theoretically and Empirically
Based Learning Trajectory.*

Abstract

This paper comprises the results of a design study that aims at developing a theoretically and empirically based learning trajectory on statistical inference for 9th-grade students. Based on theories of informal statistical inference, an 8-step learning trajectory was designed. The trajectory consisted of two similar four step sequences: (1) experimenting with a physical black box, (2) visualizing distributions, (3) examining sampling distributions using simulation software, and (4) interpreting sampling distributions to make inferences in real-life contexts. Sequence I included only categorical data and Sequence II regarded numerical data. The learning trajectory was implemented in an intervention among 267 students. To examine the effects of the trajectory on students' understanding of statistical inference, we analyzed their posttest results after the intervention. To investigate how the stepwise trajectory fostered the learning process, students' worksheets during each learning step were analyzed. The posttest results showed that students who followed the learning trajectory scored significantly higher on statistical inference and on concepts related to each step than students of a comparison group ($n=217$) who followed the regular curriculum. Worksheet analysis demonstrated that the 8-step trajectory was beneficial to students' learning processes. We conclude that ideas of repeated sampling with a black box and statistical modeling seem fruitful for introducing statistical inference. Both ideas invite more advanced follow-up activities, such as hypothesis testing and comparing groups. This suggests that statistics curricula with a descriptive focus can be transformed to a more inferential focus, to anticipate on subsequent steps in students' statistics education.

Keywords

design based research, learning trajectory, simulating repeated samples, statistical inference, TinkerPlots

Introduction

Statistical inference is at the heart of statistics, as it provides a means to make substantive evidence-based claims under uncertainty when only partial data are available (Makar & Rubin, 2018, p. 262). Interpreting inferences with associated uncertainty is difficult for students, which is why, in most countries, inferences are not taught until Grade 10 or higher. Students' difficulties in learning inferences mostly relate to limited understanding of key statistical concepts, such as sample, variability and distributions, and to problems with understanding complex formal procedures (Castro Sotos et al., 2007). Engaging in activities that involve informal inferences in the early years, within primary education or early years of secondary school, seems to facilitate learning about more complex inferential statistics later on (Makar & Rubin, 2009; Van Dijke-Droogers, Drijvers, & Bakker, 2020).

However, the pre-grade-10 statistics curriculum in most countries, including the Netherlands, focuses on descriptive statistics without paying attention to inferences—with the exception of for example New Zealand, where a full learning line including inferential activities was developed starting from primary school. Promising results for informal inferential activities encourage investigating how these can be embedded in current curricula with a descriptive focus. Within most mathematical curricula, only limited time is available for statistics. As such, we need efficient learning trajectories, and knowledge about crucial steps in such a trajectory.

In this study, we use knowledge from literature on (informal) statistical inference, and apply knowledge on learning progressions to design and evaluate an innovative learning trajectory (LT) on introducing 9th-grade students in the pre-university stream—the 15% best performing students of the Dutch educational system—to the key concepts for statistical inference. Following Duschl et al. (2011), we will address the following aspects: how the design process included the selection of the core idea of the LT; how theories on statistical inference inform the design of the LT; the identification of the starting and end point of the LT; how the successive learning steps of the LT mediate learning; and how the LT aligns with current curricula. To empirically verify the effects of the designed LT, we implemented the LT at five different Dutch schools with eleven participating teachers and a total of 267 students. We analyzed both students' performance after the intervention, and their progress during the learning process.

Theoretical Background

Statistical Inference

Statistical inference concerns interpreting sample results, drawing data-based conclusions, and reasoning about probability. For students, it is difficult to understand formal procedures to substantiate their inferences. Many difficulties involve a poor understanding of the key statistical concepts: sample, variability and distributions. These key concepts, including the understanding of the effect of sample size and the idea that a sample characteristic—such as mean or median—can be used to compare distributions, are essential for understanding inferences (Bakker, 2004; Chance, delMas, & Garfield, 2004; Konold & Pollatsek, 2002; Saldanha & Thompson, 2002; Watson & Kelly, 2008). There is a strong relationship between these concepts: understanding the sampling distribution relies on understanding the key concept of a sample, in particular on understanding the balance between sample representativeness and sample variability (Batanero et al., 1994). Common misconceptions involve neglecting the effect of sample size on the variance of sample mean or sample proportion (Tversky & Kahneman, 1971). Another common difficulty involves probabilistic reasoning, as students tend to provide deterministic explanations and not to consider the variability involved (Rossman, 2008).

To help students overcome difficulties involved in statistical inference, informal approaches have been sought in recent decades. In general, this informal approach focuses on making inferences about unknown populations based on observed samples without using formal techniques, such as hypothesis testing. Makar and Rubin (2009) define informal statistical inference in main principles: generalization beyond data, data as evidence for these generalizations, and probabilistic reasoning about the generalization. Informal inferences include data-based claims that go beyond the collected data, in which the uncertainty involved can be expressed in informal probabilistic reasoning about the likelihood of the claim. Offering informal activities at an early age—before the more formal activities in Grade 10 or higher—facilitates the understanding of key concepts and probabilistic reasoning required for statistical inference (Paparistodemou & Meletiou-Mavrotheris, 2008; Van Dijke-Droogers et al., 2020).

The Design of a Learning Trajectory

The design of an LT entails a conjectured route through a set of educational activities to support students to achieve the intended learning goals. Although learning is a personal process, unique for each student, a conjectured LT intends to describe a “possible taken-as-shared learning route for the classroom

community” (Gravemeijer et al., 2003, p. 52); a learning route needs empirical validation. Successful implementation of theory in educational practice involves the design and evaluation in real classrooms of powerful LTs that embody our present understanding of effective learning (De Corte, 2000).

The theory of Realistic Mathematics Education (Cobb, 2011; Freudenthal, 1983) provides design heuristics for the development of learning activities in an LT. First, the learning activities should be set in a context that enables students to immediately engage and develop associated mathematical concepts. As such, the learning activities support students in progressing towards a toolkit of key concepts associated with the learning goals of the LT. Second, the activities should be structured to support students in developing *models* of their concrete mathematical activity that can be used as *model* for a network of mathematical objects and relationships (Gravemeijer, 1999; Streefland, 1991).

The Current Study

The study is part of a larger study to gain knowledge about a theoretically and empirically based learning trajectory to introduce 9th-grade students to the key concepts of statistical inference. From another study (Van Dijke-Droogers, Drijvers, & Bakker, submitted) on the overall effects of the LT on students’ statistical literacy, we know *that* the LT had a significant positive effect on students’ understanding of statistical inference as measured by comparing pre- and posttest results. In the study reported here, we want to know *how* students learned something about statistical inference in terms of the intended LT-step related learning goals of the trajectory. When it comes to experimental studies that only report pre-post results, a common concern is that the reader may still not know how to benefit from the intervention reported (Savelsbergh et al., 2016). We therefore consider it worth spelling out in more detail the design of the 8-step LT, and its effects on students’ understanding of LT-step related goals for statistical inference and analyze students’ progression *during* the large-scale intervention. As such, we address the following research questions:

What are the specific effects of the designed LT on students’ understanding of statistical inference, in terms of the intended LT-step related learning goals?

How do the designed LT steps foster students’ learning processes?

Methods

The designed LT aims at introducing students to key concepts of statistical inference by using theories of informal statistical inference. We first outline the design of the LT. We incorporated two main ideas: repeated sampling with a black box and statistical modeling with a digital tool. Second, we describe the intervention characteristics and data analysis.

An Outline of the LT

This study comprises the results of a third cycle of design based research. During cycle 1 and 2, the LT was (re)designed, implemented and evaluated, to identify the feasibility of the LT, and to further define the starting and ending points of the LT.

The design of the LT consists of two similar sequences of four learning steps. Sequence I concerns only categorical data and includes the following steps: (1) experimenting with a physical black box, (2) visualizing distributions, (3) examining sampling distributions using simulation software, and (4) interpreting sampling distributions to make inferences in real-life contexts. In Sequence II, following Rossman (2008), more complex numerical data are addressed during LT steps 5 to 8. The first three steps of Sequences I and II involved 45 minutes each. In the last step of Sequences I and II, three different real-life contexts were offered with a time duration of 45 minutes per context. An outline of each LT step including a brief description, examples of learning activities, and the intended learning goals, is presented in Table 4.1. A more detailed description can be found in Supplementary Material A.

Repeated Sampling with a Black Box

Repeated sampling with a black box serves as a guiding activity through all steps of the LT. A black box refers to a box of which only part of the content is visible—for example, a box with a viewing window that is filled with marbles or a box filled with notes (see the pictures in Table 1 at LT Steps 1 and 5, respectively). The black box activities instantiate design heuristics of Realistic Mathematics Education (Cobb, 2011; Freudenthal, 1983). Starting within the engaging context of a physical black box experiment—in both Sequences I and II—enables students to immediately involve and orient towards developing key statistical concepts (Van Dijke-Droogers et al., 2020). In Sequence I, activities with a physical black box filled with marbles in LT steps 1 and 2 enable students to explore the sampling variability involved in repeated sampling. Varying the size of the viewing window in the physical black box activities allows students to explore the effects of sample size. These activities



incorporate ideas of the growing sample task (Bakker, 2004) and repeated sampling that make key statistical concepts more accessible for students (Van Dijke-Droogers et al., 2020). Specifically, when those activities are accompanied by classroom discussions for exchanging and comparing sample results (Wild & Pfannkuch, 1999). The idea of repeated sampling with a physical black box is extended in statistical modeling activities in LT steps 3 and 4. In Sequence II, the activities evolve in a similar way from starting with a physical black box filled with notes in LT steps 5 and 6 to statistical modeling in LT steps 7 and 8.

Statistical Modeling with Digital Technology

Statistical modeling activities with educational digital tools facilitate—on an informal level—the exploration of key concepts for statistical inference (Biehler et al., 2013; Garfield et al., 2015; Manor & Ben-Zvi, 2015; Rossman, 2008; Saldanha & Thompson, 2002; Watson & Chance, 2012). Digital environments such as TinkerPlots provide opportunities to easily simulate and visualize (repeated) samples. In the designed LT, the statistical modeling activities start within the familiar context of a black box, where students build a model of a black box—for example filled with 200 red and 400 blue marbles—to simulate sample results. By visualizing sample and sampling distributions, at varying sample sizes and at varying number of repeated samples, students explore (un)likely sample results. The modeling activities within the black box context gradually evolve to modeling real-life contexts. Modeling activities include building a model, simulating (repeated) samples, visualizing and interpreting the results, to solve a given problem. As with the physical black box activities, these modeling activities attend all stages of the statistical investigation cycle several times, as students collect data, analyze their data using sample and sampling distributions, and interpret the results to answer the question posed. Subsequent modeling activities involve applying gained knowledge into new contexts, where students deploy modeling activities to solve real-life problems.

Applying similar digital techniques within varying contexts encourages students to identify context-independent patterns of technical actions (Van Dijke-Droogers, Drijvers, & Bakker, in press). These context-independent technical patterns combined with a context-independent understanding of key statistical concepts, facilitate the conceptual shift from a model of to a model for, known as emergent modeling (Gravemeijer, 1999; Streefland, 1991). As such, statistical modeling enhances the use and understanding of context-independent statistical models, which is essential for interpreting inferences.

Table 4.1. Overview of Steps 1–8 of the Learning Trajectory

LT Step	Description	Example of activities	Learning Goal	Construction of LT steps
<i>Categorical data</i>				
1. Experimenting with physical black box	Physical black box with <i>marbles</i> experiment (with small and large viewing window)	 Estimate the number of yellow marbles in a black box filled with 1,000 marbles (<i>balletjes</i> in Dutch) by shaking and observing visible marbles	Students draw inferences and become accessible to concepts as sample, sample size, sampling variability, frequency and measures of center and spread, within the context of a physical black box	Students experience that sample results vary and that a larger sample size and more repeated samples lead to a better population estimate. Next question: What happens when we further increase the size and number of repeated samples? Conducting larger and more samples is time consuming: a thought experiment can help.
2. Visualizing distributions	Graph as a model (or visualization) of the frequency distribution from repeated sampling with the black box	 Make a sketch of the frequency distribution you expect when the black box experiment with a large viewing window is repeated 100,000 times	Students can draw the visualization of an expected sampling distribution from repeated samples. Students interpret sampling distributions given to make inferences about a certain range of sample results	The sampling distribution from repeated sampling can be used to determine the probability of certain sample results. Next question: How can we get the sampling distribution of repeated sampling in a quick and easy

way? Using technology can help.

Statistical modeling—including interpreting the sampling distribution from repeated sampling—can be used to determine the probability of certain sample results, within the context of the black box. Next question: Can statistical modeling be used more generally, in other situations and contexts?

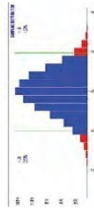
From steps 1 to 4 emerges the question of how to use statistical modeling with other—not categorical—data.

Students use statistical modeling within the digital environment of TinkerPlots to determine (un)likely sample results, within the context of a black box [Statistical modeling includes building a model, simulating (repeated) samples, visualizing the sampling distribution and interpreting the results]

Using simulation of repeated samples, from a modeled black box, in a sampling distribution as a model for interpreting probability

Use TinkerPlots to determine the most common sample results for a black box filled with 750 yellow and 250 orange marbles, and sample size 40

3. Modeling a black box (ICT)



Build and run a model of a real-life situation in TinkerPlots and use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real-life situation and the probability involved.

Use TinkerPlots to determine most common sample results when a sample of 30 is taken from a school with 300 students to determine the number of students having daily breakfast (given that on average 70% of the students have breakfast daily)

4. Modeling real-life contexts (ICT)



Numerical data

5. Experimenting with physical black box



Physical black box with notes experiment. (The box is filled with 4,000 notes. Each note contains information about one students' gender and height, for example boy—155 cm)

Take a sample of 40 notes and summarize the sample data found (calculate measures of center and spread, use a visualization).

Estimate the gender (proportion) and height (center and spread) of the 4,000 students

Students draw inferences within the context of the physical black box with notes (students' gender and height) considering sample size, sample variability, and measures of center

Students discussed how to use numerical data from repeated samples to draw inferences about the population. Next question: how can the population distribution at stake—the content of the black box filled with 4,000 notes on students' gender and height—be visualized based on the varying sample results found?

6. Visualizing distributions



Summarize and visualize the expected population (height of 4,000 students) based on the sample data found in LT step 5

Sketch the frequency distribution you expect for the whole population, based on the sample results found in step 5

Students draw a visualization of the population distribution they expect from the sample results found. Student draw inferences about the population, considering distribution, mean, sample variability and probability

Students draw inferences about the population mean and population distribution using samples found. Next question: what are the effects of larger and more repeated samples on

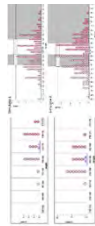
the estimate of the population mean and distribution? Using technology can be helpful to explore the effects.

Students use statistical modeling within the digital environment of TinkerPlots to determine (un)likely sample results, within the context of the black box with notes [Statistical modeling includes simulating (repeated) samples from a *given* model, visualizing the sampling distribution for the sample *mean*, and interpreting the results]

Use TinkerPlots to determine most common sample results—and extraordinary high/low results—from the (given) modeled black box of step 5

Experimenting with simulations of repeated samples (using the mean) at varying sample sizes and number of repetitions, from the modeled black box with notes of LT step 5

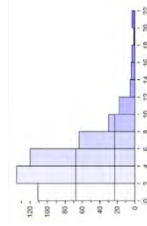
7. Modeling a black box



From step 7 emerges the question of how to apply statistical modeling with numerical data in other contexts and situations.

Run a model of a real-life situation in TinkerPlots and use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real-life situation.

8. Modeling real-life contexts



Students use statistical modeling within the digital environment of TinkerPlots to make inferences, within the context of a real-life problem

Use TinkerPlots to simulate repeated samples (size 200) from a hidden dataset of 4,000 students to determine the time students spent on sport

Participants

Eleven teachers participated in the intervention, with a total of 267 students (Grade 9, aged 14–15 years) from thirteen classes at five different schools. The teachers were trained for the intervention in two 3-hour sessions in which they worked through students' lessons and materials themselves, guided by the researcher. The teachers decided to replace all regular 9th-grade statistics lessons with the LT to save time. The students had no experience with using digital tools during their mathematics lessons. Students were instructed in using TinkerPlots in LT step 3 through a demonstration by a teacher and they received an instruction sheet for modeling black boxes that they could use during LT steps 4 to 8. They had some basic knowledge of descriptive statistics: center and distribution measures, such as mean, quartiles, class division, absolute and relative frequencies, and boxplot. A comparison group with students who followed the regular curriculum was used to interpret students' performance on statistical inference. The comparison group consisted of 217 students from ten classes. All students in the comparison group attended 10–16 regular 9th-grade statistics lessons during their mathematics lessons. The participating students, for both the intervention and comparison group, belonged to the 15% best performing students in our educational system.

Data Collection and Analysis

For Phase 1, addressing the first research question, we developed a pre- and posttest for Statistical Inference (SI) at the school level, inspired by Watson and Callingham's (2003, 2004) work on statistical literacy. The pre- and posttest can be found in Supplementary Material C and D. Both tests were part of a broader study on the effects of the designed LT on students' statistical literacy (Van Dijke-Droogers et al., submitted). For the study presented here, we focused on the SI Items of the posttest. The posttest contained 18 SI Items. We selected four Items from Watson and Callingham (2004), and we designed 14 new Items related to concepts of SI as addressed in the LT. For the design of the new Items, we used the structure and phrasing of their Items. To analyze the validity of the designed test, we conducted two pilot tests in different classrooms, each consisting of 25 students. Concerning the concurrent validity of the new designed SI Items, students' average level scores in the pilots on new designed and existing SI Items were not significantly different ($M_{\text{new}} = 2.49$, $SD_{\text{new}} = 0.71$, $M_{\text{ex}} = 2.78$, $SD_{\text{ex}} = 1.38$, $n = 50$, $t(49) = -1.6$; $p = .11$). To assess the content and construct validity of test Items, the results of each pilot were used for in-depth discussion with experts in this area on content, construct, vocabulary, and clarity. Cronbach's alpha value was .81, indicating a good reliability (Taber, 2018). For the data collection, the participating teachers from

both the intervention and comparison group conducted the test, according to a clear instruction for testing, from their own students during their regular 45-min mathematics lessons.

For the data analysis in phase 1 on the posttest results, we defined six SI levels, based on Watson and Callingham's levels for statistical literacy (see Table 4.2). Given that LT steps 1 to 4 and 5 to 8 involve similar concepts and approaches, we defined specific levels for couples of two: steps 1 & 5, steps 2 & 6, steps 3 & 7, steps 4 & 8 (see Supplementary Material B). By pairing the LT steps, we were able to analyze at least four test Items per couple. For the coding, we developed Item-specific level-codes (e.g., Figure 4.2 and 4.3). Two assessors coded test data from the participating students with the SI level scores 0–6. To indicate students' performance on the test, we compared students' test scores for both the intervention and comparison group, and as such, for attending the LT or regular statistics curriculum. Students' results on the pretest were used to identify students' initial level. Although the comparison group attended the regular statistics lessons prior to the pretest, we conjectured similar pretest results for both groups on statistical inference as the regular lessons only concerned descriptive statistics. For statistical significance, we used one-way ANOVA for comparing results from both groups, paired *t* test for analyzing students' progression between the pre- and posttest, and chi-squared test for comparing students' distribution over the levels. For reliability of the analysis, a third coder was asked to process independently a random set of 5% (80 Items) of the data with students' reasoning. The third coder agreed on 83% of the codes. Deviating codes, which were limited to one or two levels difference at most, were discussed until agreement was reached. Adjustments in the coding were also applied to the rest of the data.

Table 4.2. Levels for Statistical Inference based on Levels for Statistical Literacy by Watson and Callingham (2003)

Level	General level description
1 Idiosyncratic	Idiosyncratic engagement with context, tautological use of terminology
2 Informal	Only colloquial or informal engagement with context often reflecting intuitive non-statistical beliefs, single element of complex terminology and setting, and basic one-step table and graph readings and calculations, not referring to statistical information given

3	Inconsistent	Selective engagement with context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas, not always referring to statistical information given
4	Consistent Non-critical	Appropriate but non-critical engagement with context, multiple aspects of statistical terminology usage, and statistical skills associated with simple probabilities, and graph characteristics, not always referring to statistical information given
5	Critical	Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, appreciation of variability, explicitly referring to statistical information given
6	Critical Mathematical	Critical, questioning engagement with context, using proportional reasoning, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language, explicitly referring to statistical information given

In phase 2, addressing the second research question, we used two principles by Wilson (2009) for assessing learning progression. The first principle outlines a developmental perspective regarding the development of students' understanding of particular concepts and skills over time—that is, during the LT instead of assessing final performance. This perspective requires clear definitions and a theoretical framework of what and how students are expected to learn. In our study, these are embedded in the description of the designed 8-step LT. The second principle involves the match between the LT and assessment. To establish a strong match, we formulated indicators for success of each LT step. In the design of the learning activities on students' worksheets, specific tasks were included that correspond to these indicators. Table 4.3 displays the indicators and corresponding learning activities on students' worksheet, for each LT step (see Table 4.1 for corresponding learning goals in each LT step).

Data included students' worksheets 1 to 8 from each LT step, accompanied by teachers' and researchers' notes. We collected 267 worksheets

from Sequence I, LT step 1 to 4, and 224 worksheets from Sequence II. The teacher took notes about each lesson. After each lesson, the researcher contacted the teacher—through email, call or a meeting in person—to evaluate the lesson given and to discuss the following steps. In addition, we used researchers’ observation data from two visits in each class about how the teacher and students interacted with the intervention materials. For the data analysis in Phase 2, we coded students’ reasoning on their worksheets, for the specific tasks in each LT step, according to the indicators. Students were explicitly asked to clearly motivate their answers on their worksheets. Teachers’ and researchers’ notes were included in the analysis.

Table 4.3. Overview of Indicators and Corresponding Learning Activities on Students’ Worksheet, for each LT step

LT step	Indicators	Task description per indicator [Worksheet Task]
1.	a. Making inferences about content physical black box	a. Students make inferences about the content of the physical black box using a small and large viewing window [W1.3; W1.6]
	b. Interpreting effect of larger viewing window	b. Students mention that an inference based on a larger viewing window is more reliable as it provides more information about the content [W1.8]
2.	a. Drawing expected sampling distribution from repeated samples	c. Students draw the expected sampling distribution from 100,000 repeated samples, with sample size 40, from a black box filled with 250 yellow and 750 orange marbles [W2a]
	b. Using (given) sampling distribution to determine the probability of sample results	d. Students use a given sampling distribution from 1500 repeated samples (size 50) to determine the probability of a certain range of sample results [W2b.5]
3.	Using statistical modeling in TinkerPlots to determine the probability of sample results	e. Students determine most likely sample results for a black box filled with 300 orange and 200 yellow marbles and samples size 50, using statistical modeling in TinkerPlots [W3.15]
4.	Using statistical modeling in	

TinkerPlots for	
a. Interpreting effect of sample size in real-life contexts	a. Students argue that it is a smart decision of the school management to take a larger sample size [W4.10]
b. Probabilistic reasoning in real-life contexts	b. Students argue that the school management cannot be certain about the breakfast habits of students, based on a sample result [W4.11]
c. Determining the probability of sample results, in real-life contexts	c. Students use their simulated sampling distribution to determine the probability of (un)likely sample results [W4.18]
d. Informal hypothesis testing	d. Students determine at what sample results a school can conclude that the breakfast habits of students have improved, using statistical modeling in TinkerPlots— informal hypothesis testing [W4.18]
5. Making inferences about content physical black box	Students make inferences about the height of the population based on samples from a physical black box filled with 4,000 notes— each note contains the height and gender of one person [W5]
6. Drawing expected population distribution	Students draw a visualization of the population distribution (height of 4,000 persons in the physical black box with notes) they expect, based on the sample data found [W6]
Using statistical modeling in TinkerPlots (<i>given</i> model) for	
7. a. Making inferences about the population distribution - using a small sample size - using a large sample size	a. Students sketch the expected population distribution (height of 4,000 persons in the physical black box with notes) using statistical modeling in TinkerPlots with a <i>given</i> model for varying sample sizes [W7.1; W7.8]
b. Interpreting effect of sample size on expected population	b. Students mention that a larger sample size better reflects the population distribution [7.15]

	distribution	
	c. Making inferences about the population mean	c. Students make inferences about the expected population mean [W7.1; W7.8]
	d. Interpreting effect of sample size on the expected population mean	d. Students mention that a larger sample size leads to a better estimate of the population mean [7.16]
	e. Determining the probability of sample results (concerning the sample mean)	e. Students determine the probability of certain sample results [W7.6; W7.13]
8.	Using statistical modeling in TinkerPlots to determine the probability of sample results, in real-life contexts	Students make inferences about the population proportion of students that spent more than 12 hours per week on sports, using statistical modeling with a <i>hidden</i> model of the population (size 4,000) and sample size 500 [W8.5; W8.6]

Results

We first present students' results on the posttest to answer research question 1. Next, we present students' progress during the intervention to address research question 2.

Posttest Results on Students' Understanding of Statistical Inference

With regard to students' Statistical Inference (SI) level at the posttest, we reported in another study (Van Dijke-Droogers et al., submitted) that a one-way ANOVA between both groups indicated that the level score for the intervention group who attended the LT was significantly higher than for the comparison group ($+0.67$, $F(1, 482) = 75.0$, $p < .0005$). The results in the study presented here indicate that the intervention group scored significantly higher than the comparison group on each coupled LT steps 1 and 5 on using samples, LT steps 2 and 6 on visualizing distributions, LT steps 3 and 7 on repeated sampling and effect of sample size, and LT steps 4 and 8 on solving real-life problems. The results are displayed in Table 4.4. Although we conjectured a similar pretest score for both groups, the results showed that the initial level of the intervention group on statistical inference was significantly lower than for the comparison group—probably because the comparison group followed their (descriptive)

statistics lessons prior to the pretest. The comparison group was not taught statistics between the pre- and posttest, which explains their similar scores on SI at both tests.

Table 4.4. Students' Mean Level Scores on the coupled LT steps at the Pre- and Posttest

		Intervention (n = 267)	Comparison (n = 217)	Intervention minus Comparison
		M (SD)	M (SD)	M(inv.) – M(comp.)
Pretest	SI ¹	2.45 (0.65)	2.72 (0.71)	– 0.27***
	Step 1 & 5	2.10 (1.34)	2.43 (1.41)	– 0.33**
	Step 2 & 6	2.54 (0.91)	2.77 (0.96)	– 0.23**
	Step 3 & 7	2.48 (0.68)	2.75 (0.66)	– 0.27***
	Step 4 & 8	2.62 (0.94)	2.83 (0.92)	– 0.21*
Posttest	SI ¹	3.34 (0.84)	2.67 (0.84)	+ 0.67***
	Steps 1 & 5	3.52 (1.26)	2.94 (1.26)	+ 0.58***
	Steps 2 & 6	3.44 (1.31)	2.84 (1.42)	+ 0.60***
	Steps 3 & 7	2.39 (1.04)	1.85 (0.97)	+ 0.54***
	Steps 4 & 8	3.65 (0.97)	2.91 (1.00)	+ 0.74***
Progress Pre to Post	SI ¹	+ 0.89 (0.92)***	– 0.04 (0.71)	+ 0.93***
	Step 1 & 5	+ 1.42 (1.71)***	+ 0.52 (1.57) ***	+ 0.90***
	Step 2 & 6	+ 0.91 (1.50) ***	+ 0.06 (1.48)	+ 0.85***
	Step 3 & 7	–0.09 (1.15)	– 0.89 (1.00) ***	+ 0.80***
	Step 4 & 8	+ 1.04 (1.18)***	+ 0.08 (1.05)	+ 0.96***

*** $p < .0005$, ** $p < .005$, and * $p < .05$

¹Main results for SI (Chapter 5)

Mrs. Jones wants to buy a new car, either a Honda or Toyota. She wants whichever car will break down the least. She read in Consumer Report that for 400 cars of each type, the Toyota had more breakdowns than the Honda. She talked to three friends. Two were Toyota owners, who had no major breakdowns. The other friend used to own a Honda, but it had lots of breakdowns, so he sold it. He said he'd never buy another Honda.

Which car should Mrs. Jones buy? Explain your answer

Level	Code	Description
5	3	Honda based on larger sample size, admitting uncertainty
1	2	Doesn't matter due to uncertainty Honda, without mentioning sample size
1	1	Toyota, because of her friends' experiences
0	0	Other

Group	Average score	Percent of students per level		
		Level 5	Level 1	Level 0
Comparison group	1.87	5.9%	71.0%	23.1%
Intervention group	3.54	63.7%	35.6%	0.7%

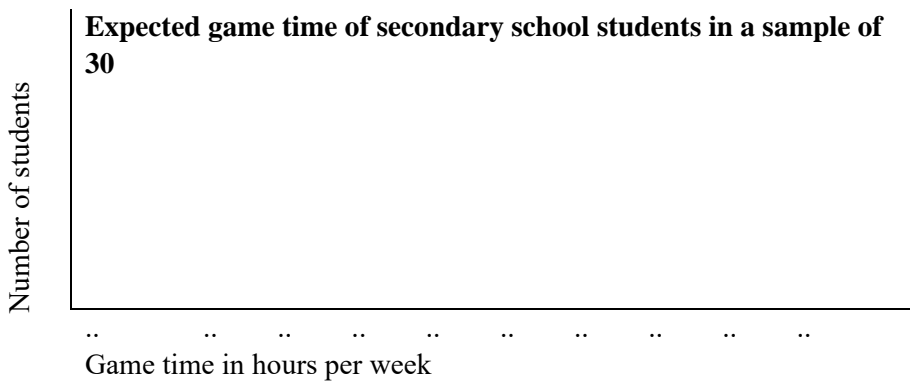
Figure 4.1. Students' achievements on posttest Item 1, taken from Watson and Callingham (2004)

We now elaborate on three posttest Items for which the results of the intervention and comparison group were quite different. The first Item is from Watson and Callingham (2004) and the second and third are newly designed Items. First, we present the results for posttest Item 1 (see Figure 4.1). Most students of the intervention group (63.7%) based their advice on data from research by Consumer Report among 400 participants, and only the minority based their advice on the personal experiences of Mrs. Jones' friends (35.6%). However, in the comparison group, we observed an inverse situation. Here, the majority of the students based their advice on the experiences of the friends (71.0%), and only a few students based their opinion on the Consumer Report survey (5.9%). A chi-squared test on the distribution over levels in percentages between both groups, confirmed a significantly higher score for the intervention group ($\chi^2(2) = 80.84, p < .0005$). The results show that students who attended the LT drew their conclusion on data-based claims. Preferring statistical

information over personal intuition and bias is an important step towards statistical inference.

To investigate the game time of 1500 students at a secondary school, a sample is taken. The students in the sample are asked how much time in hours per week they spend on gaming. They decide to randomly question 30 students at the entrance of the school.

6a. Draw in the graph below the sample results you expect.



6b. What average game time(s) do you expect for a sample of 30 students?

Level	Code	Description
6	2	Range of values corresponding to (the peak of) the graph (at Item 6a) Single value corresponding to the graph, with a measure of uncertainty
4	1	Single value corresponding to the graph
0	0	Other values (not corresponding to the graph)

Group	Average score	Percent of students per level		
		Level 6	Level 4	Level 0
Control group	3.18	19.3%	50.4%	30.3%
Intervention group	3.94	28.1%	56.1%	15.7%

Figure 4.2. Students’ achievements on posttest Item 6b, newly designed Item

Second, we present the results on posttest Item 6b, a newly designed Item (see Figure 4.2). Most students from both groups noted one specific value as their estimate of the sample result. In the intervention group, more students

considered sampling variability (28.1%) than in the comparison group (19.3%). In the comparison group, almost one-third of the answers (30.3%) did not match their answer given in Item 6a, while for the intervention group, only a smaller one-sixth (15.7%) did so. A chi-squared test on the distribution over levels confirmed a significantly higher score for the intervention group ($\chi^2(2) = 6.57$, $p < .05$).

6c. Explain your answers to Items 6a and 6b.

.....

Level	Code	Description
5	5	Statement considering the effect of small sample sizes on <i>variability</i> to explain the shape of the graph and/or the <i>range of values</i> for the average at respectively Item 6a and Item 6b; the statement is bedded in the <i>context</i>
4	4	Statement considering <i>variability</i> to explain the shape of the graph and/or the <i>range of values</i> for the average at respectively Item 6a and Item 6b; the statement is bedded in the <i>context</i>
3	3	Statement considering <i>variability</i> to explain the shape/peak of the graph and/or the average (one value or a range) at respectively Item 6a and Item 6b; the statement is bedded in the <i>context</i>
2	2	Statement <i>without variability</i> to explain the shape/peak of the graph and/or the average (one value or a range) at respectively Item 6a and Item 6b; the statement is bedded in the <i>context</i>
1	1	Vague statement of variability, using context
0	0	Statement without variability, only using context

Group	Average score	Percent of students per level			
		Level 5	Level 4	Level 3	Level 2
Comparison group	1.19	0.4%	2.1%	13.0%	20.6%
Intervention group	1.97	0.4%	9.7%	20.6%	31.8%

Figure 4.3. Students' achievements on posttest Item 6c, newly designed Item

Third, we regard the results for posttest Item 6c, a newly designed posttest Item related to Items 6a and 6b (see Figures 4.2 and 4.3). Most students in the comparison group (63.9% for levels 0–1) focused on the context, without referring to the data from their sketched graph in posttest Item 6a or their

average in Item 6b, and without taking variability into account. For the intervention group, most students (62.6% for levels 2–5) did relate data from their graph or average to the context, however, half of these students (31.8%, level 2) argued a specific sample value without taking variability into account. A chi-squared test on the distribution over levels, confirmed a significantly higher score for the intervention group ($\chi^2(5) = 28.19, p < .0005$). As such, the results for Items 6b and 6c show that students who were taught using the LT performed better on making data-based claims with reference to statistical information and accompanied by probabilistic reasoning.

Results on Students' Learning Progression

This section describes whether the supporting indicators for LT steps 1 to 8 were observed in students' worksheets (see Table 4.5). Column 3 presents the percentage of students that correctly elaborated the indicator in their work. In the following part, we highlight results from LT steps 2, 3, 4 and 7, that provided us with insight into how each of these LT steps fostered or hindered the students' learning process.

Table 4.5. Overview of Results for LT Steps 1 to 8

LT step in Sequence I <i>Categorical data</i>	Indicator	Observed result (N = 267)
1. Experimenting with physical black box	a. Making inferences about content physical black box	100%
	b. Interpreting effect of larger viewing window	88%
2. Visualizing distributions	a. Drawing expected sampling distribution from repeated samples	91%
	b. Using (given) sampling distribution to determine the probability of sample results	99%
3. Modeling a black box	Using statistical modeling in TinkerPlots to determine the probability of sample results	77%
4. Modeling real- life contexts	Using statistical modeling in TinkerPlots for	
	a. Interpreting effect of sample size in real- life contexts	98%
	b. Probabilistic reasoning in real-life contexts	83%

	c. Determining the probability of sample results, in real-life contexts	73%
	d. Informal hypothesis testing	30%
LT step in Sequence II <i>Numerical data</i>	Indicator	Observed result (N = 224)
5. Experimenting with physical black box	Making inferences about content physical black box	100%
6. Visualizing distributions	Drawing expected population distribution	76%
7. Modeling a black box	Using statistical modeling in TinkerPlots (<i>given</i> model) for	
	a. Making inferences about the population distribution	
	using a small sample size	52%
	using a large sample size	81%
	b. Interpreting effect of sample size on expected population distribution	57%
	c. Making inferences about the population mean	100%
	d. Interpreting effect of sample size on the expected population mean.	69%
	e. Determining the probability of sample results (concerning the sample mean)	32%
8. Modeling real- life contexts	Using statistical modeling in TinkerPlots to determine the probability of sample results, in real-life contexts	80%

Step 2: Visualizing the black box sampling distribution to make inferences (categorical data)

In LT step 2, for indicator 2a, most students (91%) drew a correct visualization of the expected sampling distribution as a global bell-shaped curve with a peak at 30. These students' drawings could be divided in four types (see Figure 4.4). For indicator 2b, 99% of the students correctly determined the probability of a sample result of more than 34 orange marbles based on the sampling

distribution given. Students' drawings and statements demonstrate their emerging understanding of the sampling distribution—that is, understanding the visualization of the frequency distribution from repeated sampling and using the distribution as a model for determining the probability of certain sample results—in the context of a black box. Although high deviating results were overestimated in some students' drawings and incorrect local peaks appeared, most students correctly drew a bell-shaped curve with a peak at the population proportion. Furthermore, most students correctly determined the probability of a certain range of sample results using the sampling distribution given. In a short period of time, after just one lesson, students were able to draw and interpret the (expected) sampling distribution. We assume that the physical experiments from LT step 1, combined with classroom exchange and discussion, facilitated students for LT step 2. As such, we consider LT steps 1 and 2 as essential elements to foster students' learning progress.

Step 3: Modeling a black box to make inferences (categorical data)

For step 3, the findings evidence that 77% of the students were able to use statistical modeling in TinkerPlots to determine most likely sample results within the context of a black box. The other 23% of the students incorrectly noted a vague or deterministic answer, for example: "According to TinkerPlots probably more orange than yellow marbles" or "A sample will contain 30 orange and 20 yellow." Teachers noted that most students independently deployed the required statistical modeling processes in TinkerPlots. Only a few students needed help in applying the correct digital techniques or interpreting the displays on their screen, for example the sample and sampling distributions. Teachers' feedback for those students mainly consisted of referring to the physical black box experiment and TinkerPlots instruction sheet, in particular by making explicit the similarities between the experiment and the TinkerPlots environment. As such, the initial physical black box activities in LT step 1 and 2 proved meaningful for introducing statistical modeling activities in step 3.

Step 4: Modeling real-life contexts to make inferences (categorical data)

In LT step 4, for indicator 4a and 4b, most students were able to use statistical modeling for interpreting the effect of larger sample size (98%) and for probabilistic reasoning in real-life contexts (83%). We observed more context-independent terminology than in steps 1 to 3, as students' statements involved samples, sample size, probability and variability. Teachers indicated that in the first of three lessons in step 4 about one-third of the students had difficulties applying statistical modeling in new contexts. Teachers' instruction with reference to the black box context worked well for those students with

problems. During lessons two and three of step 4, these difficulties hardly occurred. Teachers mentioned that students were inclined to refer back to the black box context in their (verbal) reasoning while working on their tasks with real-life contexts.

W2a Task description (for indicator 2a). This task is about a black box filled with 250 orange and 750 yellow marbles with a viewing window of 40. The number of observed yellow marbles per sample is noted. Consider what sample results you expect from 100,000 repeated samples. Make a sketch below, the horizontal axis displays possible sample results 0 to 40 and the vertical axis (without values) the frequency

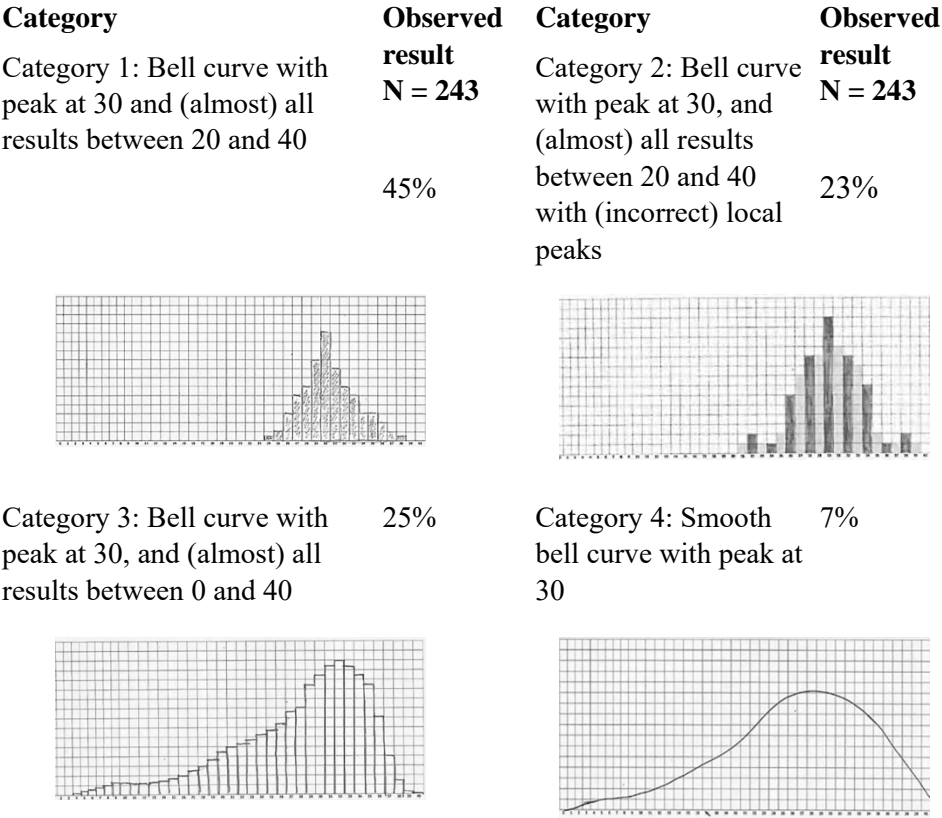


Figure 4.4. Four types of correct student drawings ($N = 243$) of the expected results of repeated sampling (100,000 repetitions) with sample size 40 in a sampling distribution, with percent per type

For indicator 4c and 4d, 73% of the students substantiated their statement with data found by statistical modeling in TinkerPlots. Of all students, 31% correctly stated that the school management can conclude that the breakfast habits of pupils are improved for unlikely high sample results—that is, for sample results above 80—and 42% incorrectly mentioned improvement for results higher than the common ones of 70, based on their TinkerPlots data found. Of all students, 27% did not refer to their TinkerPlots data found (see Figure 4.5).

Students' inferences within new real-life contexts accompanied by more sophisticated probabilistic reasoning—that is, more context-independent language and statistical terminology—confirmed their emerging understanding of key concepts. Students used their simulated sampling distribution as a model for probabilistic reasoning in real-life contexts, which is an important step towards emergent modeling. Regarding indicator 4d, using the sampling distribution to determine at what sample results it is likely that a given model can be rejected—an informal approach of hypothesis testing—appeared difficult for students. Although from steps 1 to 3, students were familiar with sampling variability, they did not transfer this knowledge to their claim and tend to use the deterministic approach that any sample proportion found, higher than the population proportion, indicates a change of population. These results confirm earlier studies about students' difficulties in understanding hypothesis testing (Stalvey et al., 2019). Nevertheless, 30% of the students correctly indicated when a given model should be rejected.

Step 7: Modeling a black box to make inferences (numerical data)

In LT step 7, for indicator 7a: making inferences about the population distribution, students tended to reflect the shape of one sample distribution found in TinkerPlots directly to the population (see Figure 4.6). However, when using a small sample size, a strict reflection often results in an incorrect irregular shape of the expected population distribution. Sample distributions for small sample sizes are less stable—sometimes even called dancing distributions—than for larger sample sizes. About half of the students (52%) compensated for these irregular shapes by comparing several (simulated) sample distributions, probably based on their experiences in LT steps 5 and 6—concerning classroom exchange and discussion of varying sample distributions found from the physical black box experiment.

<p>W4 Task Description. At the beginning of the school year, 210 out of 300 pupils had breakfast daily. At the end of the school year, the school management wants to investigate whether pupils' breakfast habits have improved (e.g., more pupils are having breakfast daily). They decide to take a sample of 30.</p>			
Specific Task	Category of answers	Examples of students' work	Observed result N = 267
<p>W4.18 (for indicator 4c) The school management decides to take a sample of 100. At which sample result (size 100) is it likely that pupils' breakfast habits have improved?</p>	Correct: referring to TinkerPlots data and considering sampling variability	"At unlikely high samples results. In TinkerPlots most common results are between 60 and 80, so for results higher than 80"	31%
	Correctly referring to TinkerPlots data, but incorrect conclusion	"For sample results higher than 70, cause in TinkerPlots most results were around 70"	42%
	Incorrect, not referring to data	"For sample results higher than 70, cause at the beginning of the school year 210 out of 300 had breakfast daily"	27%

Figure 4.5. Percent of students per category of answers on Worksheet 4 Task 18

For indicator 7b, regarding the effect of sample size on the expected population distribution, most students (78%) correctly stated that the distribution from a larger sample better reflects the population distribution. Most of these students (73%) explicitly mentioned that larger sample sizes lead to more stable distributions: less variability, smoother bell-curve, a peak at the population mean, and fewer local peaks; the other 27% of these students stated that a larger sample contains more information which results in a 'bigger' distribution: has a wider range of results and higher bars. For 22% of the students we found incorrect statements, for example: "The distributions for small and large sample sizes are quite similar." For making inferences about the population mean using

small and large samples, most students (69%) stated that a larger sample leads to a better estimate of the population mean: more stable, precise and reliable. The other 31% stated that for the expected population mean, using small or large samples sizes were quite similar.

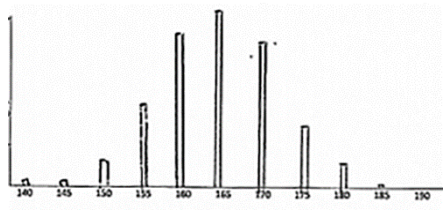
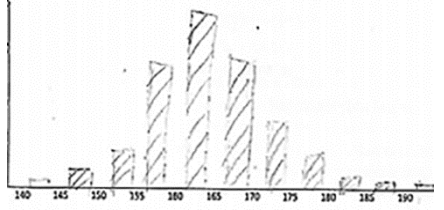
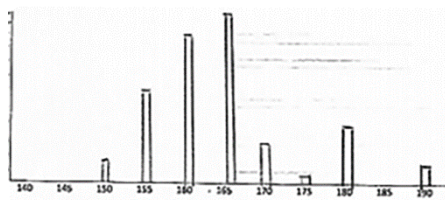
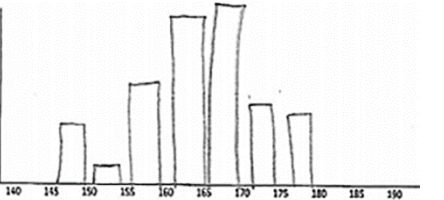
Task description. Sketch of expected population distribution for the content of the black box (height of 4,000 students) (for indicator 7a)			
[W7.1] ... , using a small sample size (40)		[W7.8] , using a large sample size (500)	
Examples of students' work at W7.1	Observed result	Examples of students' work at W7.8	Observed result
Sketch of the expected population distribution (correct)	52%	Sketch of the expected population distribution (correct)	81%
			
Sketch of expected irregular-shaped distribution with local peaks	48%	Sketch of expected irregular-shaped distribution with local peaks	19%
			

Figure 4.6. Percent of students per category of answers on Worksheet 7 Tasks 1 and 8

Regarding indicator 7e, most students (68%) had difficulties determining the probability of certain sample results. Students' problems mainly consisted of confusing the sample and sampling distribution. For example, when students were asked to determine the probability of a sample mean below 1.55 m, students tended to refer to their simulated sample distribution instead of the sampling distribution; we also observed the other way around, when students were asked to determine the probability that a person's height is below 1.55 m. We assume that the emphasis on three distributions—that is, sample, population and sampling distribution—in LT steps 5 and 6 caused confusion.

Overall, teachers explicitly mentioned that the black box as guiding activity through the learning trajectory was clear and useful, especially the strong similarities between the physical black box and statistical modeling in TinkerPlots. Furthermore, teachers described the black box as a concrete, engaging activity that is free of bias—meaning not related to students' personal preference or prior knowledge. The learning of digital techniques for using TinkerPlots in a short period of time took some time and effort. Teachers indicated that investing in these techniques was worthwhile, and that most students deployed the techniques rather easily.

Conclusion and discussion

This article reports on a design study that aimed for a theoretically and empirically underpinned design of an LT for introducing statistical inference in Grade 9. We addressed several aspects involved in design research on LT's as advised by Duschl et al. (2011). To evaluate the designed LT, we analyzed the progression made by 267 students. First, the analysis of the posttest results indicates that students' understanding of statistical inference as addressed in the coupled LT steps—in LT steps 1 and 5 on using samples, in LT steps 2 and 6 on visualizing distributions, in LT steps 3 and 7 on repeated sampling and effect of sample size, and in LT steps 4 and 8 on solving real-life problems—was significantly higher among students who took part in the LT than among students who followed the regular curriculum. These results demonstrate a higher score for all eight learning steps and, with that, a deeper understanding of the statistical concepts offered in each step. As such, it appears that all eight steps combined led to students' higher performance on statistical inference. Second, the analysis of students' worksheets, accompanied by teachers' and researcher's notes, confirms that all eight steps of the learning trajectory combined contributed in fostering students' learning. In addition to developing the statistical concepts addressed within each learning step, we also observed

progress across the eight successive learning steps—for example, in using more abstract statistical terminology, data-based reasoning, and context-independent use of statistical concepts and models. As such, the results empirically substantiate the theoretically designed learning trajectory.

Although research shows that reasoning and interpreting sampling distributions is difficult (Batanero et al., 1994; Castro Sotos et al., 2007; Chance, del Mas, & Garfield, 2004), the findings show that students can develop key concepts of statistical inference—sample, variability, and distributions—in a short period of time by using the black box sampling as guiding activity. Starting from LT steps 1 and 2, students developed an emerging understanding of the sampling distribution, initially as a visualization or model of their results, and gradually as a model for determining the probability of certain sample results. The strong similarity between the physical black box activities and the modeling activity in the digital environment of TinkerPlots facilitated the connection of the model to the real world (Konold & Kazak, 2008; Patel & Pfannkuch, 2018). In following LT steps, the black box served as a guiding paradigm for students' reasoning and teacher instructions about key concepts, in particular while modeling real-life phenomena.

Based on the promising results of this study into an LT for introducing statistical inference—designed on the basis of current ideas and theories in this area—we identify the following design heuristics as useful. First, the learning activities should be placed in a context that allows students to develop statistical concepts directly related to the learning goals of the LT—that is, a context that is recognizable to students, engaging, activating, and representative for the concepts at stake. Second, although activities may focus on specific statistical concepts, they should be viewed within the broader perspective of the entire statistical investigation cycle. Here, it is essential that students go through this cycle repeatedly, using different contexts with increasing levels of abstraction and complexity. Third, visual and enactive similarity between material and digital sources must be ensured for performing statistically identical procedures. Fourth, explorative and iterative activities with simulation software should be embedded to facilitate the development of context-independent conceptual understanding. Fifth, activities should be structured to support learners in developing a *model* of their concrete statistical activity that can then be used as a *model* for a network of statistical concepts and relationships.

However, when higher order thinking activities were addressed in the LT, such as informal hypothesis testing or reasoning about population distributions,

we saw confusion among students. Apparently, more time and more iterations are needed to anchor the key concepts before proceeding to more complex statistical concepts and ideas. We therefore suggest in Sequence II—steps 5 to 8—to focus on repeated sampling using the sample mean and to omit making inferences about the population distribution. In this way, the key concepts for statistical inference from Sequence I that emerge from the sample proportion of categorical data for repeated sampling can be further elaborated in Sequence II by using the sample mean of numerical data.

To address more complex statistical concepts in a follow-up LT, repeated sampling with a black box (or boxes) may also be used as guiding activity. With regard to hypothesis testing, which is difficult for many students (Stalvey et al., 2019), a hypothesis concerning the black box content can be used to introduce the idea of hypothesis testing. For example, by providing a physical black box filled with marbles and letting students test whether the given proportion is likely to be true. This also holds for other statistical concepts and ideas, such as determining the critical area and comparing groups, where the black box provides opportunities for engaging and guiding activities.

Concerning the use of digital technology in the LT, investigating in learning to use a digital tool—which took time and effort from both teachers and students—appeared fruitful for students’ understanding of statistical inference. The digital techniques for using the tool enabled students to identify context-independent patterns in action that seemed to facilitate the transition towards emergent modeling. This transition was reflected in students’ worksheets when they referred to similar previous technical actions and in students’ terminology that evolved from concrete terms to more abstract statistical terminology, for example from the term “viewing window” to “sample.” The development of a statistical vocabulary is essential for students’ understanding of concepts (Watson & Kelly, 2008).

The results of this study can be positioned within the findings of our larger study. The findings from the larger study using all assessment Items of the pre- and posttest indicate that the LT also stimulated other domains of statistical literacy (Van Dijke-Droogers et al., submitted). These findings suggest that the current Dutch pre-10th grade curriculum can be enriched with informal statistical inference; we assume that this also holds for other countries with a focus on descriptive statistics in lower secondary mathematics curricula.

Of course, this study comes with some limitations. Teachers’ implementation of the LT varied, for example in the amount of teacher guidance

and instruction during the teaching sequence. These differences were visible in students' worksheets, with the reasoning of students with the same teacher being more or less similar. Furthermore, we encountered practical limitations during the intervention, such as difficulties with installing TinkerPlots on the school's computer network and lesson shortening due to extremely high temperatures. The installation problems caused some delay but did not affect our study. Due to the lesson shortening, we collected 224 completed worksheets in Sequence II, instead of the 267 in Sequence I.

On a final note, the findings suggest that curricula with a strong descriptive focus can be enriched with an inferential focus—at least for this type of student population—with the benefit of students learning more about inference, but not less about descriptive statistics. We recommend that educators and researchers involved in the design of teaching materials consider the embedding of black box activities combined with statistical modeling, to anticipate subsequent steps in the students' statistics education.



Effects of a Learning Trajectory for Statistical Inference on 9th-grade Students' Statistical Literacy

This chapter is based on

Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted).
Effects of a learning trajectory for statistical inference on 9th-grade students'
statistical literacy.

Abstract

In our data-driven society, it is essential for students to become statistically literate. A core domain within statistical literacy is statistical inference, the ability to draw inferences from sample data. Acquiring and applying statistical inference is difficult for students and, therefore, usually not included in the pre-10th-grade curriculum. However, recent studies suggest that developing a good understanding of key statistical concepts at an early age facilitates the understanding of inferences later on. This study evaluates the effects of a learning trajectory for statistical inference on Dutch 9th-grade students' statistical literacy. Theories on informal statistical inference and repeated sampling guided the learning trajectory's design. For the evaluation, we used a pre-post research design with an intervention group ($n = 267$). To interpret the significant learning gains of this group, we compared students' results with national baseline achievements from a comparison group ($n = 217$) who followed the regular 9th-grade curriculum, and with international studies using similar test items. Both comparisons indicated that the learning trajectory had a significant positive effect on students' statistical literacy and on the ability to make inferences in particular, but also on the other domains of statistical literacy. These findings suggest that current statistics curricula for Grades 7–9, usually with a strong descriptive focus, can be enriched with an inferential focus.

Keywords

statistical literacy, statistical inference, learning trajectory, assessment instrument, learning effects

Introduction

In our data-driven society, it is essential for citizens to be statistically literate. Both our daily activities and professional practices increasingly rely on statistical information we obtain, either from taking measurements or through media reports. Statistical literacy concerns the ability to interpret, critically evaluate, and communicate about statistical information and messages (Gal, 2002). The growing use of and dependence on statistical data requires an educational approach in which students learn to create and critically evaluate data-based claims (Ben-Zvi et al., 2015) and, as such, to become statistically literate.

A core domain of SL is drawing inferences from sample data. However, learning and applying statistical inferences (SI) is difficult for students (Castro Sotos et al., 2007; Konold & Pollatsek, 2002). Therefore, in many countries, including the Netherlands, it is not offered in the pre-10th-grade curriculum. Recent studies suggest that developing, at an early age, a good understanding of key statistical concepts of sample, variability and distributions facilitates the understanding of SI later on (Ben-Zvi et al., 2015; Zieffler et al., 2008). Innovative educational software for simulating samples and repeated sampling offers opportunities to make these key concepts accessible (Biehler et al., 2013).

To support students' SI, a learning trajectory (LT) for 9th-grade students (14–15-years old) was designed to introduce the key concepts of SI (Van Dijke-Droogers et al., 2020). Theories of informal statistical inference (Makar & Rubin, 2009) complemented by ideas of growing samples and repeated sampling (Bakker, 2004), constituted the design of the LT. This simulation-based LT comprises an investigative approach that includes all stages of the statistical investigation cycle—from collecting data to interpreting the results—with an emphasis on interpreting sample data and reasoning about probability. Although the focus of the LT is on SI, the approach concretizes broader underlying statistical concepts, such as measures of center and spread, distribution and correlation, by means of visualizations. As such, our conjecture is that the designed LT for introducing SI will also have a stimulating effect on the other, more descriptive-focused, domains of statistical literacy. Currently, the typical Dutch pre-10th-grade curriculum is mainly focused on those descriptive domains. In this regard, the purpose of the LT is to expand the 9th-grade curriculum with SI, the more complex domain of statistical literacy, without neglecting the current educational goals on the other domains.

The aim of the study reported here is to evaluate the effects of the designed LT for introducing statistical inference on students' statistical literacy. Therefore, we wanted to assess students' performance on SI, and their achievements on the other descriptive-focused domains of statistical literacy as offered in the regular curriculum. Because such assessment instruments with a specific focus on SI hardly exist for our age group, we developed a pre- and posttest, by adapting and expanding already validated tests. This assessment instrument enabled us to establish students' performance on both tests, and hence to evaluate the effects of the designed LT for statistical inference on students' statistical literacy, and on the SI domain in particular.

Theoretical Background

Domains of Statistical Literacy

Statistical literacy (SL) concerns critical thinking that uses statistical information as evidence (Schield, 2004). This includes the ability to read and interpret numbers in statements, surveys, tables and graphs and studies how statistical associations are used as evidence for causal connections. Although SL has several definitions, the most-used one comes from Gal (2002), where SL is portrayed as the ability to interpret, critically evaluate, and communicate about statistical information and messages. According to Rumsey (2002), SL includes the understanding of basic statistical concepts and ideas in data awareness, production, understanding, interpretation and communication.

Three domains of SL can be distinguished (Watson & Callingham, 2003). The average and chance (AC) domain covers determining measures of center and spread, and calculating and interpreting chance issues, as reflected in the mathematics curriculum in most Western countries (Watson & Callingham, 2004). The graphing and variation (GV) domain entails creating and interpreting visual representations of data with the variation involved. The sampling and inferences domain focuses on statistical inference and, as such, can be considered as the statistical inference domain within SL. This SI domain covers working with samples and drawing inferences, where interpreting the relationship between these two is particularly important in the process of statistical decision making.

Many secondary school curricula make a distinction between statistics without probability (descriptive statistics, exploratory data analysis), as addressed in the GV and AC domains, and statistics with probability (inferential statistics) as addressed in the SI domain. The latter is usually taught at upper levels (Burrill & Biehler, 2011). This also holds for the Dutch secondary school

curriculum, in which statistics education progresses from descriptive statistics in the early years, to preparing for a more formal approach to inferential statistics from Grade 10 and in higher education (Van Dijke-Droogers et al., 2017; Van Streun & Van de Giessen, 2007). In the Dutch curriculum for Grades 7–9, the first two domains of SL are embedded in the descriptive statistics, whereas the SI domain is not addressed at all.

Statistical Inference

Statistical inference (SI) is at the heart of statistics as “it provides a means to make substantive evidence-based claims under uncertainty when only partial data are available” (Makar & Rubin, 2018, p. 262). As such, SI can be considered both an outcome and a reasoned process for probabilistic generalizations from data (Makar & Rubin, 2009). SI concerns interpreting sample results, drawing data-based conclusions, and reasoning about probability. For most students, it is difficult to understand SI and the uncertainty involved. Several studies focused on the introduction and conceptualization of SI. The offering of educational activities of SI at an early age on informal level, combined with the frequent recurrence of such activities later on, seems to make SI accessible for students, in particular at the school level (Makar & Rubin, 2009; Paparistodemou & Meletiou-Mavrotheris, 2008; Van Dijke-Droogers et al., 2020; Zieffler et al., 2008). In general, this informal approach focuses on ways in which students without knowledge of formal statistical techniques, such as hypothesis testing, use their statistical knowledge to underpin their inferences about an unknown population based on observed samples. A widely used framework for informal statistical inference identifies three main principles: generalization beyond data, data as evidence for these generalizations, and probabilistic reasoning about the generalization (Makar & Rubin, 2009).

SI requires an understanding of the key concepts of sample, variability and distribution—including frequency distribution and (simulated) sampling distribution. These concepts can be introduced at the school level by using ideas of simulating repeated samples (Garfield et al., 2015; Manor & Ben-Zvi, 2017; Rossman, 2008; Saldanha & Thompson, 2002; Watson & Chance, 2012) and growing samples (Bakker, 2004; Ben-Zvi et al., 2012; Wild et al., 2011). Digital tools such as TinkerPlots offer opportunities for simulating repeated samples and to visualize concepts, such as random behavior, distribution and probability (Garfield et al., 2012; Konold et al., 2007; Pfannkuch et al., 2018). Working with such simulations stimulates the understanding of statistical models and modeling processes, that are essential for SI. In the LT we designed, students start with interpreting the sampling distribution obtained from repeated

sampling with a physical black box filled with marbles. As a follow-up, students build and run a model of a real world situation in TinkerPlots and use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real world situation, and to draw inferences. An overview of the LT can be found in Table 5.2.

Assessing Statistical Literacy and Inference

Assessment instruments at the secondary school level for SL, with a focus on SI, are scarce. The situation is very different at the tertiary level; think of the web-based ARTIST project—Assessment Resource Tools for Improving Statistical Thinking—by Garfield, delMas and Chance (2002), the CAOS project—Comprehensive Assessment of Outcomes in a First Statistics Course—by delMas et al. (2007), the GOALS project—Goals and Outcomes Associated with Learning Statistics—by Garfield et al. (2012), and the BLIS project—Basic Literacy in Statistics—by Ziegler (2014). The latter project, BLIS, involves a compilation of existing Items from the other projects supplemented with simulation-based questions. The Items in these projects require students to think and reason, not to compute, use formulas, or recall definitions.

The only studies that seemed useful for our students were the ones by Watson and Callingham (2003, 2004) and the LOCUS project (Whitaker et al., 2015), as both focused on Grades 6 to 12. Watson and Callingham's studies appeared to be particularly suited, as they specifically distinguished between the three domains of SL. Their approach allowed to identify students' SL, and also their performance on the domain of SI in particular. Using archived data from 1993–2000, Watson and Callingham empirically developed a 6-level hierarchy of SL that helped to identify the distribution of Australian middle school students' SL across the levels. Their hierarchical levels for SL are presented in Table 5.1. A follow-up study by Callingham and Watson (2017) showed that the level construct had remained appropriate and stable over time. This finding suggests that the identified levels provide a good basis for determining the level of SL in secondary education. In addition, their longitudinal analysis indicates that the statistical literacy hierarchy can be used to monitor students' progress.

Research Question

This study focuses on the question:

What are the effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy?

To answer this question, we examined the effects of the LT on students' proficiency in the domains of SL, SI in particular. Although the designed LT concentrates on statistical inference—the SI domain of SL—we conjectured that a focus on more complex learning activities for SI would also have a positive effect on students understanding of the other domains of SL.

Table 5.1. Levels of Statistical Literacy as presented by Watson and Callingham (2003, p. 14)

Level	Characteristic of level
6. Critical Mathematical	Critical, questioning engagement with context, using proportional reasoning particularly in media or chance contexts, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language.
5. Critical	Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, qualitative interpretation of chance, and appreciation of variation.
4. Consistent Non-critical	Appropriate but non-critical engagement with context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities, and graph characteristics.
3. Inconsistent	Selective engagement with context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.
2. Informal	Only colloquial or informal engagement with context often reflecting intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph and chance calculations.
1. Idiosyncratic	Idiosyncratic engagement with context, tautological use of terminology, and basic mathematical skills associated with one-to-one counting and reading cell values in tables.

Methods



To evaluate the effects of the LT, we used a pre-post research design with an intervention group ($n = 267$) who engaged with the LT. To interpret the learning gains of the intervention group, we compared their results with national baseline achievements from a comparison group ($n = 217$) who followed the regular Dutch curriculum at an earlier stage, and compared the results with those of Australian students (Callingham & Watson, 2017).

An Outline of the Learning Trajectory

A Learning Trajectory (LT) is a design and a research instrument to structure and connect all elements involved in learning a particular topic. An LT consists of a set of learning goals for students, learning activities that will be used to achieve these goals, and conjectures about the students' learning process. It includes the simultaneous consideration of mathematical goals, student thinking models, teacher and researcher models of students' thinking, sequences of teaching tasks, and their interaction at a detailed level of analysis of processes (Clements & Sarama, 2004).

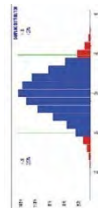
The designed LT introduces the key concepts for statistical inference to 9th-grade students by using an investigative approach with a physical black box and simulation-based methods (Van Dijke-Droogers et al., 2020), see Table 5.2. Ideas of repeated sampling and growing samples instantiate the design, both for working with the physical black box filled with marbles and for simulating samples using TinkerPlots. All stages of the statistical investigation cycle are addressed in the LT, as students collect both physical and simulated data, analyze their data using the sampling distribution, and interpret the results to answer the question posed. The emphasis is on interpreting sample data and reasoning about probability. Recent views on statistical models and modeling (Büscher & Schnell, 2017; Manor & Ben-Zvi, 2017; Patel & Pfannkuch 2018), and educational guidelines on the use of context, digital tools, exchange and comparison of sample results, making predictions, and engagement in both physical and simulation-based activities, are embedded in the design. The investigative approach and learning activities in the more complex SI domain also attend to the other domains of SL. For example, the AC domain, average and chance, is addressed as students summarize their obtained sample data in measures of center and spread. As another example, the graphing part of the GV domain is given attention in the visualizations of both sample results and population models, and the variation part is targeted as students explore results of repeated samples.

Table 5.2. Overview of Steps 1–8 of the Learning Trajectory

LT Step	Description	Example of activities	Learning Goal	Construction of LT steps
<i>Categorical data</i>				
1. Experimenting with physical black box	Physical black box with <i>marbles</i> experiment (with small and large viewing window)	 Estimate the number of yellow marbles in a black box filled with 1,000 marbles (<i>balletjes</i> in Dutch) by shaking and observing visible marbles	Students draw inferences and become accessible to concepts as sample, sample size, sampling variability, frequency and measures of center and spread, within the context of a physical black box	Students experience that sample results vary and that a larger sample size and more repeated samples lead to a better population estimate. Next question: What happens when we further increase the size and number of repeated samples? Conducting larger and more samples is time consuming: a thought experiment can help.
2. Visualizing distributions	Graph as a model (or visualization) of the frequency distribution from repeated sampling with the black box	 Make a sketch of the frequency distribution you expect when the black box experiment with a large viewing window is repeated 100,000 times	Students can draw the visualization of an expected sampling distribution from repeated samples. Students interpret sampling distributions given to make inferences about a certain range of sample results	The sampling distribution from repeated sampling can be used to determine the probability of certain sample results. Next question: How can we get the sampling distribution of repeated sampling in a quick and easy

way? Using technology can help.

3. Modeling a black box (ICT)



Using simulation of repeated samples, from a modeled black box, in a sampling distribution as a model for interpreting probability

Use TinkerPlots to determine the most common sample results for a black box filled with 750 yellow and 250 orange marbles, and sample size 40

Students use statistical modeling within the digital environment of TinkerPlots to determine (un)likely sample results, within the context of a black box [Statistical modeling includes building a model, simulating (repeated) samples, visualizing the sampling distribution and interpreting the results]

Statistical modeling—including interpreting the sampling distribution from repeated sampling—can be used to determine the probability of certain sample results, within the context of the black box. Next question: Can statistical modeling be used more generally, in other situations and contexts?

4. Modeling real-life contexts (ICT)





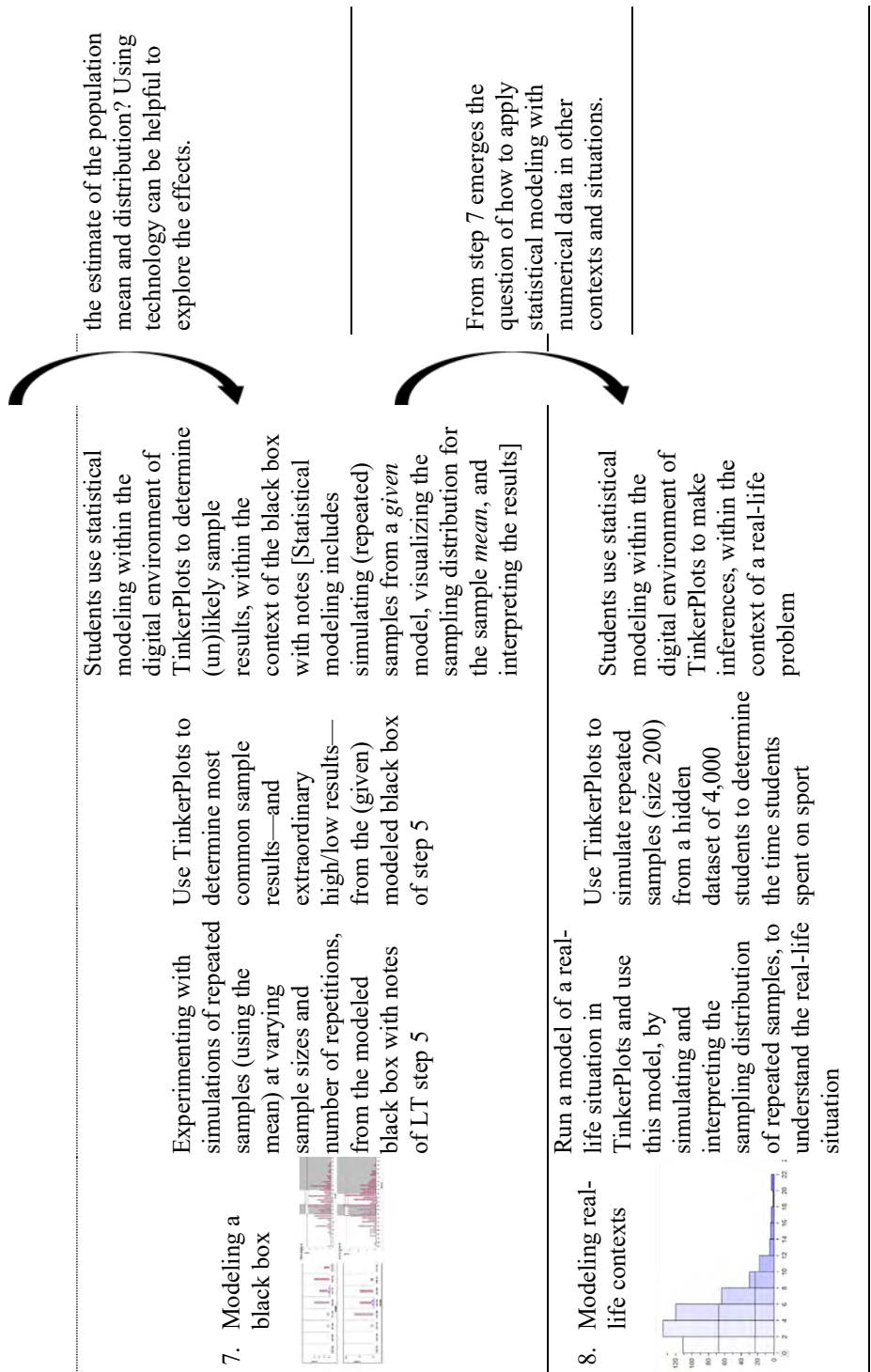
Build and run a model of a real-life situation in TinkerPlots and use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real-life situation and the probability involved

Use TinkerPlots to determine most common sample results when a sample of 30 is taken from a school with 300 students to determine the number of students having daily breakfast (given that on average 70% of the students have breakfast daily)

Students use statistical modeling within the digital environment of TinkerPlots to make inferences, within the context of a real-life problem

From steps 1 to 4 emerges the question of how to use statistical modeling with other—not categorical—data.

Numerical data	
<div>5. Experimenting with physical black box</div> <div></div>	<div>Take a sample of 40 notes and summarize the sample data found (calculate measures of center and spread, use a visualization).</div> <div>Estimate the gender (proportion) and height (center and spread) of the 4,000 students</div> <div>Students draw inferences within the context of the physical black box with notes (students' gender and height) considering sample size, sample variability, and measures of center</div>
<div>6. Visualizing distributions</div> <div></div>	<div>Summarize and visualize the expected population (height of 4,000 students) based on the sample data found in LT step 5</div> <div>Sketch the frequency distribution you expect for the whole population, based on the sample results found in step 5</div> <div>Students draw a visualization of the population distribution they expect from the sample results found. Student draw inferences about the population, considering distribution, mean, sample variability and probability</div>
<div>Students discussed how to use numerical data from repeated samples to draw inferences about the population. Next question: how can the population distribution at stake—the content of the black box filled with 4,000 notes on students' gender and height—be visualized based on the varying sample results found?</div> <div>Students draw inferences about the population mean and population distribution using samples found. Next question: what are the effects of larger and more repeated samples on</div>	



The LT comprises eight learning steps that are split into two similar parts of four. Part one considers only categorical data and includes the following steps: (1) experimenting with a physical black box, (2) visualising distributions, (3) statistical modeling using TinkerPlots, (4) applying models in new real-life contexts. Subsequently, in part two, LT steps (5) to (8) include similar steps, now using more complex numerical data. The eight steps of the LT were organized in two sequences of six 45-minutes lessons, with a total of twelve lessons.

Design of the Assessment Instrument

To evaluate the effects of the designed LT, we needed an assessment instrument to measure 9th-grade students' SL, and SI in particular. To measure the effects of the LT on students' proficiency—i.e., students' progress when working with the LT—we developed an assessment instrument consisting of a pre- and posttest, inspired by Watson and Callingham (2003, 2004). Following Ziegler (2014), we used existing items from validated tests for the design of the tests, supplemented by simulation-based items. As such, we used the approach of a pre- and a posttest from delMas et al. (2017), test items for statistical reasoning with levels from Watson and Callingham (2004), and expanded these with newly designed items on statistical inference and simulation.

The pre- and posttest each contained ten clusters of items. Each cluster included two to six sub items, with a total of 39 and 34 items on the pre- and posttest respectively. Both tests had a similar composition and a time-duration of 45 minutes. For each test, we selected five clusters of items from Watson and Callingham (2004) that covered the three domains of SL. We selected one cluster item applicable for secondary level from the CAOS test (delMas et al., 2007). As context was found to be an important factor affecting the difficulty of items for students, the selection of items was based on educational background, as well as on familiarity with the context. Table 5.3 provides an overview of the composition of the pre- and posttest, with reference to sources and accompanying domains of SL.

Figure 5.1 shows an example of an item from a validated test, in the AC domain. The level scores in this item refer to Watson and Callingham's (2003) hierarchical levels 1 to 6 for SL, supplemented with the null level for incorrect or uncompleted items. As Figure 5.1 shows, the answers could not be given on each level: It was not possible to formulate an answer on levels 1 and 2, the informal and inconsistent level, as all possible answers include the context information given—level 3 or higher—or the answer is incorrect—level 0.

Table 5.3. Overview of Clusters and Items in the Pre- and Posttest

Number of Items (clusters)		Source	Domain of SL
Pre	Post		
17 (5)	18 (5)	Watson & Callingham	AC – GV – SI
3 (1)	2 (1)	CAOS	AC
19 (4)	14 (4)	Newly designed	SI

Note. SL = Statistical Literacy, AC = Average and Chance, GV = Graphing and Variation, SI = Statistical Inference.

Similarly, based on the item context, some items could only be coded to a maximum level score of 4 instead of 6. As such, for the selection of items, the chosen items had to be similar in maximum level score on the pre- and posttest, for each domain of SL, to compare students' scores on both tests. The average maximum scores for SI items on the pre- and posttest were similar, both around 5.6, and, for the GV items, the average maximum scores were also similar, with around 3.7 for both tests. For AC, however, the maximum scores on the selected items in the pre- and posttest were rather different, with 5.7 and 4.6, respectively. To compensate for this difference, a correction was applied to the posttest results, so that students' level scores on the pre- and posttest could be properly compared. Using the corrected AC scores, the average maximum score on SL was about 5.5 for both tests. As such, we considered the selected items on the pre- and posttest comparable for both tests, on all domains of SL.

As we were specifically interested in the effects of the LT on students' understanding of the concepts of SI as addressed in the LT, four additional items were designed for this study, focusing on the SI domain. For the design, we chose recognizable contexts and used the structure and phrasing of items from the two previously described tests. Figure 5.2 shows an example of a newly designed item with its levels. The level scores of these new items were, as with the existing items, based on Watson and Callingham's (2003) level descriptions, and on the exemplary Items they formulated on the SI domain (2004).

To analyze the validity of the designed assessment instrument for our Dutch 9th-grade students, we conducted two pilot tests in different classrooms, each consisting of 25 students, for the pretest. Concerning the concurrent validity of the new designed SI items, we expected the students to score on the

newly designed SI items at a similar level to the existing SI items from Watson and Callingham (2004). Students' average level scores in the pilots on newly designed and existing SI items were not significantly different ($M_{\text{new}} = 2.49$, $SD_{\text{new}} = 0.71$, $M_{\text{ex}} = 2.78$, $SD_{\text{ex}} = 1.38$, $n = 50$, $t(49) = -1.6$; $p = .11$). For the other domains, GV and AC, all items were from already validated tests. To assess the content and construct validity of all test Items for our students, the results of each pretest pilot were used for in-depth discussion with experts in this area on content, construct, vocabulary, and clarity. In a similar way, the posttest was piloted in two other classrooms. The posttest pilots took place after the large-scale implementation of the pretest. Based on our pretest experiences, the initial designed posttest was modified slightly—for example, the number of items was reduced from 38 to 34. The results of the two posttest pilots, each consisting of 25 students who did not follow the LT or other statistics education in the intervening weeks, were thoroughly examined to ensure the pre- and posttest were comparable.

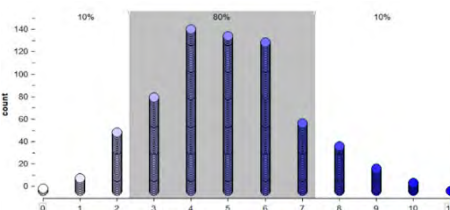
Pretest Item		
<p>Nine students in a science class weighed a small object separately on the same scales. The weights (in grams) recorded by each student are: 6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3. The students had to decide on the best way to summarize these values. Ben said, "I'd use the most common value to get the mode. That's 6.3."</p> <p>Is Ben's approach a good way to summarize the information? Explain your answer.</p>		
Level	Description	Examples of students' reasoning
6	Statistical and contextual responses incorporating both positive and negative aspects of method	"Yes, because Ben is using the most common weight for the Item. However, he does not look at the other weights and if the most common weight was an extreme value it would be inaccurate"
5	Statistical response – positive evaluation	"Yes, the majority of times it was weighed at 6.3"
	Statistical response – negative evaluation	"No, doesn't take into account the other weights"

4	Claims of inaccuracy but with no statistical response – negative evaluation	“No, the mode might weigh more than the others”. No, it’s not accurate”. “No, three people might have weighed wrong”
	Claims of accuracy but with no statistical response – positive evaluation	“Yes, it’s the average weight”
3	Recommendation of other methods	“No, he should have added them up and divided by 9”
	Tautological but positive evaluation based on majority or “most common”	“Yes, because he is using the most common”
	Methodological reasons – positive and/or negative evaluations	“Yes, it’s easy”. “No, too much calculating”
0	No reason or apparent logic regardless of evaluation No response	

Figure 5.1. Item with corresponding level description from Watson and Callingham (2004, p. 138)

Concerning the reliability of the tests, Cronbach’s alpha values were .84 and .85 on the pre- and posttest respectively, indicating a good reliability (Taber, 2017). To assess the difficulty of the items, p values were calculated. To assess the discrimination of the items, we used Rit (Item–test correlation) and Rir (item–rest correlation), using classical test theory. See Table 5.4 for an overview of the reliability of item characteristics on the pre- and posttest, with accompanying ratings. For the pretest, we observed moderately difficult items with four easy Items (p value $> .80$) and one difficult item (p value $< .20$). Rit and Rir values $> .30$ are indicated as good items, scores between .20 and .30 as medium, and scores $< .20$ as poor items (Ebel & Frisbie, 1991). The pretest Rit values indicated five poor items, twelve moderate and twenty-two good items, and, the Rir scores indicated eight poor items, sixteen moderate and fifteen good items. For the posttest, we observed moderately difficult items with four easy items and no difficult items. The Rit values indicated one poor item, nine moderate and 24 good items, and the Rir scores indicated two poor items, thirteen moderate and nineteen good items. We considered these item scores on the pre-

and posttest to be most acceptable. The pre- and posttest can be found in Supplementary Material C and D.

Pretest Item To analyze the number of candies with strawberry taste in a roll of ‘Minitos’, 700 rolls were checked. Each roll contained 20 candies. From each roll the number of candies with strawberry flavor was counted. The results of these counts are shown in the graph. Pieter claims that he had a roll in which half the candies were strawberry-flavored last week. Explain what you think of his claim.	Results for 700 candy rolls  Number of strawberry candies in one roll	
Level	Description	Examples of students' reasoning
6	Statement admitting possibility, but also acknowledging the unlikelihood of the event, based on graph Statement of low likelihood based on being an outlier, with reference to the graph	“Well, it is possible that Pieter is telling the truth, but it is very unlikely. According to the graph, there is less than 2% chance” “The story of Pieter is very unlikely. According to the graph, there is very little chance of having 10 strawberry candies in one roll, however, maybe he was extremely lucky”
4	Statement of impossibility or possibility based on being an outlier without mentioning the graph	“Maybe Pieter was lucky, it seems very unlikely to have that number of strawberry candies in one roll”
3	Definite statement of impossibility or possibility, without explicitly referring to the graph	“Pieter is exaggerating, it is impossible to have that number of strawberry candies in a role”
2	Statement of possibility without acknowledging unlikelihood or reference to the graph	“That is a large number of strawberry candies”

0	Statement based on personal experience No reason or apparent logic regardless of evaluation	“I hope Pieter likes strawberry flavor”
---	--	---

Figure 5.2. Newly designed Item with corresponding level description on the SI domain of statistical literacy (SI = Statistical Inference)

Table 5.4. Reliability and Item Characteristics of the Pre- and Posttest

	Pretest		Posttest	
	Average measure	Rating	Average measure	Rating
<i>p</i> value	.54	Moderately difficult	.62	Moderately difficult
Rit value	.35	Good	.42	Good
Rir value	.30	Medium/good	.36	Good
Cronbach’s α	.84	Good	.85	Good

Participants

Figure 5.3 provides an overview of participants and data collection. The participating students from both the intervention and comparison group were in the pre-university stream, and thus belonged to the 15% best performing students in our educational system.

For the intervention group, through a national call, in for instance newsletters for math teachers and on Social Media, we invited Dutch teachers who were willing to implement the LT in their regular mathematics lessons. Eleven of them applied, with a total of 267 9th-grade students (aged 14–15 years) from thirteen classes in five different schools. Two teachers participated with two of their classes. The teachers were instructed for the LT during two similar 3-hr sessions. The first session focused on LT steps 1–4 and included the 45-min lessons 1 to 6. The teachers worked through students’ lessons and materials themselves, guided by the researcher. The second session was similar to the first one and concentrated on LT steps 5–8, lessons 7 to 12. The project materials consisted of a teacher guidebook and students’ materials, such as worksheets, datasets, and physical black boxes with marbles. The teachers of the intervention group decided to eliminate all the regular 9th-grade statistics lessons to save time for the LT.

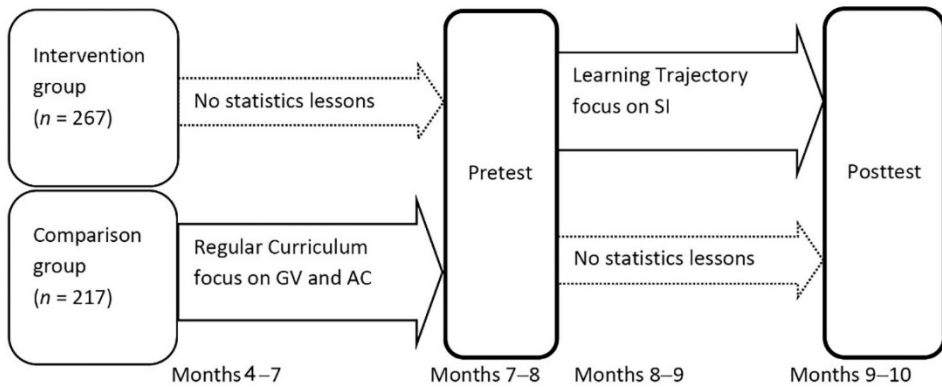


Figure 5.3. Overview of data collection and statistics education for the intervention and comparison group in grade 9. GV, AC and SI refer to the three domains of SL: graphing and variation (GV), average and chance (AC), and statistical inference (SI)

Subsequently, we invited a comparison group through a national call, with teachers who did not participate in the intervention, but who were interested in using the assessment instrument to identify the SL of their students. The effort for teachers in the comparison group was considerably lower than for the intervention group. The six teachers of the comparison group only administered the pre- and posttests on their students (217 in ten classrooms). All students in the comparison group attended 10–16 regular statistics lessons during their mathematics lessons before the pretest. The regular curriculum focused on the AC and GV domains of SL, as described earlier in the section on the domains of SL. The comparison group attended no statistics lessons in between the pre- and posttest, and therefore, we expected their results on both tests to be similar. As such, the average results on the pre- and posttest for the comparison group could be used as Dutch baseline achievements for the SL of 9th-graders.

One might suggest that posttest results of the two groups are not comparable because the posttest of the intervention group was taken within 1–2 months after the intervention, and in the comparison group not until 3–4 months after education. Taking into account the time span between education and assessment, we could also compare the pretest results of the comparison group (1–3 months after education) with the posttest of the intervention group (1–2 months after education). However, in retrospect, the comparison group's results on the pretest, which was taken 1–3 months after their education, was not significantly different from their results on the posttest. Since the results of the

comparison group on the pre- and posttest hardly differ, we considered the consequences of the intervening time to be negligible.

We are aware that teachers from the intervention group who were willing to ‘go the extra mile’ were possibly more motivated for teaching statistics. However, the teachers of the comparison group also volunteered, mainly because they were interested in the performance of their students in the field of statistics. In this regard, the teachers from both groups had an above-average interest in teaching statistics. Students in both groups belonged to the 15% best achieving students in the Dutch educational system. They all successfully completed the regular statistics curriculum in Grades 7 and 8. Students’ grade level, from both the intervention and the comparison group, was described as average according to their performance on mathematics and statistics tests. As such, we assumed both groups to be comparable.

Data Collection

The data consisted of pre- and posttests from the intervention and comparison group. The pretest was taken in months 7–8 of the school year 2019–2020, from the participating students of both groups. The participating teachers took the test, according to a clear instruction for testing, from their own students during their regular 45-min mathematics lessons. The posttest was taken in months 9–10 of the school year in a similar way, by the teachers during their regular lessons in their own school, see Figure 5.3.

Data Analysis

For the analysis, we first graded the pre- and posttest level scores for the intervention group on the domains of SL with two assessors, and we compared the scores of the intervention group with Dutch baseline achievements from the comparison group. Second, we compared the level scores for both groups with findings by Callingham and Watson (2017).

First, for assessing students’ proficiency on the domains of SL, the pre- and posttest data from the participating 9th-grade students were coded with the level scores 0–6 for SL (Watson & Callingham, 2003), as described in the section on the assessment instrument. To indicate students’ progress for the intervention group, we compared changes in students’ pre- and posttest scores. To indicate students’ achievements, we compared the posttest scores of the intervention and comparison group, and as such, for being taught through the LT or the regular statistics curriculum. Graphical representations were used for data exploration. Several statistical measures were calculated, such as center and spread, and proportions for level scores. For significance, we used paired *t*

tests for comparing pre- and posttest results, one-way ANOVAs for comparing results from both groups, and chi square tests for comparing students' distribution over the levels. For students' proficiency level at SL, we calculated the mean of students' average scores on the AC, GV and SI domain, allowing us to compensate for the inequality in the number of Items per domain.

Second, to further interpret the effects of the LT on students' SL, we compared our findings with the studies by Watson and Callingham and with their distribution of Australian students from Grades 6 to 9 found across the levels for SL. As our assessment instrument was mainly based on their validated tests and hierarchical level construct for SL, we considered the results for our students to be comparable to theirs. In this regard, we expected the distribution in levels for our 9th-graders to be broadly similar to their distribution found for grade 9, and also expected that most students would score on level 3–4 for SL. Concerning the comparison of our students' average level scores with those of Australian students (Callingham & Watson, 2017), estimates for the Australian students' average level score per grade were calculated using the distribution of students across the levels.

For reliability of the analysis, a second coder was asked to independently grade a random set of 5% (250 Items) of the pre- and posttest data with students' reasoning. The third coder agreed on 83% of the codes. Deviating codes, which were limited to one or two levels difference at most, were discussed until agreement was reached. Adjustments in the coding were also applied to the rest of the data.

Results

In this section, we first present the level scores for the intervention group on the domains of SL at the pre- and posttest, and we compare these results with Dutch baseline achievements from the comparison group. Second, to further interpret students' level scores, we compare our results with findings from Watson and Callingham (2017).

Students' Level Scores for SL

Table 5.5 displays students' proficiency on the domains of SL in level scores for the pre- and posttest for the intervention and comparison group, including their progress from pre to post.

When comparing the results for SL on the posttest, a one-way ANOVA between both groups indicated significantly more proficiency on SL for students who followed the LT in comparison with Dutch baseline achievements

from the comparison group, who followed the regular curriculum (+0.33; $F(1, 482) = 24.6, p < .0005$). On the pretest, a one-way ANOVA between both groups indicated the average level score for the intervention group on SL was significantly lower than the score for the comparison group (-0.37 ; $F(1, 482) = 34.9, p < .0005$). The lower score was to be expected, as the intervention group, unlike the comparison group, did not have 9th-grade statistics lessons prior to the pretest. Furthermore, the lower level score of -0.37 for the intervention group relative to the comparison group on the pretest turned out to be almost equal in size to their higher level score of $+0.33$ on the posttest. Since the intervention group had an educational disadvantage of about one school year relative to the comparison group at the pretest, their score on the posttest could be interpreted as almost one school year advantage.

Table 5.5. Students' Mean Level Scores on the Domains of SL at the Pre- and Posttest for the Intervention and Comparison group, Including their Progress from Pre to Post

		Intervention (<i>n</i> = 267)	Comparison (<i>n</i> = 217)	Intervention minus Comparison M(I) – M(C)
		M (SD)	M (SD)	
Pretest	SL	2.60 (0.61)	2.97 (0.68)	-0.37^{***}
	SI	2.45 (0.65)	2.72 (0.71)	-0.27^{***}
	GV	2.07 (0.63)	2.29 (0.58)	-0.22^{***}
	AC	3.29 (1.38)	3.92 (1.31)	-0.63^{***}
Posttest	SL	3.28 (0.69)	2.95 (0.78)	$+0.33^{***}$
	SI	3.34 (0.84)	2.67 (0.84)	$+0.67^{***}$
	GV	2.59 (0.81)	2.38 (0.88)	$+0.21^*$
	AC	3.92 (0.88)	3.80 (1.06)	$+0.12$
Pre to Post	SL	$+0.68 (0.86)^{***}$	$-0.02 (0.73)$	0.70^{***}
	SI	$+0.89 (0.92)^{***}$	$-0.04 (0.71)$	0.93^{***}
	GV	$+0.52 (0.98)^{***}$	$+0.09 (0.94)$	0.43^{***}
	AC	$+0.63 (1.53)^*$	$-0.11 (1.45)$	0.74^{***}

* $p < .05$, ** $p < .005$, and *** $p < .0005$

Note. SL = Statistical Literacy; SI, GV, AC are domains of SL; SI = Sampling and Inference, GV = Graphing and Variation; AC = Average and Chance.

Regarding students' progress on SL, a paired t test between the pre- and posttest for the intervention group indicated the average posttest score was significantly higher than the score on the pretest ($+0.68$, $t(266) = 13.0$, $p < .0005$). The average level score for the comparison group on the pretest was, as expected, not significantly different from their score on the posttest (-0.02 , $t(216) = 0.4$, $p = .65$). Students' results on SL confirmed our conjecture that following the LT had a clear positive effect on students' SL.

Students' Level Scores on the Specific Domains of SL

With regard to the SI domain of SL, on the posttest, a one-way ANOVA between both groups indicated that the level score for the intervention group who followed the LT was considerably higher in comparison with the Dutch baseline achievements from the comparison group ($+0.67$, $F(1, 482) = 75.0$, $p < .0005$). For the comparison group, the pre- and posttest scores on SI were again, as for SL, not significantly different, using a paired t test for differences between the pre- and posttest (-0.04 , $t(216) = 0.9$, $p = .40$). On the pretest, however, when comparing both groups, the score for the intervention group was slightly, but significantly, lower than the level score for the comparison group (-0.27 , $F(1, 482) = 18.5$, $p < .0005$). We did not expect this lower score. Although the comparison group followed the regular statistics curriculum, the SI domain was not offered in the regular lessons, so we expected a similar score for both groups. Concerning students' progress for the intervention group, a paired t test between the pre- and posttest indicated that their average level score on the posttest was considerably higher than on the pretest ($+0.89$, $t(266) = 13.0$, $p < .0005$). The results for the intervention group were in line with our expectations, as we hypothesized that the investigative approach and more complex learning activities for SI as embedded in the LT would support all domains of SL, and SI in particular.

Concerning the GV domain of SL, a one-way ANOVA between both groups indicated that the posttest score for the intervention group was slightly, but significantly, higher than the score for the comparison group ($+0.21$, $F(1, 482) = 7.4$, $p = .01$). Although we expected the intervention group that followed the LT with a focus on SI to progress in the other domains, we did not expect them to reach higher scores than the baseline achievements from students who followed the regular curriculum with a focus on GV and AC. The pre- and posttest scores for the comparison group on the GV domain were not significantly different ($+0.09$, $t(216) = 1.4$, $p = .18$). With respect to students' progress on GV, a paired t test between the pre- and posttest for the intervention group indicated that their posttest score was significantly higher than their

pretest score ($+0.52$, $t(216) = 8.7$, $p < .0005$). Regarding students' level for the GV domain, it is important to note that the average maximum scores for the test Items used in this domain were, as elaborated earlier in the methods section, considerably lower than for Items in the other domains. Therefore, the GV level score cannot be used for comparison with other domains.

For the AC domain, a one-way ANOVA between both groups indicated that the posttest score for the intervention group that followed the LT was comparable with the Dutch baseline achievements from the comparison group ($+0.12$, $F(1, 482) = 1.8$, $p = .18$). As for the GV domain, the posttest score on the AC domain for the intervention group was higher than we expected, as the LT focused on SI. Concerning students' progress, a paired t test between the pre- and posttest for the intervention group indicated that their posttest score was significantly higher than their pretest score ($+0.63$, $t(266) = 15.8$, $p < .0005$). The comparison group scored similar on both the pre- and posttest (-0.11 , $t(216) = 1.1$, $p = .26$). The findings on the domains for SL confirmed our conjecture that following the LT had a clear positive effect on students' SL and SI, and more moderate effects on the GV and AC domains.

Students' Level Score on SL in Comparison with those of Australian Students

To further interpret the proficiency of students, we compared our results with those of Australian students (Callingham & Watson, 2017). In doing this, we compared the distribution of students over the levels for SL, and we compared students' average level scores on SL. The distribution of students over the levels of SL on the pre- and posttest, is presented in Table 5.6.

Table 5.6. Students' Distribution over Levels of SL

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Pretest SL intervention ($n = 267$)	11.2%	25.5%	56.6%	6.7%	-	-
Posttest SL intervention ($n = 267$)	1.1%	13.5%	44.6%	39.3%	1.5%	-
Pretest SL comparison ($n = 217$)	4.1%	17.5%	53.9%	23.5%	-	-
Posttest SL comparison ($n = 217$)	6.5%	20.3%	46.1%	27.2%	-	-

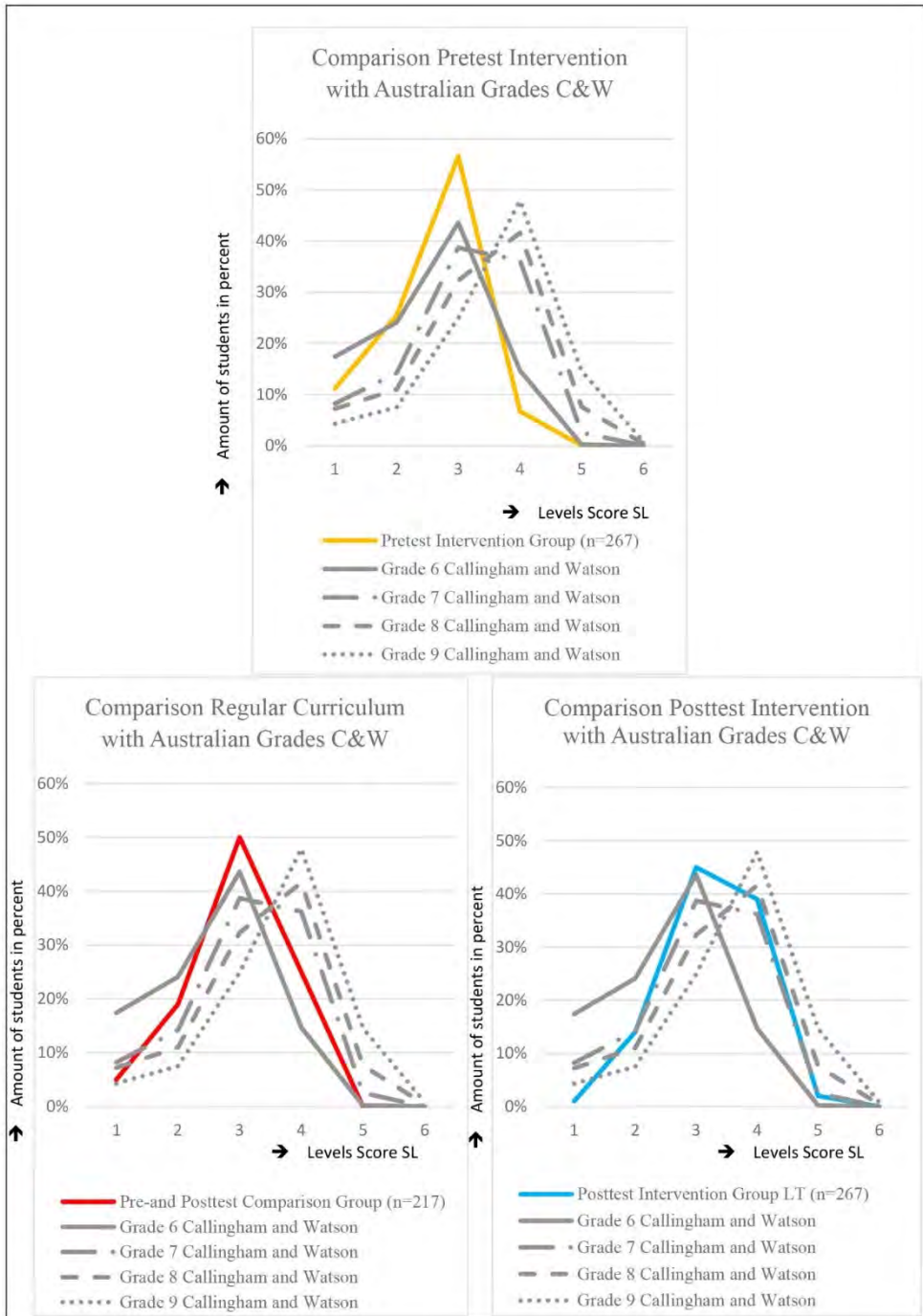


Figure 5.4. Comparison of students' level scores on statistical literacy (SL) with Australian grade results from findings by Callingham and Watson (2017)

The pretest scores for the intervention group corresponded most closely to the performance of Australian students in grade 6 (Callingham & Watson, 2017) and, as such, were lower than we expected. Figure 5.4 visualizes the comparison of students' distribution over the levels. A chi-squared test on the distribution over levels in percentages, between the pretest score for the intervention group and each Australian grade 6 to 9, confirmed the highest p value, and with that the best fit, for grade 6 ($\chi^2(4) = 6.26, p = .18$). The pretest mean level score for the intervention group 2.60 (0.70) also corresponded to the estimate of the mean level score for Australian grade 6. The estimates per grade were calculated using the distribution of their students across the levels. Table 5.7 summarizes the comparison of both groups with Australian grade results, based on the distribution of students over the levels and average level scores. Regarding the posttest score for the intervention group, the results corresponded most closely to Australian Grades 7–8. The chi-squared test confirmed the similarity between the posttest scores for the intervention group and Grades 7–8, as the highest p values found were $\chi^2(4) = 6.2, p = .184$ and $\chi^2(5) = 11.3, p = .05$, for Grades 7 and 8 respectively. The posttest average level score for the intervention group 3.28 (0.69) also corresponded most closely to the estimate of the level score for Australian Grade 8 (3.3). The pre- and posttest scores for the comparison group were quite similar. According to the findings by Callingham and Watson, the scores for the comparison group corresponded most closely to Australian Grades 6–7. The chi-squared test confirmed the similarity, as the highest p values found were for Australian Grades 6 and 7 ($\chi^2(4) = 9.3, p = .05$ and $\chi^2(4) = 5.8, p = .22$ respectively). The mean level score for the comparison group on the pretest 2.97 and the posttest 2.95 also corresponded to the estimate of the level score for Australian Grades 6-7, respectively 2.6 and 3.1.

Concerning the effects of the LT, the posttest score on SL for the intervention group that followed the LT appeared to be more advanced than the score for the comparison group. Moreover, from the comparison with findings by Callingham and Watson (2017), the advantage for the intervention group on SL corresponded again, as in our earlier findings, with about one school year higher. Furthermore, the calculated estimates of students' average level score per grade from the study of Callingham and Watson indicated that students' progress per year from Grades 6 to 9 is roughly 0.25. When we compare the posttest SL level score for the intervention group 3.28 (0.78) with the score for the comparison group 2.95 (0.69), the difference of 0.33 between both groups again corresponds to a level difference of more than one school year.

Table 5.7. Students' Proficiency in Comparison to Grade-Results of Australian Students (Callingham & Watson, 2017), Based on the Distribution of Students over the Levels and the Average Level Scores

Dataset	Statistics education	Distribution and average level similar to that found in grade X by C&W
Pretest Intervention group ($n = 267$)	No 9th-grade statistics lessons	Grade 6
Pre- and posttest Control group ($n = 217$)	Regular 9th-grade curriculum	Grade 6-7
Posttest Intervention group ($n = 267$)	Learning Trajectory	Grade 7-8

Conclusion and discussion

The aim of this study was to evaluate the effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy, and on their SI in particular. Theories of informal statistical inference complemented by ideas of growing samples and repeated sampling, guided the design of the LT.

Based on students' level scores on the pre- and posttest and the comparison with Dutch baseline achievements from the comparison group, we conclude that the LT had a significant positive effect on students' SL, and in particular on the SI domain. Furthermore, students who were taught using the LT showed significant improvements on the other domains of SL as well. With regard to SL, the posttest results showed significantly more proficiency for students who followed the LT in comparison to the Dutch baseline achievements from the comparison group. Regarding the domains of SL, students' level score on the SI domain for the intervention group was significantly higher than for the comparison group. Furthermore, the scores for the intervention group on the GV domain—graphing and variation—were slightly, but significantly higher than for the comparison group, and their scores on the AC domain—average and chance—were comparable with the national baseline achievements from the comparison group. In comparing our results with those of Australian students (Callingham & Watson, 2017), the posttest results for the intervention group corresponded most closely to Grades 7–8, while the national baseline achievements from the comparison group equaled Grades 6–7.

Furthermore, the lower level score on SL on the pretest for the intervention group relative to the comparison group turned out to be almost equal in size to their higher posttest score. Since the intervention group had an educational disadvantage of about one school year relative to the comparison group at the pretest, their score at the posttest could be interpreted as an almost one school year lead. Moreover, the comparison with findings by Callingham and Watson also reflected a one school year lead, as the results for the intervention group were most similar to Grades 7–8, while those for the comparison group were more equivalent to Grades 6–7.

In discussing these conclusions, there are a few points to consider. The first involves the low level of proficiency of our students on SL relative to Australian students (Callingham & Watson, 2017). We expected our students to score one the posttest on grade 9 level, and not on Grades 6–7 and Grades 7–8, for the intervention and comparison group respectively. These lower scores may be due to the fact that our Dutch pre-10th-grade statistics curriculum is more limited than the Australian curriculum for students in Callingham and Watsons' research (<https://www.australiancurriculum.edu.au/>). Another issue in this respect is that the average maximum attainable score on the GV Items on both tests was lower (about 3.7) than for the other domains (about 5.5), which negatively affected students' overall SL scores. When we compensate for the lower GV Item scores, the SL average level scores of participating students increase by about 0.3. When we then compare the adjusted SL scores with the Australian grade-results, the grade-results for our students increase with almost one school year, and, as such, were closer to our expectations.

The second point considers effect sizes. The use of effect sizes is complex and disputed, and only makes sense for comparing similar studies (Bakker et al., 2019; Cohen, 1988; Schäfer & Schwarz, 2019; Simpson, 2017). The only study we could find that is similar enough to judge the differences found is Novak (2014), since it shares content and design with ours. Novak's study involved the evaluation of a simulation-based intervention for an introductory statistics course at the university level. A pre-post research design was used with two random intervention groups and a total of 64 students, where both groups followed a slightly different simulation-based intervention. By comparing the pre- and posttest, Novak found a significant learning effect on students' statistical knowledge with Cohen's $d = 0.45$, and the effect on students' conceptual knowledge was approaching significant with Cohen's $d = 0.18$. In comparing our results with theirs, the effects of the LT on students' SL and on the SI domain appeared considerably positive with Cohen's $d = 0.90$ and

Cohen's $d = 1.12$ respectively, and we also found clear positive effects on the GV and AC domains.

Limitations of our study are the following. First, we worked with students from the pre-university level, the 15% best performing students of our educational system. As such, the results in this research are not generalizable to regular classrooms without further research. Second, the intervention group took the posttest close after following education based on the LT. The comparison group completed their 9th-grade statistics lessons in the first part of the school year. By the time of the pretest, conducted 1–3 months after completing their statistics lessons, the students from the comparison group had possibly forgotten specific topics that not often recur, such as the median. To identify possible changes in their performance due to the time interval of a few months, the pre- and posttest were taken at two separate moments with an intervening time period of about two months. The tests were taken at the same time of the school year as for the intervention group to limit influences such as natural growth, in months 7–8 and 9–10 respectively. Taking into account the time span between education and assessment, we could also compare the pretest results of the comparison group (1–3 months after education) with the posttest of the intervention group (1–2 months after education). However, as the performances of the comparison group on both tests were comparable, this does not affect our conclusion. Third, we did not examine differences due to instructors' or students' background. We recommend taking both issues into account in future research.

We present two points for recommendations. First, in this study, the identified levels of SL by Watson and Callingham (2003, 3004) proved well applicable for evaluating the effects of the LT. The development of a pre- and posttest, consisting of Items from validated tests—mainly from Watson and Callingham—supplemented by equivalent newly designed SI Items, enabled us to assess students' SL, and their SI in particular. Both newly designed and existing test Items were found appropriate, with a Cronbach's alpha greater than .84 on the pre- and posttest. In analyzing the results, the levels of SL appeared useful to examine students' proficiency. Furthermore, the findings by Callingham and Watson (2017) proved useful for interpreting students' results, and, with that, the effect of the LT. Therefore, we recommend researchers and educators who intend to investigate the SL of secondary school students to use the levels of SL by Watson and Callingham for assessing and evaluating students' results.

Second, for the participating teachers of the intervention group, implementing the LT required considerable effort. In our study, 11 teachers from five different schools were willing to invest in the LT. The load for teachers from the comparison group was limited to administering two tests, making it easier for teachers to participate. Using a comparison group was of added value to interpret the intervention group results. Therefore, we recommend researchers and educators interested in the effects of an LT, who are for practical reasons confined to an intervention group with considerable effort for participating teachers, to consider the use of national baseline achievements from a comparison group. Furthermore, as highlighted by several researchers, much work remains to be done to obtain a good understanding of how to assess the practical and substantive effects of educational interventions, this study contributes by presenting a pre-post research design in which students' results were compared with Dutch baseline achievements from a comparison group and with findings from international studies.

To end with, the LT highly affected students' performance on SL and SI, and we also indicated significant positive effects for the AC and GV domains. Although the LT was not focused on the latter two, the investigative approach and more complex learning activities for SI as embedded in the LT appeared to have a positive effect here as well. These findings suggest that current statistics curricula for grades 6–9, usually with a strong descriptive focus, can be enriched with an inferential focus—at least for the pre-university level. The benefit will be that students learn more about inference and not less about the other domains of statistical literacy, to anticipate subsequent steps in students' statistics education.



General Discussion

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write (attributed to H. G. Wells (1866–1946), in Watson, 2006, p. vii)

Introduction

The field of statistics (education) is changing rapidly. Over the past decades, the use of data in society increased tremendously due to technological innovations that provide opportunities to easily collect, store, analyze and represent data. As citizens and professionals, we are confronted daily with statistical information, which requires us to be statistically literate: to be able to interpret, critically evaluate, and communicate about statistical information and messages (Gal, 2002). These changes in the field of statistics necessitate an educational emphasis on developing statistical literacy and learning from and with technology.

Making inferences is the main goal of statistics. As such, the ability to draw conclusions about processes and populations based on samples is essential. However, research in statistics education shows how challenging this is for students (Castro Sotos et al., 2007; Konold & Pollatsek, 2002). In many countries, including the Netherlands, statistical inference is not taught until Grade 10 or higher. Most students' difficulties relate to a limited understanding of key concepts required for statistical inference—such as sample, variability, and distribution. A way to overcome these problems involves offering informal statistical inference, before the transition to more formal inferential statistics (Makar & Ruben, 2009; Paparistodemou & Meletiou-Mavrotheris, 2008; Van Dijke-Droogers et al., 2020; Zieffler et al., 2008). In general, this informal approach focuses on ways in which students without knowledge of formal statistical procedures, such as hypothesis testing, use their statistical knowledge to underpin their inferences about an unknown population based on observed samples. Statistical modeling activities with educational digital tools facilitate—on an informal level—the exploration of concepts for statistical inference (Biehler et al., 2013; Manor & Ben-Zvi, 2015). These digital tools offer opportunities to easily visualize and explore concepts as sampling, variability and distribution.

In many countries, including the Netherlands, the statistics curriculum is evolving from descriptive statistics in the early years to more complex inferential statistics later on. Little is known about how to embed (informal) statistical inference earlier in current curricula. As such, there is a need for efficient learning trajectories, and knowledge about crucial steps in such a

trajectory, that can extend the descriptive curricula in early years with inferential activities. To address this, the aim of this study was to gain knowledge about a theoretically and empirically based learning trajectory to introduce 9th-grade students to statistical inference. We addressed the following guiding research question:

How can a theoretically and empirically based learning trajectory introduce 9th-grade students to statistical inference?

The formulated research question involved both the design and evaluation of a learning trajectory. A design-based research approach seemed suitable to address this dual question. Three cycles were completed evolving in size of the trajectory and implementation scope. Furthermore, between cycles 2 and 3, a domain-specific case study was conducted into learning from and with technology.

Research Overview and Main Findings

Currently, one of the five Content Standards of the US National Council of Teachers of Mathematics (NCTM) encompasses the following specific expectations on statistical inference for grades 9–12:

Each and every student should use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions. Furthermore, students should understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference (NCTM, n.d.)

In Chapter 2 we discussed the first research cycle concerning the first three steps of a learning trajectory for introducing statistical inference. In this starting phase of the research, defining design guidelines for a learning trajectory appeared challenging. Based on a literature study, personal experience as a teacher–researcher, and brainstorm sessions with a focus group, design guidelines were distilled. The focus group consisted of an experienced teacher, two teacher-researchers, a teacher educator, two experienced researchers, a statistician, and an educational developer. A hypothetical learning trajectory was developed, based on the guidelines distilled. The two main ideas incorporated in the design of the trajectory were repeated sampling with a black box and the use of simulation software for statistical modeling. The trajectory

aimed to introduce students to the key concepts for statistical inference. As such, we addressed the following research question:

RQ1: How can repeated sampling with a black box introduce 9th-grade students to the concepts of sample, frequency distribution, and simulated sampling distribution?

To empirically evaluate the hypothetical learning trajectory, we conducted a teaching experiment with twenty 9th-grade students. Indicators of observable learning behavior of students that supported the hypotheses were drawn up for each learning step of the trajectory. The results showed that most indicators were observed. We assume that the strong coherence and construction between the three learning steps stimulated the students to go through the steps fluently. From their concrete black box experiences in step 1, by visualizing the scaling up of this experiment in step 2, students could easily make the transition to interpreting the simulated sampling distribution in step 3. Figure 6.1 illustrates the similarity between steps 1 to 3. These first three steps of the learning trajectory provided students insight into how a sampling distribution can be constructed and how it can be used as a model for interpreting variation and uncertainty. These findings suggested a promising way to introduce students to (informal) statistical inference.

In Chapter 3, we presented a case study into learning from and with technology. Earlier studies indicated that the use of digital tools for statistical modeling offers means for introducing statistical inference, as those tools have the potential to deepen students' conceptual understanding of statistics and probability (Pfannkuch, Ben-Zvi, & Budgett, 2018). Such educational digital tools, for example TinkerPlots, provide opportunities for statistical reasoning with data, as students build statistical models and use these models to simulate sample data (Biehler et al., 2017). As such, the use of statistical modeling seemed promising. In this study, we focused on *how* students' statistical modeling processes in TinkerPlots fostered their development of statistical concepts. We particularly examined 9th-grade students' intertwined development of learning techniques for using TinkerPlots and their understanding of statistical concepts, by using the theoretical perspective of instrumental genesis (Artigue, 2002). In this study, we addressed the following question:

RQ2: Which instrumentation schemes do 9th-grade students develop through statistical modeling processes with TinkerPlots

and how do emerging techniques and conceptual understanding intertwine in these schemes?

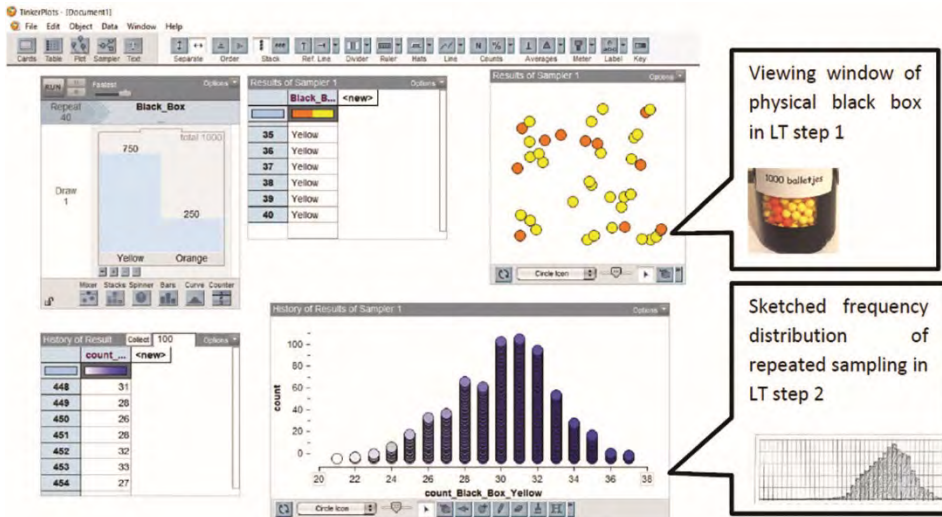


Figure 6.1. Similarity between the digital environment of TinkerPlots in step 3 and the physical black box experiment in steps 1 and 2

A suitable phase to examine students' instrumental genesis was after the introduction of the tool and the concepts, when acquired knowledge is applied in new situations, in step 4 of the learning trajectory. The data for this study consisted of video and audio recordings of two laboratory sessions with a total of 28 students. In particular, we analyzed how the development of digital techniques and the learning of statistical concepts were intertwined in the instrumentation schemes that students developed. We observed a strong intertwining between students' emerging technical and conceptual understanding. Techniques for using TinkerPlots helped students to explore context-independent technical patterns that facilitated a conceptual shift from a *model of* to a *model for* (Gravemeijer, 1999). Vice versa, students' conceptual understanding led them to explore more advanced digital techniques. These findings demonstrated that investing in learning digital techniques in the meantime had a positive effect on the development of statistical concepts.

Chapter 4 considered the third research cycle and reported on the design, implementation and evaluation of the whole 8-step learning trajectory. Findings from the first two research cycles and the case study were elaborated in the (re)design of the trajectory in cycle 3. Research cycle 2—the implementation of

the trajectory in three classes at different school—is not elaborated in this thesis; however, the findings were incorporated in cycle 3. The aim of this third cycle was to empirically substantiate the designed learning trajectory by analyzing students' progression during a large-scale intervention. We were specifically interested in how the eight steps of the trajectory fostered students' understanding of statistical inference. Our focus was on both students' learning processes and on their achievements for statistical inference. As such, the following research questions were addressed:

RQ3.1: What are the specific effects of the designed Learning Trajectory (LT) on students' understanding of statistical inference, in terms of the intended LT-step related learning goals?

RQ3.2: How do the designed steps of the learning trajectory foster students' learning processes?

The designed learning trajectory included eight learning steps, divided into two similar sequences of four: (1) experimenting with a physical black box, (2) visualizing distributions, (3) investigating sampling distributions using simulation software, (4) interpreting sampling distributions for inferences in real-life contexts. Steps 1 to 4 included only categorical data and steps 5 to 8 regarded numerical data.

Finding participating teachers for the intervention was challenging, as the curriculum in Grade 9 allowed little time for adding an extensive learning trajectory like this. However, after some promotional activities, we were able to implement the learning trajectory in an intervention among 267 students in 13 classes at different schools. A pre- and posttest were developed to evaluate students' performance on the intended LT-step related goals for statistical inference, and a comparison group of 217 students—who attended the regular 9th-grade curriculum—was used to indicate the results found. The analysis of test results demonstrated that students' understanding of statistical inference as addressed in the coupled LT steps—in LT steps 1 and 5 on using samples, in LT steps 2 and 6 on visualizing distributions, in LT steps 3 and 7 on repeated sampling and effect of sample size, and in LT steps 4 and 8 on solving real-life problems—was significantly higher among students who took part in the LT than among students who followed the regular curriculum. In addition, the analysis of students' worksheets, accompanied by teachers' and researcher's notes, showed that the eight steps of the learning trajectory fostered students' learning processes. As such, the results empirically substantiated the

theoretically designed learning trajectory. Again, ideas of repeated sampling with a black box and statistical modeling proved fruitful for introducing statistical inference. Both ideas also have potential for embedding in more complex follow-up activities, such as testing hypotheses and comparing groups. These findings suggested that current statistics curricula with a descriptive focus can be extended with an introduction to statistical inference.

In Chapter 5 we presented the effects of the learning trajectory on students' proficiency on all domains of statistical literacy, and inferences in particular. Although the trajectory concentrated on statistical inference, we conjectured that the focus on more complex inferential activities would have a positive effect on students' understanding of all domains of statistical literacy. In this chapter, we addressed the following research question:

RQ4: What are the effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy?

For the evaluation of the effects of the learning trajectory, a pre-post research design with the intervention group ($n = 267$) from the third research cycle was used. To indicate the learning effects, students' test results were compared with a national baseline and international findings. For the national baseline, we used the results of a comparison group ($n = 217$) that followed the regular 9th-grade curriculum, and the international comparison was done using an Australian study with similar test design. The comparison with the national baseline showed that the intervention group scored significantly higher on statistical literacy, and in particular on the domain of statistical inference. The comparison with the international study showed that the posttest results of the intervention group were similar to the results for Grades 7–8 of the international study, while the results of the comparison group were similar to those of Grades 6–7.

Finally, the results indicated that the learning trajectory had a strong positive effect on students' statistical literacy, and in particular on the domain of statistical inference. We also found significant positive effects for the other two domains of statistical literacy—graphing and variation, and average and chance. We assumed that the inquiry-based approach and the more complex learning activities for statistical inference, as embedded in the learning trajectory, brought about the positive effect on the other domains. These findings suggest that current statistics curricula for grades 6–9, usually with a strong descriptive focus, can be enriched with an inferential focus—at least for preparatory university education (VWO). The benefit will be that students learn more about inference and not less about the other domains of statistical literacy, to

anticipate subsequent steps in students' statistics education. See Figure 6.2 for an impression of the intervention(s).



Figure 6.2. Impressions of the intervention

Contributions

This research contributes theoretical insights that are closely related to a practical educational design. As Lewin (1952) wrote: “There is nothing more practical than a good theory” (p. 169). Scientific knowledge is gained about learning and teaching statistical inference, through designing and evaluating a learning trajectory for 9th-grade students. In addition, methodological knowledge is gained about how to design and evaluate an innovative learning trajectory through design-based research, ending with a quantitative analysis in the last cycle.

Scientific Contribution: Introducing Statistical Inference

As the field of statistics and its education are changing rapidly, knowledge about efficient learning trajectories is needed for the successful and sustainable implementation of curriculum changes (Biehler et al., 2018). In this regard, Ben-Zvi, Gravemeijer, and Ainley (2018) express the need to think about

learning environments and their design to support sustainable change in students' understanding of key statistical ideas.

Engaging in Statistical Inference Impacts on Statistical Literacy

Although statistical inference is considered a more complex domain of statistical literacy, this study demonstrated that the designed learning trajectory for statistical inference had a significant positive effect on all domains of statistical literacy. As such, engaging in (informal) inferential activities also promoted students' capacity on other statistical literacy domains. This insight into a joint development of (informal) statistical inference and literacy, allows in educational practice for an early introduction of statistical inference. An early introduction can support a sustainable change in students' understanding of statistical concepts required for both making inferences and statistical literacy.

Currently, the Dutch curriculum, as in many other countries, evolves from descriptive statistics in the earlier years to an inferential focus later on. In early years—pre-10th grade—the focus is on the statistical literacy domains of graphing and variation, and average and chance. Later on, the domain of statistical inference is given attention. The results of this research advocate an earlier introduction of statistical inference. The positive effects of the learning trajectory on the other domains of statistical inference are presumably due to the inquiry-based approach of the learning trajectory, in which all phases of the statistical investigation cycle are addressed several times—that is, posing a question, collecting data, analyzing data, to answer the question posed. This is consistent with previous studies and theories that advocate a holistic approach (Ainley, Pratt, & Hansen, 2006; Franklin et al., 2007; Lehrer & English, 2017; Van Dijke-Droogers, Drijvers, & Tolboom, 2017).

Networking Theories

Statistics education has matured into a discipline distinct from mathematics education, with its own perspectives on teaching and learning (Groth, 2015). Although statistics education has its own character, in many countries it is part of the secondary mathematics curriculum, including in the Netherlands. Coordinating perspectives from statistics and mathematics through boundary interactions between the two can strengthen both areas of education (Groth, 2015). Given the landscape of strategies for connecting theoretical perspectives (Prediger, Bikner-Ahsbahr, & Arzarello, 2008), this research contributes by locally integrating mathematical ideas into statistics education research, that is, Realistic Mathematics Education theory and the perspective of Instrumental Genesis.

Connecting mathematics and statistics education: Realistic Mathematics Education

As stated by Ben-Zvi et al. (2018), theories of constructivism and Realistic Mathematics Education (RME) (Freudenthal, 1983) provide a conceptual foundation to guide the design of learning environments for statistics education. According to the constructivist theory, new knowledge and understandings are grounded on students' prior experiences, understandings, and practices (Cobb, 1994; Piaget, 1978; Vygotsky, 1978). RME provides domain-specific design heuristics that encompass guided reinvention, didactical phenomenology, and emergent modeling (Gravemeijer, 2004); guidelines that serve for the design of mathematical learning experiences and that proved useful in our study.

Based on these theories, repeated learning experiences with statistical concepts were incorporated in the design of the learning trajectory. Within and between each sequence of four learning steps, learning experiences with the key concepts of sample, variability and distributions, were embedded using the black box paradigm. Starting in learning steps 1 and 2 with the physical black box experiment, students developed a beginning understanding of the key concepts. Initially, the distribution was used as a visualization or model *of* sample results found, and gradually in steps 3 and 4, students were able to use the distribution as a model *for* determining the probability of particular sample results. The strong similarity between the physical black box activities and the modeling activities in the digital environment of TinkerPlots facilitated the connection of the model to the real world (Konold & Kazak, 2008; Patel & Pfannkuch, 2018). In the following learning steps, the black box served as a guiding paradigm in students' reasoning and in the teacher's instruction of key concepts, particularly during modeling real-life phenomena.

Several studies have indicated that reasoning and interpreting sampling distributions is difficult (Batanero et al., 1994; Castro Sotos et al., 2007; Chance, delMas, & Garfield, 2004). From the findings in this research, it appeared that students could develop the key concepts of statistical inference, including interpreting sampling distributions, in a short period of time by using black box sampling as a guiding activity. The design of the black box paradigm was based on the RME design heuristics for guided reinvention—for example, exploring sampling variability and using repeated sampling; for didactical phenomenology—exploring context-independent patterns; and emergent modeling—the conceptual shift from a *model of* to a *model for*. As such, the RME perspective strengthened the design of the learning trajectory for statistics education.

Connecting theories on statistics education and on instrumental genesis

The use of digital tools is a shared problem space (Groth, 2015) between mathematics and statistics education. It is known from research in mathematics education that once digital tools are used during the learning process, the development of conceptual understanding becomes intertwined with the emergence of techniques for using the digital tool. A theoretical perspective that is useful to investigate this intertwining is instrumental genesis (Artigue, 2002; Drijvers, Godino, Font, & Trouche, 2013).

Using the theoretical perspective of instrumental genesis enabled us to unravel students' development of instrumentation schemes consisting of digital TinkerPlots techniques and conceptual understanding. The scheme developments revealed a strong intertwining in both directions between learning digital techniques and developing conceptual knowledge. The instrumental genesis perspective appeared helpful to demonstrate *that* and *how* investing in learning digital techniques simultaneously had a positive effect on the development of statistical understanding (see Chapter 3). Although we focused on statistical modeling processes using TinkerPlots, we consider our findings on the intertwining of emerging digital techniques and conceptual understanding applicable to the broader field of statistics education, and to other educational digital tools as well. Digital tools for other areas in statistics education also structure and guide students' thinking by providing specific options for entering parameters and commands and by facilitating explorative options that may strengthen students' conceptual understanding. As such, the perspective of instrumental genesis seems applicable for research into learning from and with technology in statistics education.

To conclude, this research contributes to insights into the joint development of statistical inference and statistical literacy by demonstrating that engaging in (informal) inferential activities simultaneously may promote students' capacity in other statistical literacy domains. Furthermore, this research presents fruitful insights by connecting theories of mathematics education research, that is, Realistic Mathematics Education and instrumental genesis, into statistics education research.

Methodological Contribution: Design-Based Research

A design-based research approach (Bakker, 2018; McKenney & Reeves, 2012) proved effective for the design and evaluation of the innovative learning trajectory. Design-based research consists of a cyclical process in which educational materials are designed, implemented in teaching practice, and evaluated, for subsequent cycles of redesign and testing. Starting from a theoretically informed design, the trajectory was empirically tested in several cycles for further improvement.

Cyclic Scaling Up the Length of the Learning Trajectory

A cyclic scaling up in length of the trajectory allowed for a constructivist approach in the development of the learning trajectory. This research aimed at both the design and the evaluation of a learning trajectory. A constructivist approach enabled us to answer initial questions of: “What do we as educational designers want students to construct?” and “How do we create learning trajectories in which students construct what we want them to construct?” (Cobb, 1994). An initial focus on the first learning steps, allowed for monitoring students’ changing conceptions that provided insights as starting points for following steps. Although the design included a complete 8-step learning trajectory from the start in cycle 1, our focus for analysis and evaluation was initially on the first three learning steps. These three steps introduced the key concepts: sample, variability, and distributions, which were fundamental to subsequent steps in the trajectory. This focus on the initial steps enabled a specific examination of whether and how the paradigm of the black box and statistical modeling promoted students’ learning in steps 1 to 3. The results from these steps informed about the starting point in learning step 4. In cycle 2, the full learning trajectory was again conducted, with a focus on step 4. From the results, the need emerged to further investigate learning with and from technology in the fourth step—more specifically into the application of statistical modeling with TinkerPlots and students’ development of statistical concepts. As the construct and coherence in learning steps 1 to 4 were similar to steps 5 to 8, regarding categorical and numerical data respectively, the first two cycles and the case study provide knowledge about the whole trajectory.

Cyclic Scaling Up the Number of Participants

As addressed by Arnold et al. (2018), scalability is important in research regarding learning trajectories. However, experimenting with innovative learning trajectories in educational practice is complex, in particular on a large scale. Maass et al. (2019) stated: “Implementing innovations in one classroom can be a challenging endeavor, and it is even more demanding across a whole

school. However, it becomes exponentially more challenging when scaling up an innovation aims to reach many schools” (p. 304). In the initial cycle(s) of design-based research, a design is still in the experimental phase and it is unclear whether and how the learning trajectory will work. Teaching time with students is scarce and teachers want to fill their teaching time efficiently. Experimental aspects make it hard to ensure the effectiveness of the trajectory. To reduce this problem, we chose a small-scale start in cycle 1 with subsequent scaling up in cycles 2 and 3. For this purpose, cycle 1 was conducted in one class with 20 students, taught by the teacher-researcher. The teacher-researcher was able to make adjustments during the teaching practice to ensure the students’ learning efficiency. Based on the results, a (re)design was developed for scaling up in cycle 2 to three classes with a total of 60 students. The three participating teachers were not involved in the design of the trajectory and were aware of the experimental aspect. The researcher was present during each lesson as an observer and the teacher(s) and researcher discussed extensively before and after each lesson, to ensure the intended learning goals were addressed. Cycle 2 was not elaborated as a separate study in this thesis, but the results were incorporated in the (re)design for cycle 3.

In cycle 3, the learning trajectory was evaluated on a larger scale. However, scaling up was an intensive process, as it required all educational materials to be unambiguous, complete, and feasible to minimize discrepancies in implementation. Also, the participating teachers had to be trained in several sessions for implementing the trajectory as intended. Despite the fact that researchers and teachers both strived for an effective learning trajectory, they aimed for slightly differing goals. On the one hand, the teachers’ goals were specifically focused on their students’ learning achievements, for which a fully developed learning trajectory was preferred. The researchers, on the other hand, wanted to gain new knowledge about crucial elements of the trajectory, which meant that experimental components—the effectiveness of which was not yet certain—were also incorporated in the design. To identify possible tension or misunderstanding due to these differing goals, we kept in close contact with the participating teachers during the large-scale intervention.

A Quantitative Evaluation in the Final Cycle

The third cycle aimed at quantifying the effects of the learning trajectory on students’ learning. Measuring students’ performance on pre- and posttests, for an intervention and comparison group, with the use of statistical methods is a convincing way to make claims about the effects of a learning trajectory. However, in design-based research, a quantitative approach is not commonly

used. In experiments, the focus is typically on the learning product rather than the process. As indicated by Savelsbergh et al. (2016), a common concern is that for experimental studies that only report pre-post results, it remains unclear to the reader how to benefit from the intervention reported. Evaluating the effects of a learning trajectory requires finding out *that* the trajectory works and also *how* it works; a focus on both process and product. To this end, the quantitative analysis in cycle 3 focused on both students' global achievements on statistical literacy—and inferences in particular—and on specific step-related goals for statistical inference.

For the evaluation of the trajectory in cycle 3, a pre- and posttest were developed, inspired by the work of Watson and Callingham (2003, 2004) on testing statistical literacy at the school level. A total of 267 students in 13 classes at different schools participated in the intervention. Students' performance on the test were used to verify *that* the trajectory works. To indicate students' learning progress, the results obtained were compared to both national and international findings. Both comparisons confirmed that participating in the designed trajectory had a significant positive effect on students' statistical literacy, and in particular on the domain of statistical inference (see Chapter 5). To analyze *how* the trajectory works, we specifically examined the effects of the 8-step learning trajectory on students' understanding of step-related goals for statistical inference. As such, we analyzed students' progression during the large-scale intervention, using students' worksheets and their test scores on learning step-related test items (see Chapter 4).

To conclude, this research contributes to methods of educational research by presenting how complexities involved in experimenting with innovative educational materials can be overcome by using design-based research with cyclic scaling up—in number of participants and length of the learning trajectory. Evolving from a small-scale qualitative focus in cycles 1 and 2 to a more quantitative large-scale approach in cycle 3 enabled us to develop an empirically based learning trajectory—that is, to design a learning trajectory and to evaluate *that* and also *how* it works.

Limitations

This thesis presents a learning trajectory for introducing statistical inference that proved to be effective for Dutch 9th-grade students in the pre-university stream. In designing this trajectory, we opted for an approach with a black box experiment combined with statistical modeling. This was an approach that proved beneficial. However, one might wonder whether other approaches for

(informal) statistical inference may also provide positive results, such as starting from meaningful data contexts (Franklin et al., 2007; Pfannkuch, 2011) or expanding the growing samples principle (Bakker, 2004). Our research did not look at this question. However, we demonstrated that it is possible to introduce a complex domain of statistical literacy, such as inferences, at a younger age.

Evaluating the effect of the trajectory by using pre- and posttests for an intervention and comparison group, raises issues of generalization and causality. The use of evidence-based randomized controlled trials on the effectiveness of educational materials has its limitations (Olsen, 2004). Although no strict statistical claims from sample to population and causal effects can be derived, the quantitative analysis in research cycle 3 does provide insight into the effects of the learning trajectory on the achievements of the students we worked with. For the intervention group, we worked with teachers who volunteered to participate. These teachers were willing to invest time and effort in the implementation of an innovative statistics project, and as such were above average motivated. These teachers were inexperienced in teaching statistical inference, as this is not offered in the current pre-10th curriculum, and inexperienced in teaching from and with technology. They implemented the learning trajectory for the first time, which made them inexperienced and unfamiliar with the learning materials. When repeated in a following year, with the same teachers, it will probably be easier for them to implement. The effect of a learning trajectory strongly depends on the way it is implemented by the teachers. We consider the positive effects found, for 267 students with thirteen teachers at different school, as a strong indication that the learning trajectory works—when implemented as intended. To investigate the effect of the learning trajectory, we focused on students' cognitive achievements. We did not address the effects of the trajectory on other aspects related to students' learning, such as involvement, autonomy, relevance, commitment, engagement, motivation and expectations.

Implications for future research and educational design

Based on the findings in this study, we suggest the following directions for future research and educational design.

Joint Development of Statistical Inference and Statistical Literacy

The results in this research demonstrated a joint development of statistical inference and statistical literacy, for the group of students we worked with—that is, for 9th-grade students in pre-university education. These students had basic statistical knowledge from their descriptive statistics lessons in Grades 7 and 8,

such as using graphs and calculating measures of center and spread. Engaging in (informal) inferential activities in earlier years or in other educational levels may also promote a joint development. However, our research was focused on a specific educational level and age. More research is needed to investigate this joint development for students with less prior knowledge and on other educational levels.

Networking Theories for Mathematics and Statistics Education

This research presented fruitful insights by connecting theories of mathematics education research—that is, Realistic Mathematics Education and instrumental genesis—into statistics education. When integrating perspectives from two educational areas, insights into both areas can be strengthened. Networking theories is especially urgent for mathematics and statistics education, where there are several shared problem spaces (Groth, 2015). On top of that, in educational practice, mathematics and statistics lessons are often taught by the same mathematics teacher, for whom integrating knowledge from both areas can be beneficial. We therefore recommend more research with networking theories to strengthen insights for both mathematics and statistics education.

The Lens of Instrumental Genesis on Using Technology in Statistics Education

In this research, we focused on students' statistical modeling processes using TinkerPlots. The perspective of instrumental genesis helped to gain insight into students' learning from and with technology. Revealing students' instrumentation schemes provided insight into how the learning of the tool related to the development of statistical concepts. We consider our findings on the intertwined development of digital techniques and conceptual understanding, to be applicable to the broader field of statistics education, and perhaps also when using other digital tools. More research is needed to explore the applicability of instrumental genesis for other topics, other educational levels, and with other digital tools.

Recommendations for Educational Practice

In this section we highlight recommendations for educational practice that appeared from our findings.

Addressing Statistical Inference in Early-grade Curricula

This research presents a learning trajectory for statistical inference in descriptive-oriented pre-grade 10 curricula. The findings suggest that current statistics curricula for grades 6–9, usually with a strong descriptive focus, can be enriched with an inferential focus—at least for preparatory university

education (VWO). The benefit will be that students learn more about inference, and not less about the other domains of statistical literacy, to anticipate subsequent steps in students' statistics education. Introducing (informal) statistical inference in these early years of secondary school seems feasible, effective, and beneficial for students' follow-up statistics education. However, implementing innovations within educational curricula is complex. A combined approach of top-down and bottom-up seems most effective (Fullan, 1994). More research is needed into how innovations can be successfully addressed in current statistics curricula with a descriptive focus.

Broadening of the Black Box Paradigm

The black box activities, combined with statistical modeling, proved engaging and promoted students to achieve the intended learning goals. We assume that these activities can also be applicable for younger students or students in other streams of secondary education—that is, not pre-university level. Furthermore, the ideas of the black box and modeling also seem applicable to more complex statistical concepts, such as comparing groups or hypotheses testing—which are difficult for many students (Stalvey et al., 2019). For example, providing a physical black box filled with marbles and having students test whether the given ratio is likely to be true, can be an informal approach to hypotheses testing. We recommend teachers and educators involved in the design of teaching materials for introducing statistical inference to consider these ideas.

Preparing Mathematics Teachers for Innovations in Statistics Education

To successfully implement learning trajectories for statistics, we recommend to carefully prepare participating mathematics teachers. In many countries, including the Netherlands, secondary statistics education is part of the mathematics curriculum. The differing nature of statistics—more contextual and less deterministic—makes it less popular among many mathematics teachers. In addition, many mathematics teachers in the early years of secondary education are inexperienced and not trained to teach inferential statistics. On top of that, most mathematics teachers are not used to work with technology in class, and they are not accustomed to an inquiry-based teaching approach that differs from the often instruction-based regular lessons.

Using Technology in Statistics Education

Technology is indispensable for doing and learning statistics. However, many mathematics teachers are insufficiently trained to teach statistics by digital means. For the participating teachers in our research, learning how to use a new digital tool themselves, as well as learning how to teach with a digital tool and

how to teach students to use a digital tool, appeared challenging. In daily educational practice, teachers typically lack time and opportunities to develop their proficiency for teaching and learning from and with digital tools. More research is needed into how teachers can be adequately prepared for using technology in statistics education.

In many schools, teaching with digital tools is also limited due to practical issues. During the implementation of our learning trajectory, we were confronted with several practical problems, such as computer shortage, difficulties with scheduling in computer rooms, problems with installing new software on a school network and poor internet facilities. More research is needed into how schools can be sufficiently facilitated in both materials and knowledge for effective deployment of technology, especially within statistics education.

The use of technology in education increased tremendously in the past year, due to the COVID pandemic. The abrupt school closure in many countries, including the Netherlands, resulted in a disorderly explosion of using all kinds of digital learning environments. The vast body of experiences gained provided a new impulse to teaching and learning with technology. These developments call for research into sustainable educational innovations in which the use of technology can be integrated into the regular educational system.

Personal Reflection as a Teacher-researcher

Starting this research project involved transitioning from a familiar educational world into an unfamiliar scientific world. Combining both worlds, in the role of a teacher-researcher, is identified by Bakx et al. (2016) as *boundary crossing*. In this regard, boundaries encompass socio-cultural differences, which lead to discontinuity in action or interaction (Akkerman & Bakker, 2011). Boundary crossing is defined by Bakker and Akkerman (2014) as efforts made by individuals or groups at boundaries to establish or restore continuity in action or interaction across practices. As a beginning researcher, the assimilation into the scientific culture, the novelty of academic knowledge and skills, and the unfamiliarity with fellow researchers, were challenging aspects. Balancing time and flexible switching between the two worlds remained a concern throughout the project. When implementing and coordinating the intervention(s), both worlds—and with that both roles—intersect. On the one hand, it was challenging to observe and analyze intervention data as a researcher, and not as a teacher. On the other hand, teacher experiences were beneficial in designing

the intervention, organizing it practically, and guiding participating teachers. Research by Bakx et al. (2016) indicates that more teacher-researchers recognize the challenges described in boundary crossing. However, they did not mention the ambiguity in roles that might occur when both worlds intersect, as is the case with intervention studies—research in educational practice.

This research project made a rich contribution to my professional development as a teacher at a micro, meso and macro level (Akkerman & Bruining, 2016). At the micro level in my own teaching practice, this research project provided insight into students' learning processes and how to promote these. For example, the designed learning trajectory was implemented in my classes and knowledge gained was also integrated into the teaching of other mathematics topics. At the meso level as a teacher in the school, this research provided an advanced analytical view on the school as educational system. For example, this research provided insight into integrating (inter)national educational theories, materials and approaches at the school level, and also insight into the coherence between groups within the school and the educational system, with varying goals and perspectives—for example teachers, students, authors of textbooks and educational designers. At the macro level of the (regional and national) mathematics education community, the research findings were disseminated to mathematics teachers by arranging workshops, and by publishing findings in journals for mathematics teachers. As a result, several teachers implemented the designed learning trajectory in their classrooms, in a variety of educational levels and grades—for example in Grades 10–12 and higher education. The informal exchange of their experiences was a valuable continuation and addition to this research project.

As a researcher, this project enabled me to develop and increase my competencies and passion for conducting research. Functioning within the scientific community deepened my perspective on teaching and research. Working with experts at the Freudenthal Institute was a unique learning experience. Also, collaboration with international colleagues broadened my view on education in many ways. In summary, this research project strengthened my professional development in a broad scope—as a professional in the classroom, within the school, and within the (inter)national world of education and research.

References

- Ahuvia, A. (2001) Traditional, interpretive, and reception based content analyses: Improving the ability of content analysis to address issues of pragmatic and theoretical concern, *Social Indicators Research*, 54(2), 139–172.
- Ainley, J., Pratt, D., & Hansen, A. (2006). Connecting engagement and focus in pedagogic task design. *British Educational Research Journal*, 32(1), 23–38.
- Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of educational research*, 81(2), 132–169.
- Akkerman, S., & Bruining, T. (2016). Multi-level boundary crossing in a professional development school partnership. *Journal of the Learning Sciences*, 25(2), 240–284.
- Arnold, P., Confrey, J., Jones, R. S., Lee, H. S., & Pfannkuch, M. (2018). Statistics learning trajectories. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 295–326). Cham, Switzerland: Springer.
- Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, 7(3), 245–274.
- Bakker, A. (2004). *Design research in statistics education. On symbolizing and computer tools*. Utrecht, the Netherlands. Utrecht University.
- Bakker, A. (2007a). Diagrammatic reasoning and hypostatic abstraction in statistics education. *Semiotica*, 164, 9–29.
- Bakker, A. (2007b). Diagrammatisch redeneren als basis voor begripsontwikkeling in het statistiekonderwijs [Diagrammatic reasoning as a basis for concept development in statistics education]. *Pedagogische Studiën*, 84(5), 340–357.
- Bakker, A. (2018). *Design research in education. A practical guide for early career researchers*. London, UK: Routledge.
- Bakker, A., & Akkerman, S. F. (2014). A boundary-crossing approach to support students' integration of statistical and work-related knowledge. *Educational Studies in Mathematics*, 86(2), 223–237.
- 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.
- Bakx, A., Bakker, A., Koopman, M., & Beijgaard, D. (2016). Boundary crossing by science teacher researchers in a PhD program. *Teaching and Teacher Education*, 60, 76–87.
- Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts.

- International Journal of Mathematics Education in Science and Technology*, 25(4), 527–547.
- Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. [CD-ROM]. Voorburg, the Netherlands: International Statistical Institute.
- Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM—The International Journal on Mathematics Education*, 44(7), 913–925.
- Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics*, 88(3), 291–303.
- 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). New York, NY: Springer.
- Biehler, R., Ben-Zvi, D., Bakker, A., & Maker, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third international handbook of mathematics education* (pp. 643–690). Cham, Switzerland: Springer.
- Biehler, R., Frischemeier, & D., Podworny, S. (2017). Editorial: Reasoning about models and modeling in the context of informal statistical inference. *Statistics Education Research Journal*, 16(2), 8–12.
- Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning about data. In D. Ben-Zvi, J. Garfield, & K. Makar (Eds.), *International handbook of research in statistics education* (pp. 139–192). New York, NY: Springer.
- Bikner-Ahsbahs, A., & Prediger, S. (eds) (2014). Networking of theories as a research practice in mathematics education. *Advances in Mathematical Education*. Cham, Switzerland: Springer.
- Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. 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. 57–69). Dordrecht, the Netherlands: Springer.
- Büscher, C., & Schnell, S. (2017). Students' emergent modeling of statistical measures—a case study. *Statistics Education Research Journal*, 16(2), 144–162.
- Callingham, R., & Watson, J. M. (2017). The development of statistical literacy at school. *Statistics Education Research Journal*, 17(1), 181–201.

- Castro Sotos, A. E., Vanhoof, S., van Den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 1(2), 90–112.
- Cave, D. (2020, April 24). Vanquish the virus? Australia and New Zealand aim to show the way. *New York Times*. Retrieved from <https://www.nytimes.com/2020/04/24/world/australia/new-zealandcoronavirus.html?referringSource=articleShare>
- 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 Journal*, 1(1), 1–23.
- Chance, B., delMas, 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). Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, 6(2), 81–89.
- Cobb, P. (1994). Constructivism in mathematics and science education. *Educational Researcher*, 23(7), 4–4.
- Cobb P. (2011). Learning from distributed theories of intelligence. In E. Yackel, K. Gravemeijer & A. Sfard (Eds.), *A journey into mathematics education research: Insights from the work of Paul Cobb* (pp. 85–105). New York, NY: Springer.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.
- Davies, R. (2011). Understanding technology literacy: A framework for evaluating educational technology integration. *TechTrends*, 55(5), 45–52.
- De Corte, E. (2000). Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction*, 10(3), 249–266.
- delMas, R.C., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6, 28–58.
- Doerr, H., delMas, R., Makar, K. (2017). A modeling approach to the development of students' informal inferential reasoning, *Statistics Education Research Journal*, 16(2), 86–115.
- Drijvers, P., Godino, J. D., Font, D., & Trouche, L. (2013). One episode, two lenses. A reflective analysis of student learning with computer algebra from instrumental and onto-semiotic perspectives. *Educational Studies in Mathematics*, 82(1), 23–49.

- Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, 47(2), 123–182.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Euler, D. (2017). Design principles as a bridge between scientific knowledge production and practice design. *EDeR. Educational Design Research*, 1(1), 1–15.
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Schaeffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. Alexandria, VA: American Statistical Association.
- Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht, the Netherlands: Reidel.
- Fullan, M. G. (1994). Coordinating top-down and bottom-up strategies for educational reform. In R. F. Elmore & S. H. Fuhrman (Eds.), *The governance of curriculum* (pp. 186–202). Alexandria, VA: Association for Supervision and Curriculum Development.
- Gal, I. (2002). Adults' statistical literacy: Meaning, components, responsibilities. *International Statistical Review*, 70(1), 1–25.
- Garfield, J., Ben-Zvi, D., Le, L., & Zieffler, A. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, 88(3), 327–342.
- Garfield, J., delMas, R., & Chance, B. (1999). *Developing statistical reasoning about sampling distributions*. Presented at the First International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL), Kibbutz Be'eri, Israel.
- Garfield, J., delMas, R., & Chance, B. (2002). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project. NSF CCLI grant ASA-0206571. <https://app.gen.umn.edu/artist/>
- Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM—The International Journal on Mathematics Education*, 44(7), 883–898.
- Gravemeijer, K. (1999). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning*, 1, 155–177.
- Gravemeijer, K., Bowers, J., & Stephan, M. (2003). A hypothetical learning trajectory on measurement and flexible arithmetic. In M. Stephan, J. Bowers, P. Cobb, & K. Gravemeijer (Eds.), *Supporting students' development of measuring conceptions: Analyzing students' learning in social context* (pp. 51–66). Reston, VA: NCTM.

- Gravemeijer, K. (2004). Learning trajectories and local instruction theories as means of support for teachers in reform mathematics education. *Mathematical Thinking and Learning*, 6(2), 105–128.
- Gray, E.M., & Tall, D.O. (1994). Duality, ambiguity, and flexibility: A “proceptual” view of simple arithmetic. *Journal for Research in Mathematics Education*, 25(2), 116–140.
- Groth, R. E. (2015). Working at the boundaries of mathematics education and statistics education communities of practice. *Journal for Research in Mathematics Education*, 46(1), 4–16.
- Innabi, H., & El Sheikh, O. (2007). The change in mathematics teachers' perceptions of critical thinking after 15 years of educational reform in Jordan. *Educational Studies in Mathematics*, 64(1), 45–68.
- Kaufman, D.M. (2003). Applying educational theory in practice. *British Medical Journal*, 326, 213–216.
- Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, 12(3), 217–230.
- Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, 2(1), 1–37.
- 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.
- Lehrer, R., & English, L. D. (2017). Introducing children to modeling variability. In Ben-Zvi, D., Garfield, J., & Makar, K. (Eds.). *International handbook of research in statistics education* (pp. 229–260). Dordrecht, the Netherlands: Springer.
- Lewin, K. (1952). *Field theory in social science: Selected theoretical papers by Kurt Lewin*. London, UK: Tavistock.
- Maass, K., Cobb, P., Krainer, K., & Potari, D. (2019). Different ways to implement innovative teaching approaches at scale. *Educational Studies in Mathematics*, 102(3), 303–318.
- Makar, K. (2016). Developing young children’s emergent inferential practices in statistics. *Mathematical Thinking and Learning*, 18(1), 1–24.
- Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, 13(1–2), 152–173.
- Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82–105.
- Makar, K., & Rubin, A. (2018). Learning about statistical inference. In D. Ben-Zvi, K. Makar, J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 261–294). Cham, Switzerland: Springer.
- Manor, H., & Ben-Zvi, D. (2015). Students’ articulations of uncertainty in informally exploring sampling distributions. In A. Zieffler, & E. Fry (Eds.),

- Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 57–94). Minneapolis, MN: Catalyst Press.
- Manor, H., & Ben-Zvi, D. (2017). Students' emergent articulations of statistical models and modeling in making informal statistical inferences. *Statistics Education Research Journal*, 16(2), 116–143.
- McClain, K., McGatha, M., & Hodge, L. (2000). Improving data analysis through discourse. *Mathematics Teaching in the Middle School*, 5(8), 548–553.
- McKenney, S., & Reeves, T.C. (2012). *Conducting educational design research*. London, UK / New York, NY: Routledge.
- Meletiou-Mavrotheris, M., & Paparistodemou, E. (2015). Developing young learners' reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics*, 88(3), 385–404.
- NCTM, (n.d.). Content standards of the US National Council of Teachers of Mathematics. Retrieved from <https://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Data-Analysis-and-Probability/>
- Novak, E. (2014). Effects of simulation-based learning on students' statistical factual, conceptual, and application knowledge. *Journal of Computer Assisted Learning*, 30(2), 148–158.
- Olsen, D. R. (2004). The triumph of hope over experience in the search for “what works”: A response to Slavin. *Educational Researcher*, 33(1), 24–26.
- Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, 7(2), 83–106.
- Patel, A., & Pfannkuch, M. (2018). Developing a statistical modeling framework to characterize Year 7 students' reasoning. *ZDM Mathematics Education*, 50(7), 1197–1212.
- Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, 13(1–2), 27–46.
- Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modelling to connect data, chance and context. *ZDM Mathematics Education*, 50(7), 1113–1123.
- Piaget, J. (1978). *Success and understanding*. Cambridge, MA: Harvard University Press.
- Podworny, S., & Biehler, R. (2014). A learning trajectory on hypothesis testing with TinkerPlots – Design and exploratory evaluation. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education (Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, USA)*. Voorburg, the Netherlands: International Association for Statistical Education and the International Statistical Institute.

- Prediger, S., Bikner-Ahsbahs, A., & Arzarello, F. (2008). Networking strategies and methods for connecting theoretical approaches: First steps towards a conceptual framework. *ZDM—The International Journal on Mathematics Education*, 40(2), 165–178.
- Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, 7(2), 5–19.
- Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, 10(3).
- Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, 51(3), 257–270.
- Sandoval, W. A. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences*, 23(1), 18–36.
- Savelsbergh, E., Prins, G., Rietbergen, C., Fechner, S., Vaessen, B., Draijer, J., & Bakker, A. (2016). Effects of innovative science and mathematics teaching on student attitudes and achievement: A meta-analytic study. *Educational Research Review*, 19, 158–172.
- Schäfer, T., Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10(813), 1–13.
- Schild, M. (2004). Statistical literacy curriculum design. *IASE Curriculum Design Roundtable*. <https://www.StatLit.org/pdf/2004SchildIASE.pdf>
- Schild, M. (2010). Assessing statistical literacy: take care. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: an international perspective* (pp. 133–152). Chichester, UK: John Wiley & Sons.
- Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics*, 22(1), 1–36.
- Sfard, A. & Lavie, I. (2005). Why cannot children see as the same what grown-ups cannot see as different? Early numerical thinking revisited. *Cognition and Instruction*, 23(2), 237–309.
- Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, 32(4), 450–466.
- Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26(2), 114–145.
- Simon, M. A. (2019). Analyzing qualitative data in mathematics education. In K.R. Leatham (Ed.), *Designing, Conducting, and Publishing Quality Research in Mathematics Education* (pp. 111–123). Cham, Switzerland: Springer.
- 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.

- Stalvey, H. E., Burns-Childers, A., Chamberlain, D., Kemp, A., Meadows, L. J., & Vidakovic, D. (2019). Students' understanding of the concepts involved in one-sample hypothesis testing. *The Journal of Mathematical Behavior*, 53, 42–64.
- Streefland L (1991) *Fractions in realistic mathematics education. A paradigm of developmental research*. Dordrecht, the Netherlands: Kluwer
- Tessmer, M. (1993). *Planning and conducting formative evaluation*. London, UK: Kogan Page.
- Taber, K.S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48, 1273–1296.
- Thijs, A., Fisser, P., & Van der Hoeven, M. (2014). *21e-eeuwse vaardigheden in het curriculum van het funderend onderwijs* [21st century skills in the curriculum of foundational education]. Enschede, the Netherlands: SLO.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110.
- Van Dijke-Droogers, M.J.S., Drijvers, P.H.M., & Bakker, A. (2018). Repeated sampling as a step towards informal statistical inference. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018)*, Kyoto, Japan. Voorburg, the Netherlands: International Statistical Institute.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2020). Repeated sampling with a black box to make informal statistical inference accessible. *Mathematical Thinking and Learning*, 22(2), 116–138.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2021). Statistical modeling processes through the lens of instrumental genesis. *Educational Studies in Mathematics*. <https://doi.org/10.1007/s10649-020-10023-y>
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted). Introducing Statistical Inference: Design of a Theoretically and Empirically Based Learning Trajectory.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted). Effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy.
- Van Dijke-Droogers, M., Drijvers, P., & Tolboom, J. (2017). Enhancing statistical literacy. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10, February 1–5, 2017)* (pp. 860–867). DCU Institute of Education and ERME.
- Van Someren, M. W., Barnard, R., & Sandberg, J. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London, UK: Academic Press.
- Van Streun, A. & Van de Giessen, C. (2007). Een vernieuwd statistiekprogramma: Deel 1 [A renewed statistical program, Part 1]. *Euclides*, 82(5), 176–179.

- Vergnaud, G. (1996). Au fond de l'apprentissage, la conceptualisation. In R. Noirfalise & M.-J. Perrin (Eds.), *Actes de l'école d'été de didactique des mathématiques* (pp. 174–185). Clermont-Ferrand, France: IREM.
- Vygotsky, L. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.
- Watson, J. (1997). Assessing statistical literacy through the use of media surveys. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). International Statistical Institute IOS Press.
- Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Watson, J.M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2, 3–46.
- Watson, J., & Callingham, R. (2004). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education: International Association for Statistical Education roundtable* (pp. 116–137). International Association for Statistical Education.
- Watson, J., & Callingham, R. (2020). COVID-19 and the need for statistical literacy. *Australian Mathematics Education Journal*, 2(2), 20–25.
- Watson, J., & Chance, B. (2012). Building intuitions about statistical inference based on resampling. *Australian Senior Mathematics Journal*, 26(1), 6–18.
- Watson, J. M., & Kelly, B. A. (2008). Sample, random and variation: The vocabulary of statistical literacy. *International Journal of Science and Mathematics Education*, 6(4), 741–767.
- Whitaker, D., Foti, S., & Jacobbe, T. (2015). The levels of conceptual understanding in statistics (LOCUS) project: Results of the pilot study. *Numeracy*, 8(2). <http://dx.doi.org/10.5038/1936-4660.8.2.3>
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–265.
- Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 247–295.
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716–730.
- Ziegler, L. (2014). Reconceptualizing statistical literacy: Developing an assessment for the modern introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics: Sustainability in Statistics Education*. International Association for Statistical Education.
- Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 40–58.

Supplementary Materials

Supplementary Material A: Overview of the Eight-step LT

Step 1 starts with a physical black box experiment. Students examine the content of a black box filled with 1,000 marbles in the colors yellow and orange. By counting the number of yellow marbles in a viewing window with 20 visible marbles—and later using a larger window with 40 visible marbles—they examine how many yellow marbles there are in the whole box. By repeatedly shaking the black box and counting the visible yellow marbles, students become aware of sampling variability.

In this physical experiment of step 1, students collect sample data to estimate the population—here the content of the black box. By taking repeated measurements, i.e., samples, and exchanging them within the classroom, they are confronted with sampling variability. Students experience that an estimate can be made based on a sample, but that the content cannot be determined with absolute certainty. By repeating this experiment using a larger viewing window, students experience that the corresponding estimates vary less and provide a better picture of the population. The hypothesis for step 1 is that students get an idea of the concept of a sample with associated uncertainty. This activity incorporates theories of repeated and growing samples (Bakker, 2004; Saldanha & Thompson, 2002; Wild & Pfannkuch, 1999) and informal statistical inference (Makar & Rubin, 2009), combined with design principles of Realistic Mathematics Education (Freudenthal, 1983) and ideas of using meaningful contexts (Ainly et al., 2006). From step 1, in which students experience the variability and uncertainty of samples and the added value of using repeated and larger samples, raises the question of what happens when we further increase the size and number of repeated samples. Students experienced in this step that conducting more repetitions and using larger sample sizes requires more time and effort. As a follow-up, in step 2 they use a thought experiment to explore possible sample results for a large number of repetitions.

In step 2, students make a sketch of the sample results they expect when the black box experiment of step 1 is repeated many times. They sketch the expected frequency distribution for 100,000 repeated samples of size 40 from a black box filled with 750 yellow and 250 orange marbles. Sketched distributions are exchanged and discussed in classroom. As a follow-up within step 2, students determine the probability of a certain range of sample results from a given distribution for 1,500 repeated samples.

In sketching the frequency distribution in step 2, students visualize the sampling variability they expect for a large number of repetitions, based on their experiences with the black box in step 1. Furthermore, students use their experience in sketching the frequency distribution from repeated samples, for interpreting a given distribution to determine the probability of a certain range of sample results. The hypothesis for step 2 is that students get to understand the concept of frequency distribution for repeated samples by sketching one and, subsequently, that they understand that the distribution facilitates them to determine the probability of a certain range of sample results. In this step, theories on making predictions—or using “What if” questions—and reasoning with the frequency distribution from repeated sampling (Rossman, 2008; Watson & Chance, 2012) are incorporated. From step 2 emerges the question of how to get a distribution of repeated samples to determine the probability of certain sample results, in a quick and easy way. Therefore, in step 3, students are introduced to the digital environment of TinkerPlots.

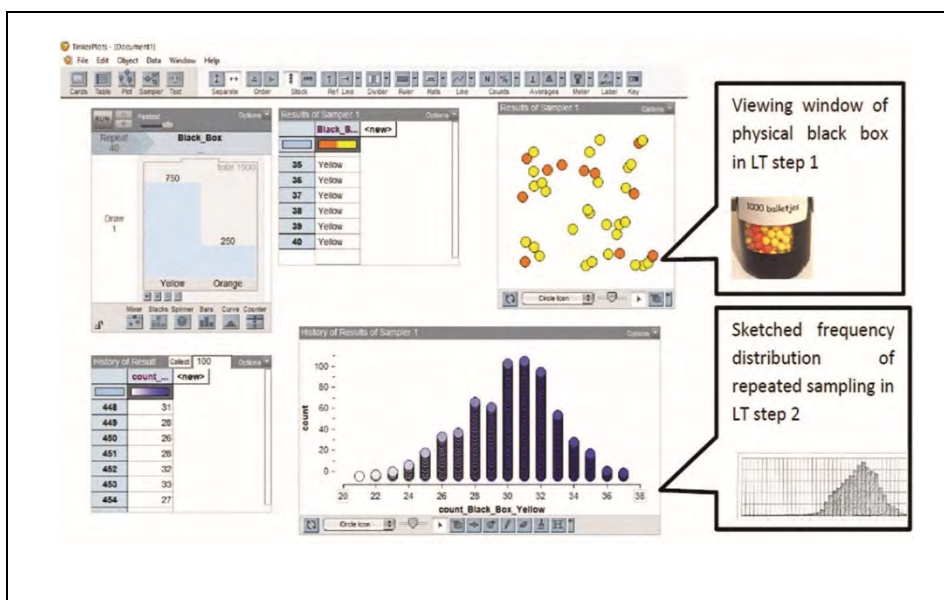


Figure 1. Similarity between the digital environment of TinkerPlots in LT step 3 and the black box experiment in LT steps 1 and 2

In step 3, students use statistical modeling within the digital environment of TinkerPlots to examine the probability of certain sample results by interpreting the simulated sampling distribution of repeated samples, within the context of a black box. Statistical modeling includes building a model (i.e., of a black box filled with marbles), simulating repeated samples, visualizing the sampling

distribution and interpreting the results. Subsequently, students experiment with varying sample sizes and number of repeated samples to investigate the effects of size and repetitions on the estimate of the population.

The digital environment of TinkerPlots has strong similarities with students' experiences from step 1—the viewing window with marbles—and step 2—their sketched visualizations of the frequency distribution of repeated samples (see Figure 1). The hypothesis in step 3 is that students recognize the digital environment of TinkerPlots from their physical black box experiments and visualizations, and that these experiences facilitate them to deploy statistical modeling and to interpret the simulated sampling distribution. According to Chance et al. (2004), ways to improve the understanding of sampling distributions include exploring samples, comparing how sample behavior mimics population behavior, and conducting both structured and unstructured explorations with the digital tool. Step 3 constitutes on theories about working with a computer model for simulations that has a strong connection with a concrete experiment (Chance et al., 2007; Konold & Kazak, 2008; Manor & Ben-Zvi, 2015)—here the black box experiment— and about working with simulations of many repeated samples to determine whether a sample result is likely (Garfield, et al., 2015; Manor & Ben-Zvi, 2015; Watson & Chance, 2012). These theories are combined with ideas about experimenting with various sample results from a given population to explore the effects of sample size and number of repetitions on sampling variability (Wild et al., 2011), and accordingly, on the probability of the inference about the population. From steps 1–3, students get to understand that statistical modeling can be helpful to determine the probability of certain sample results, within the context of the black box. However, statistical modeling with a black box is context-specific and therefore, in step 4 students examine how statistical modeling can be used more generally in other situations and contexts.

In step 4, students use their statistical modeling experiences from the black box context in new real-life situations and contexts. As such, students use statistical modeling with TinkerPlots to solve a given problem. For example, within the context of social media and in particular the use of WhatsApp, students investigate whether the use of WhatsApp within their class deviates from the national standard—according to research by Newcom: 90% of Dutch students aged 15-19 years uses WhatsApp on a daily basis. By collecting data from each student in class, for example 21 out of 25 students use WhatsApp on a daily basis, students investigate whether a sample result of 21 in a sample of 25 from a population proportion of 90% is (un)likely.

The application in several new contexts, using the same digital techniques for statistical modeling in TinkerPlots, enables students to identify general patterns and to develop a context-independent use of statistical modeling, known as emergent modeling (Gravemeijer, 2008). Using statistical modeling for solving real-life problems includes the process of abstracting the real world into a model and then using this model for understanding the real world. Patel and Pfannkuch (2018) elaborated this relationship between the real and model world in a framework that displays students' cognitive activities about understanding the problem (real world), seeing and applying structure (real world–model world), modeling (model world–real world), analyzing simulated data (model world) and communicating findings (model world–real world). In this regard, Manor and Ben-Zvi (2017) identified several dimensions: reasoning with phenomenon simplification, with sample representativeness, and with sampling distribution. These theories on statistical modeling were elaborated in step 4, where students build and run a model of a real-world situation in the model world of TinkerPlots and use this model—by simulating and interpreting the sampling distribution of repeated samples—to understand the real world situation.

In steps 1 to 4, students are introduced to the key concepts of statistical inference: sample, sampling variability, sample size, repeated sampling, frequency and sampling distributions, probability and uncertainty. During these four LT steps, students only use categorical data. From these steps emerges the question of how to use statistical modeling with other data. Therefore, in steps 5 to 8, students go through similar learning steps to steps 1 to 4, but now using numerical data. The hypothesis is that this iterative approach facilitates students to anchor, expand and deepen their understanding of the key concepts. Step 5 to 8 mainly constitute on theories mentioned in step 1 to 4; in the following we focus on the new elements.

In step 5, as in step 1, students conduct a physical experiment, but this time using a black box filled with 4,000 notes. Each note contains information on gender and height for a 14-years-old Dutch student, for example: boy – 155 cm. In couples of two, students randomly draw a sample of 40 from the black box. They summarize their sample data by calculating measures of center and spread, and visualizing their findings. The sample results are exchanged and discussed within the classroom, focusing on sampling variability and drawing inferences about the population.

Students compare and discuss the sample results found—both the visualizations and measures of center and spread—focusing on similarities and differences between samples and on what these varying sample data say about the population. Finding similarities between samples with numerical data, in particular for a small sample size, is more difficult than for the categorical data in steps 1 to 4—for example, the comparison of categorical data in step 1 only considered the number of yellow marbles. Drawing inferences about the population based on several samples requires merging the information found. Students experience that the use of a sample characteristic—such as the sample mean—is helpful to merge information found in repeated samples with numerical data. The hypothesis in step 5 is that students understand that a sample characteristic, for example the sample mean, combined with the sample distribution, can be used to obtain a picture of the population distribution. From step 5, in which students discussed how to use numerical data from repeated samples to draw inferences about the population, raises the question of how the population distribution at stake—the content of the black box filled with 4,000 notes on students’ gender and height—can be pictured based on the sample results found. As a next step, students are asked to visualize the population distribution they expect based on the numerical data from the samples found.

In step 6, students draw a sketch of the population distribution—that is, the height of the 4,000 students in the black box—they expect based on the exchanged and discussed sample results from step 5. The hypothesis in step 6 is that students use the sample mean and distributions found in step 5 to visualize the expected population distribution. During a whole class discussion, the expected population distributions are exchanged and discussed, and also compared with the real population distribution. From steps 1 to 4, students explored through statistical modeling that for categorical data, using larger sample sizes and more repetitions lead to better estimates of the population. From step 5 and 6 emerges the need for better estimates when working with numerical data. As a follow-up in step 7, students use statistical modeling with numerical data to explore the effects of larger samples on the sample mean and sample distribution, and accordingly, on the probability of the inference about the population distribution.

In step 7, students use statistical modeling in TinkerPlots with a given model of the population. The population model consists of the numerical data from the 4,000 students in the black box with notes, considering gender and height. Entering the exact population model in TinkerPlots for statistical modeling is complex and time-consuming, and therefore students use a given

model. Students use repeated sampling to explore the effects of sample size on the sample mean and sample distribution. The hypothesis is that students understand that—for numerical data—larger sample sizes better reflect the population distribution. They also experience that the sample mean for larger sample sizes varies less and better resembles the population mean. From step 7 emerges the question of how to apply statistical modeling with numerical data in other contexts and situations.

In step 8, students use statistical modeling in TinkerPlots to solve a real-life problem, by working with a given or a hidden model of the population. For example, students investigate whether the time on sports per week within their class deviates from that of 4,000 Dutch students in a given population model. By collecting data from each student in class, for example the mean sporting time for 25 students is six hours a week, students investigate whether a sample mean of six hours from the given population is (un)likely. When working with a hidden population model, students are unable to see the model. By simulating and visualizing (repeated) samples they make inferences about the population mean and distribution. The hypothesis is that the iterative process of statistical modeling, with both categorical and numerical data within varying contexts, facilitates students to make the conceptual transition to emergent modeling.

Supplementary Material B: Specified SI Levels for each LT step
Table 1. Defined Levels for Statistical Inference based on Levels for Statistical Literacy (Watson & Callingham, 2003)

Specifications and examples for each LT step				
General level description	Steps 1 and 5	Steps 2 and 6	Steps 3 and 7	Steps 4 and 8
	Introduction to samples and statistical investigation cycle, not necessarily attending variation involved.	Use of visualizations with variation involved Frequency distribution (one sample) Sampling distribution (repeated samples).	Interpreting statistical information given to reason about probability; often taking into account sample size and/or number of repetitions.	Answering a real-life problem by drawing inferences on the statistical information given.
6 Critical Mathematical	Using proportional reasoning in answering a real-life problem accompanied by full and correct statistical and contextual arguments, and attending uncertainty involved.			
Critical, questioning engagement with context, using proportional reasoning, showing appreciation of the need for uncertainty	Full and correct reasoning using statistical arguments and appropriate terminology and definitions, and addressing uncertainty if needed.	Drawing or reading visualization, including all aspects of variation involved.	Using proportional reasoning, attending uncertainty in making predictions, and interpret subtle aspects of language.	

in making predictions, and interpreting subtle aspects of language, explicitly referring to statistical information given.	<p><i>For example (concerning the description of a sample): A sample is a small, representative, randomly chosen part of the population that can be used to estimate the population, when examining the whole population is impossible.</i></p> <p><i>For example: Drawing a bell curve for the expected sampling distribution from a large number of repeated samples, with a peak at the proportion and appropriate variation.</i></p> <p><i>For example: Possible sample results with a small size roughly resemble the population, these (sample) results may vary and contain local peaks but strong deviations hardly exist.</i></p> <p><i>For example: The filling weight of the jam jars in the sample is lower than required. The simulated sampling distribution shows that the measured weight occurs in less than 5% of the samples. It is possible that the large batch meets the requirement, but that is not very likely.</i></p>	
5 Critical		
Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, appreciation of variation, explicitly referring to statistical information given.	<p>Critical reasoning using some statistical arguments, terminology and definitions, and addressing uncertainty if needed.</p> <p><i>For example (concerning the description of a sample): A sample is a small part of the population that can be used to estimate the</i></p>	<p>Using critical, but simple proportional reasoning in answering a real-life problem accompanied by some statistical and contextual arguments, and attending the uncertainty involved.</p> <p><i>For example: Possible sample results with a small size roughly resemble the population, but overestimating outliers.</i></p> <p><i>For example: Drawing a bell curve for the expected sampling distribution from a large number of repeated samples, with a peak at the population</i></p>
		<p>Using critical, but simple proportional reasoning in answering a real-life problem accompanied by some contextual arguments, and attending the uncertainty involved.</p> <p><i>For example: The filling weight of the jam jars in the sample is lower than required. The simulated sampling distribution shows that this weight hardly</i></p>

occurs. The large batch will most likely not meet the requirement.

proportion, but overestimating (or neglecting) the number of outliers.

population, when examining the whole population is impossible.

<p>4 Consistent Non-critical</p> <p>Appropriate but non-critical engagement with context, multiple aspects of terminology usage, and statistical skills associated with simple probabilities, and graph characteristics, referring to statistical information given.</p>	<p>Appropriate reasoning. Using correct but simple or incomplete aspects of statistical arguments, referring to statistical information given.</p> <p><i>For example (concerning the description of a sample): A sample is a small part of the population that can be used to estimate the population.</i></p>	<p>Drawing or reading a visualization, including simple aspects of variation, referring to statistical information given.</p> <p><i>For example: Drawing the expected sampling distribution from a large number of repeated samples with a peak at the population proportion, but also other local peaks.</i></p>	<p>Appropriate reasoning. Using simple aspects of probability related to the statistical information given, but focusing on context; often referring to personal beliefs.</p> <p><i>For example: Possible sample results with a small size resemble the population most of the time, but in the given context (I think) it is probably less.</i></p>	<p>Appropriate reasoning in answering a real-life problem accompanied by simple statistical and probabilistic arguments, but focusing on contextual arguments, referring to statistical information given.</p> <p><i>For example: The filling weight of the jam jars in the sample is lower than expected. Although the simulated sampling distribution shows varying sample results, the measured weight is too low. The large batch does not meet the requirements.</i></p>
---	--	---	--	---

3 Inconsistent			
Selective engagement with context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas, not always explicitly referring to statistical information given.	Appropriate but incomplete reasoning, without justification, always explicitly referring to statistical information given.	Drawing or reading a visualization, without considering variation involved.	Appropriate conclusions in reasoning, but without justification, and without considering variation and probability, often focusing on context.
without justification, and qualitative rather than quantitative use of statistical ideas, not always explicitly referring to statistical information given.	<i>For example (concerning the description of a sample): A sample is a small part of the population that indicates the population, when examining the whole population is impossible.</i>	<i>For example: Drawing the expected sampling distribution from a large number of repeated samples with only one to three bars at the population proportion.</i>	<i>For example: The filling weight of the jam jars in the sample is lower than expected. This hardly occurs, so the large batch does not meet the requirements.</i>
2 Informal			
Only colloquial or informal engagement with context often reflecting intuitive non-statistical beliefs, single element of complex terminology and setting, and basic	Informal reasoning with context, often reflecting intuitive, non-statistical beliefs.	Drawing or reading a visualization, based on non-statistical beliefs; sometimes with considering variation involved.	Informal reasoning in answering a real-life problem, often reflecting non-statistical beliefs; sometimes considering variation involved.
	<i>For example (concerning the</i>	<i>For example: Drawing the expected sampling</i>	<i>For example: The filling weight of the jam</i>

one-step table and graph readings and calculations, not referring to statistical information given.	<i>description of a sample): A sample is a small group of students that can be used to examine a specific issue.</i>	<i>distribution for a large number of repetitions, with a peak not corresponding to the population proportion but other intuitive ideas.</i>	<i>rely on the statistical information given, but on personal or intuitive ideas.</i>	<i>jars in the sample is lower than expected. Perhaps the sample happens to contain mainly light-weighted jars, or the large batch does not meet the requirements.</i>
1 Idiosyncratic		Drawing or reading a visualization. Tautological reasoning with focus on the context.		Tautological reasoning in answering a real-life problem with focus on the context.
Idiosyncratic engagement with context, tautological use of terminology	<i>For example (concerning the description of a sample): A sample is something you measure, such as weight or height.</i>	<i>For example: Drawing the expected sampling distribution for a large number of repeated sampling as a frequency distribution from one sample and/or based on intuitive ideas.</i>	<i>For example: Possible sample results are a small part of the population</i>	<i>For example: The filling weight of the jam jars is too low.</i>
0 Incorrect reasoning	Incorrect use of terminology and context.	Random drawing or incorrect reading a visualization.	Incorrect use of terminology and context.	Incorrect inferences.

Supplementary Material C

Pre-test 'Statistical Literacy' Vwo 3 (Grade 9)

English Version

NOTE: Do not turn the page yet. Write down the information requested and read the instruction below carefully.



Name:

Group:

School:

Date of test:

This test is designed to examine the level of statistical reasoning among Vwo 3 students. The test consists of 10 open-ended questions. For the usability of these test results, it is strongly requested to **explain your answers as clearly and completely as possible**.

You will have 40 minutes to complete this test. The answers can be written down in this booklet. It is not a problem if not all questions are answered.

Questions may be completed with pen or pencil. A calculator is not required.

Wait for your teacher to indicate that you can start the test.

Item 1

To study the breakfast habits of the 600 students at a school, the researchers decide to question a part of the students. These students are asked whether they eat breakfast every day. The question a sample consisting of 30 random students.

a. What/which number(s) of students do you expect to answer the question positively?

..... students.

b. Explain your answer to 1a.

.....
.....

c. Marieke claims that a sample of 30 students is too small and that it's better to ask 100 students. Do you agree with Marieke?

.....

d. Why do you (dis)agree with Marieke?

.....
.....

e. The researchers decide to question two samples of students. The first sample, consisting of 30 students, has 20 students that eat breakfast on a daily basis. The second sample, consisting of 100 students, has 85 students that eat breakfast on a daily basis. Estimate how many students at the school eat breakfast on a daily basis.

..... students

f. Explain your answer to 1e.

.....
.....

Item 2

You toss a fair coin five times in a row and each of those five tosses results in heads.

a. What is the probability that the next toss will also result in heads?

.....

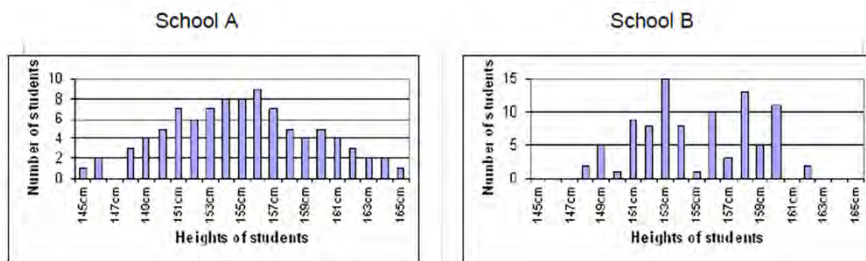
b. Explain your answer to 2a.

.....

.....

Item 3

The following graphs describe some data collected about Grade 7 students' heights in two different schools.



a. How many students are 156 cm tall in each school?

School A students and school B students.

b. Which graph shows more variability in students' heights?

Graph

c. Explain why you think this.

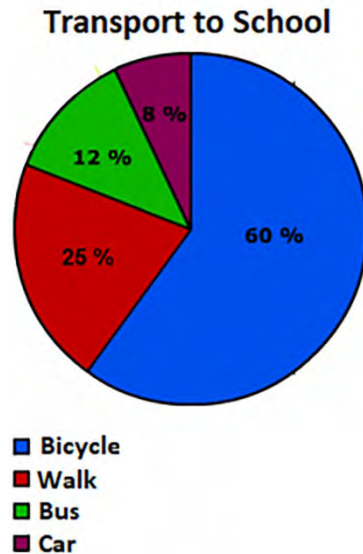
.....

.....

Item 4

This pie chart was made after a questionnaire in the exam classes.

a. What information does this chart show?



.....

.....

b. Is there something strange about this pie chart?

.....

.....

Item 5

Research centre Newcom found that 58% of Dutch adolescents between 15 and 19 years old uses Instagram on a daily basis (January 2019). A study is held with 50 of these adolescents. They are asked whether they use Instagram on a daily basis.

a. What are the results you expect from this study?

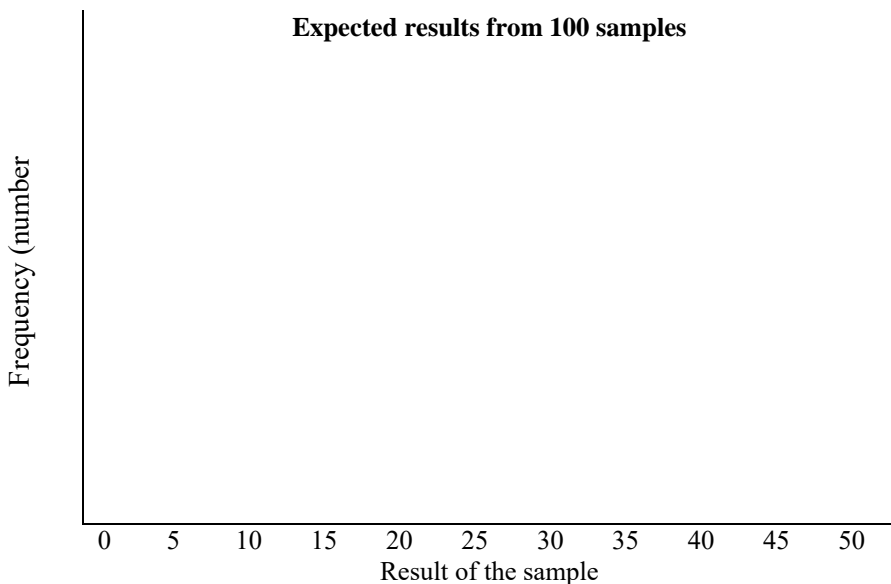
I expect that of these adolescents use Instagram on a daily basis.

b. Explain your answer to 5a.

.....

.....

c. To get a bigger picture, the study is repeated in 100 large cities. In each of these cities, the researchers ask a random group of 50 adolescents whether they use Instagram on a daily basis. Sketch a bar graph of the results you expect to receive from these samples.



d. Explain your bar graph.

.....

.....

e. At one school, it turns out that 33 out of 50 VWO-3 students use Instagram on a daily basis. Compare this result to the national results. What do you notice?

.....

.....

Item 6

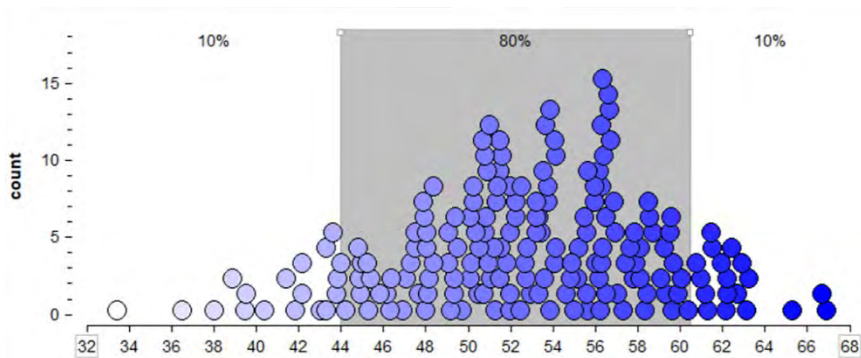
a. Explain what we mean when we talk about a probability of 2%.

.....

.....

At a middle school, 181 students are assigned to study the growth of mustard plant. Each of them receives 10 seeds and after a week they all measure the height of each of their plants in mm. Their results are shown in the graph below.

b. Explain how likely you think it is for a set of plants to have an average height below 4 cm.



Average height of 10 mustard plants after one week in mm

.....

.....

Yob, Xander and Marit missed the class and have to do the assignment later. They receive the same assignment, but each get a different type of potting soil for the seeds to grow in. Marit receives soil M, Xander soil X and Yob soil Y.

c. After a week, Marit's plants have an average height of 57 mm. Explain whether you can now conclude that mustard plants have better growth in soil M.

.....

.....

d. After a week, Xander's plants have an average height of 64 mm. Explain whether you can now conclude that mustard plants have better growth in soil X.

.....

.....

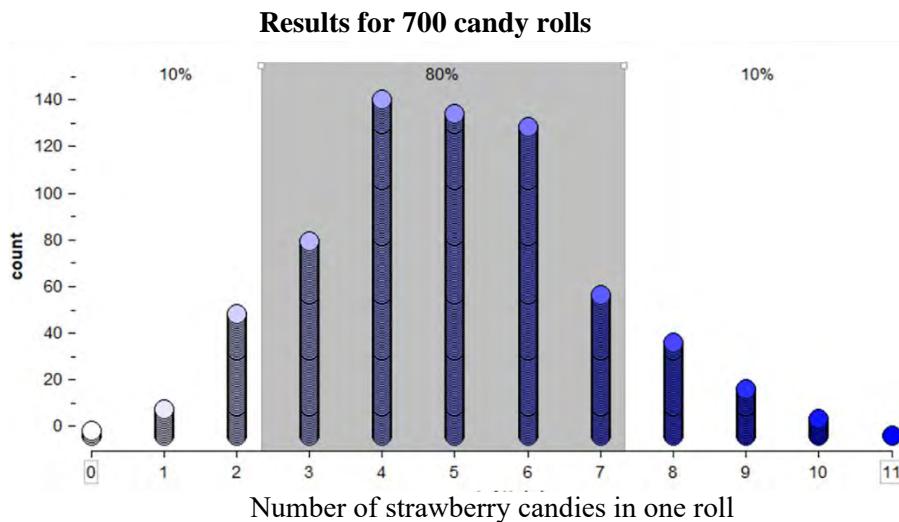
e. After a week, Yob's plants have an average height of 70 mm. Explain whether you can now conclude that mustard plants have better growth in soil Y.

.....

.....

Item 7

To analyze the number of candies with strawberry taste in a roll of 'Minitos', 700 rolls were checked. Each roll contained 20 candies. From each roll the number of candies with strawberry flavour was counted. The results of these counts are shown in the graph.



a. What was the most common result?

.....

b. Explain your answer to 7a.

.....

c. Pieter claims that he had a roll in which half the candies were strawberry-flavoured last week. Explain what you think of his claim.

.....

Item 8

Nine students in a science class weighed a small object separately on the same scales. The weights (in grams) recorded by each student are shown below.

6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3

The students had to decide on the best way to summarise these values.

a. Ben said, “I’d use the most common value to get the mode. That’s 6.3.” Is Ben’s way a good way to summarise the information? Explain your answer.

.....

b. Jane said, “I’d put them in order and use the middle value to get the median. That’s 6.2.” Is Jane’s way a good way to summarise the information? Explain your answer.

.....

c. Ron said, “I’d add them all up and divide by 9 to get the mean. That’s 7.18.” Is Ron’s way a good way to summarise the information? Explain your answer.

.....

d. May said, “I’d leave out the 15.3 and use the mean of the others. That’s 6.17.” Is May’s way a good way to summarise the information? Explain your answer.

.....

e. Which of the ways described above would you use? Why?

.....

.....

Item 9

A class wants to raise money for their school trip to Movieworld. They could raise money by selling raffle tickets for a game system. Before they decide to have a raffle, they wanted to estimate how many students in the whole school would buy a ticket. They decide to do a survey to find out first.

The school has 600 students in grades 1-6 with 100 students in each grade.

a. How many students would you survey? How would you choose them? Explain your answers.

.....

.....

b. Shannon got the names of all 600 students in the school and put them in a hat. Then she pulled out 60 names, of which 22 would want to participate. What do you think of Shannon's survey? Explain your answer.

.....

.....

c. Jake asked 10 students at an after-school computer games club, of which 5 would want to participate. What do you think of Jake's survey? Explain your answer.

.....

.....

d. Claire set up a booth at the exit of the school. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 kids to complete them, of which 37 would want to participate. What do you think of Claire's survey? Explain your answer.

.....

.....

e. How many students of the 600 students in the entire school do you think would want to participate in the raffle? You can use the results from Shannon, Jake and/or Claire. Explain your answer and which results you used.

.....

Item 10

A primary school had a sports day where every student could choose a sport to play. Here is what they chose.

	Netball	Football	Tennis	Swimming	Total
Boys	0	20	20	10	50
Girls	40	10	15	10	75

a. What was the most popular sport for boys?

.....

b. How many children were at the sports day?

..... children

c. One of the tennis players was late.

Was this player a boy or a girl. Explain your answer.

.....

End of the test

Thank you for filling it out!

Please write down your end time

Supplementary Material D

Post-test ‘Statistical Literacy’ Vwo 3 (Grade 9)



English Version

NOTE: Do not turn the page yet. Write down the requested information and carefully read the instruction below.

Name:

Group:

School:

Date of test:

This test is designed to examine the level of statistical reasoning among Vwo 3 students. The test consists of 10 open-ended tasks. For the usability of these test results, it is strongly requested to **explain your answers as clearly and completely as possible**.

You will have 40 minutes to complete this test. The answers can be written down in this booklet. It is not a problem if not all questions are answered.

Questions may be completed with pen or pencil. A calculator is not required.

Wait for your teacher to indicate that you can start the test.

Item 1

Mrs. Jones wants to buy a new car, either a Honda or Toyota. She wants whichever car will break down the least. She read in Consumer Reports that for 400 cars of each type, the Toyota had more breakdowns than the Honda. She talked to three friends. Two were Toyota owners, who had no major breakdowns. The other friend used to own a Honda, but it had lots of breakdowns, so he sold it. He said he would never buy another Honda.

Which car should Mrs. Jones buy? Explain your answer

.....,because

.....

.....

Item 2

To get the average number of children per family in a town, a teacher counted the total number of children in a town. She then divided by 50, the total number of families. The average number of children per family was 2.2.

For each of the following five statements, write down whether or not that statement is true and explain your answer.

a. Half of the families in the town have more than two children.

.....

.....

b. There are a total of 110 children in the town.

.....

.....

c. There are 2.2 children in the town for each adult.

.....

.....

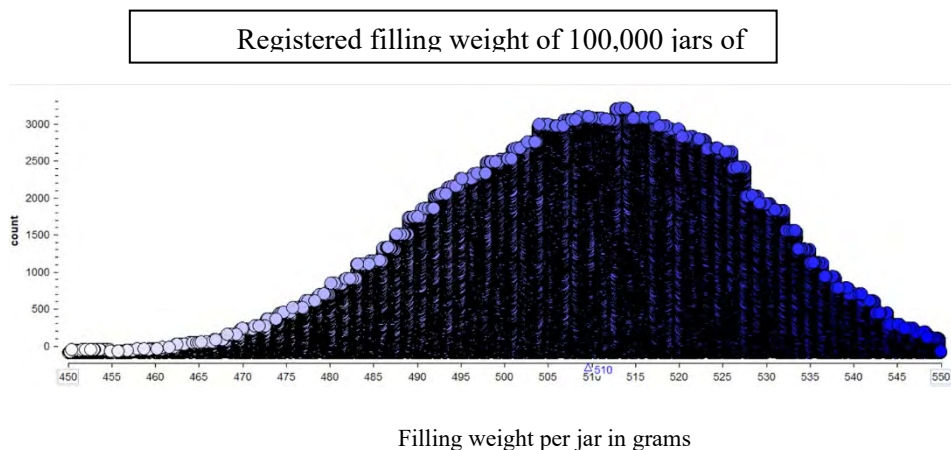
d. The most common number of children in a family is 2.

.....

.....

Item 3

According to jam manufacturer Heros, the large jars they produce contain on average 510 grams of jam. Since the filling machine cannot fill to the gram accurately, some jars contain more and others less jam. The filling weight of each jar is registered in the factory. According to the manufacturer, a printout of the filling weight of 100,000 jars looks like the chart below.

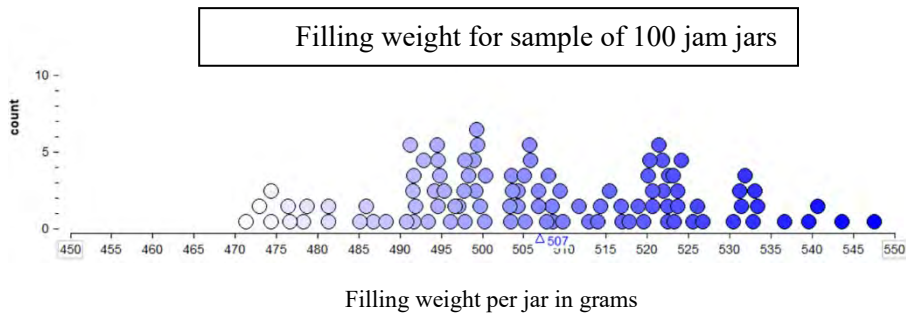


The customer who bought this batch of 100,000 jars decides to test with a sample whether the filling weight of the jars is in line with the manufacturer's registration. The customer takes a sample of 100 jars.

- a. Which sample average(s) do you expect for a sample of 100 jars?
-

- b. Explain your answer to 3a.
-
-

- c. The result of the customer's sample is visualized in the graph below.



The average in this sample is 507 grams, which is less than the promised 510 grams. Based on this sample, can the customer conclude that the manufacturer's registration is incorrect and that the manufacturer lied about the filling weight in the large batch of 100,000 jars? Explain your answer.

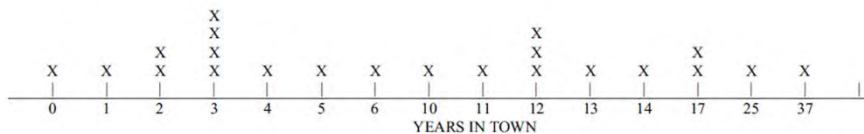
....., because

.....

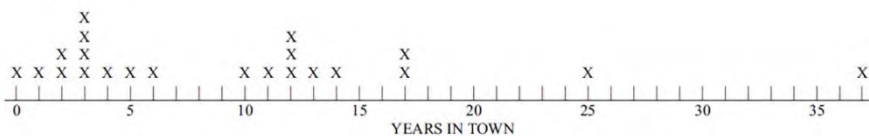
Item 4

A class of students recorded the number of years their families had lived in their town. Here are two graphs that students drew to tell the story.

Graph 1



Graph 2



a. What can you tell by looking at Graph 1?

.....

.....

b. What can you tell by looking at Graph 2?

.....
.....

c. Which Graph is better at presenting the information and “telling the story”? Explain your answer.

Graph, because
.....

Item 5

A mathematics class has 13 boys and 16 girls in it. Each student’s name is written on a piece of paper. All the names are put in a hat. The teacher picks out one name without looking.

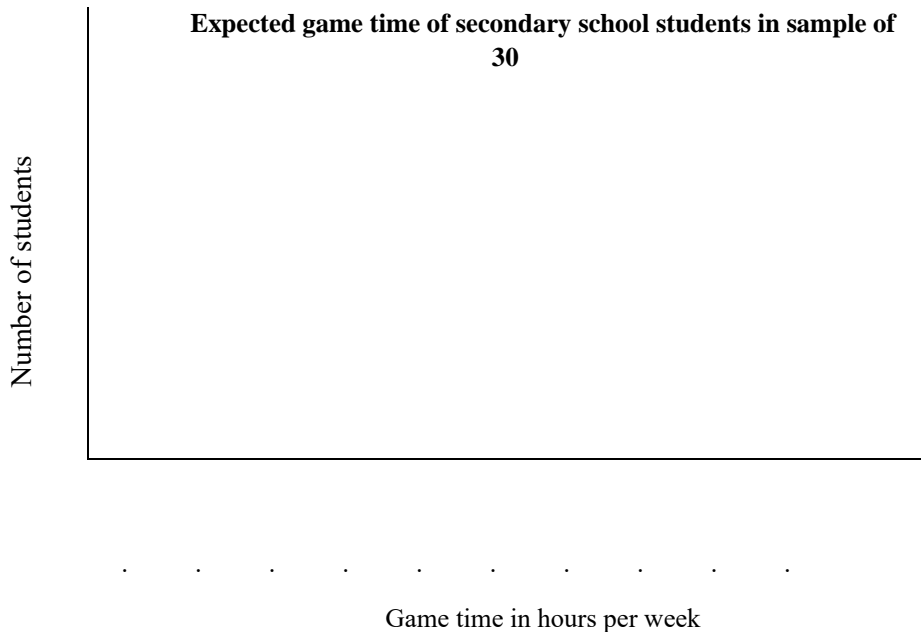
Will he pick a boy or a girl? Explain your answer.

....., because
.....
.....

Item 6

To investigate the game time of 1500 students at a secondary school, a sample is taken. The students in the sample are asked how much time in hours per week they spend on gaming. They decide to randomly question 30 students at the entrance of the school.

- a. Describe in the graph below the sample result that you expect. Choose suitable units along the horizontal axis.



- b. What average game time(s) do you expect for a sample of 30 students?

.....

.....

- c. Explain your answers to 6a and 6b.

.....

.....

- d. According to Patrick, a sample of 30 students is not enough and they have to ask at least 150 students to get a good picture of the gaming behaviour of the 1500 students.

Do you agree with Patrick? Explain your answer.

.....

.....

e. According to Mayke, there is a big difference between the game times of boys and girls. According to her, it is therefore better to examine the results of boys and girls separately. There are 10 boys in her class. Their game time in hours per week is described in the table below.

Game time per week <i>in hours</i>	7	10	14	15	17	20	35
Number of boys <i>From Mayke's class</i>	1	1	3	1	1	2	1

The result from a sample of 100 boys at the school is described in the table below.

Game time per week <i>in hours</i>	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Number of boys <i>From the sample (100)</i>	4	15	20	25	15	10	8	3

The school has 729 boys. What do you expect from the game time of the 729 boys in percentages? Describe your expectation in the table below.

Game time per week <i>in hours</i>	0-9	10-19	20-29	30-39	≥40
Percentage of boys <i>in school</i> % % % % %

f. Explain your values for the table of 6e.

.....

Item 7

The following information is from a survey about smoking and lung disease among 250 people.

	Lung disease	No lung disease	Total
Smoking	90	60	150
No smoking	60	40	100
Total	150	100	250

a. Using this information, do you think that for this sample of people lung disease depended on smoking?

.....

b. Explain your answer to 7a.

.....

Item 8

a. You throw a fair six-sided die. What result do you expect to get? Explain your answer

.....

b. You intend to throw the die until you get a 6. What is the minimum number of times you have to throw the die? And the maximum number of times?

Minimum: Maximum:

c. Explain your answers to 8b.

.....

.....

You now throw the die 60 times.

d. In the table below, fill in how many times you think each number came up.

Number on die	Times thrown
1	
2	
3	
4	
5	
6	
Total	

e. Explain why you think these numbers are reasonable.

.....

.....

Michael did the same thing with four different dice. The results can be seen in the table below.

	Times thrown			
Number on die	Die 1	Die 2	Die 3	Die 4
1	10	12	9	55
2	10	13	7	1
3	10	11	12	1
4	10	15	13	1
5	10	8	9	1
6	10	1	10	1
Total	60	60	60	60

f. Do you think these are all “fair” dice? If not, which ones do you think aren’t “fair”?

Die 1: Fair / Unfair , because

.....
.....

Die 2: Fair / Unfair , because

.....
.....

Die 3: Fair / Unfair , because

.....
.....

Die 4: Fair / Unfair , because

.....
.....

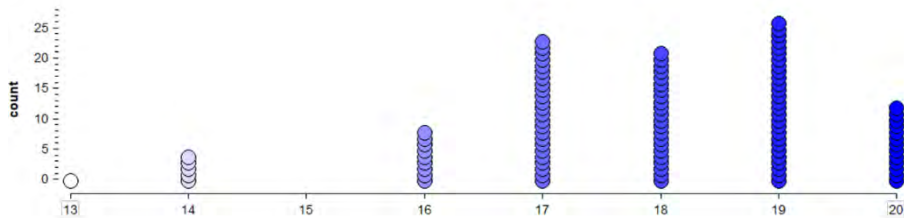
Item 9

a. Describe as clearly as possible what a sample is.

.....
.....

Do-it-yourself shop Prakkus is getting a lot of complaints about broken LED lights in the boxes of 20 lights. However, the supplier guarantees that at least 90% of the lights are in order. Prakkus decides to check the large stock of 10,000 boxes. They take a sample of 100 boxes with 20 lights each. Below you can see the number of good lights per box of 20 for a sample of 100.

Result from a sample of 100 boxes



Number of good lights per box of 20

b. What is your estimate of the probability that a random box from the large stock contains exactly 15 good lights? Explain your answer.

.....%, because
.....

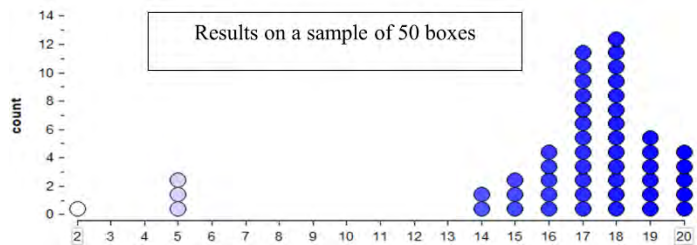
c. What is your estimate of the probability that a random box from the large stock contains less than 15 good lights? Explain your answer.

.....%, because
.....

d. Do you think that the supplier's claim is correct and that indeed 90% of the lights from the large stock are good? Explain your answer.

.....%, because
.....

e. Prakkus decides to take another sample of 50. The result is described in the graph on the right.



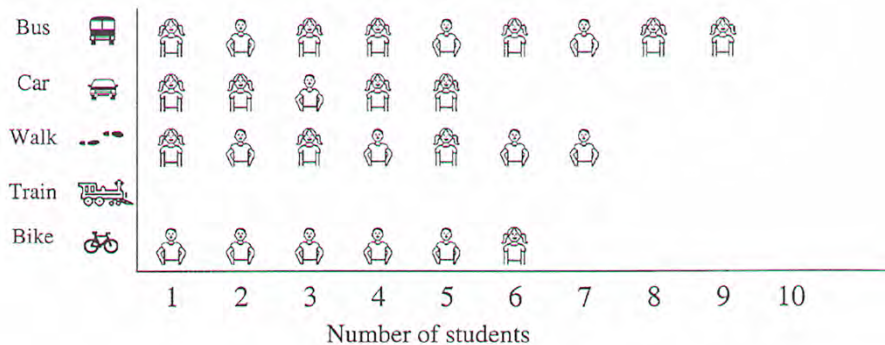
The average in this sample is 16.4, the median is 17 and the mode is 18. Which centre size (average, median or mode) gives a clear description of the sample result? Explain your answer.

.....%, because

.....

Item 10

The graph below shows how children came to school on one day.



a. How many children walked to school?

..... children

b. A new student came to school by car.

Is the new student a boy or girl? Explain your answer.

.....%, because

.....

c. Tom is not at school today.

How do you think he will come to school tomorrow? Explain your answer.

.....%, because

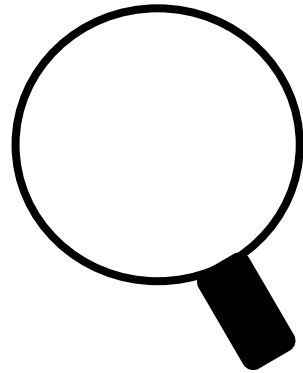
.....

End of the test

Thank you for filling it out!

Please note your end time





Samenvatting (Summary in Dutch)

De overweldigende hoeveelheid data, grafieken en voorspellingen met betrekking tot de COVID-pandemie in de media het afgelopen jaar, illustreert het essentiële belang van statistiek. De afgelopen decennia is het gebruik van data enorm toegenomen vanwege technologische vernieuwingen die het mogelijk maken om eenvoudig data te verzamelen, op te slaan, te analyseren en representeren. Op basis van data worden ingrijpende beslissingen genomen en uitspraken gedaan, zowel door burgers als professionals. Het is daarom van belang om de statistische geletterdheid van onze leerlingen te ontwikkelen. Dit houdt in dat leerlingen toegerust worden om statistische informatie te interpreteren, kritisch te beoordelen en hier conclusies uit te trekken (Gal, 2002).

Een belangrijk onderdeel van statistische geletterdheid is het werken met inferenties, oftewel met steekproeven en populaties. Bij inferenties worden op basis van steekproefdata conclusies getrokken over een groter geheel of proces. Deze conclusies gaan vergezeld van onzekerheid omdat niet alles of iedereen is onderzocht. Het interpreteren van deze onzekerheid en het duiden van de waarschijnlijkheid van de conclusie is veelzijdig en complex.

In veel landen, waaronder Nederland, wordt statistische inferentie daarom pas behandeld in de bovenbouw van het voortgezet onderwijs of in het hoger

onderwijs². Uit onderzoek blijkt dat inferentiële statistiek hier een struikelblok is voor veel leerlingen en studenten. De moeilijkheden van leerlingen worden met name veroorzaakt door een beperkt begrip van kernconcepten die nodig zijn voor inferenties (Castro Soto et al., 2007; Konold & Pollatsek, 2002) zoals steekproef, variatie en verdelingen. Deze conceptuele problemen worden verergerd door een sterke onderwijsfocus op het aanleren van complexe, formele procedures.

Om de moeilijkheden van leerlingen te overbruggen is in de afgelopen decennia gezocht naar *informele* onderwijsbenaderingen om conceptueel begrip te promoten. Het aanbieden van informele inferentiële activiteiten op jongere leeftijd zou het leren van de complexere inferentiële statistiek op latere leeftijd kunnen vereenvoudigen (Zieffler et al., 2008). Het gaat hierbij om het trekken van conclusies vanuit informele statistische kennis, dus niet vanuit formele procedures zoals hypothese toetsen of berekeningen met de normale verdeling. Makar en Rubin (2009) definiëren informele statistische inferentie in de volgende principes: het generaliseren van steekproefdata naar een groter geheel; data als bewijs van deze generalisatie; redeneren over de waarschijnlijkheid van deze generalisatie. Nieuwe digitale middelen bieden mogelijkheden voor het simuleren van steekproeven, waarmee leerlingen op informeel niveau de kernconcepten voor statistische inferentie kunnen onderzoeken.

Het gebruik van technologie is onmisbaar voor het doen en leren van statistiek (Gal, 2002; Thijs, Fisser, & Van der Hoeven, 2014). De inzet van recente digitale leeromgevingen met opties voor statistisch modelleren, zoals VUstat en TinkerPlots, biedt een informele aanpak om het begrip van statistische concepten en modellen te verdiepen (Biehler, Frischemeier, & Podworny, 2017). Inzicht in statistische modellen is van fundamenteel belang voor het interpreteren van statistische inferenties (Manor & Ben-Zvi, 2017). Statistische modellen helpen om de waarschijnlijkheid van op steekproefdata gebaseerde conclusies te duiden. Digitale middelen voor het simuleren van steekproefdata uit populatiemodellen maken concepten visueel en toegankelijk. Het modelleren met zulke digitale middelen is veelbelovend voor het statistiekonderwijs nu en in de toekomst.

² Tevens geldt voor ons land dat de leerlingen met een technisch profiel in de bovenbouw—vanaf vwo 4—helemaal geen inferentiële statistiek krijgen, tenzij ze wiskunde D kiezen.

Kortom: het onderwijzen van statistische inferentie is belangrijk maar ook moeilijk. Het inbedden van informele statistische inferentie in eerdere leerjaren lijkt veelbelovend, met name in combinatie met het gebruik van digitale leermiddelen. Er is echter nog weinig bekend over hoe we ons huidige onderbouwcurriculum kunnen uitbreiden met een goed onderbouwd leertraject. Dit onderzoeksproject beoogt kennis te verwerven over een theoretisch en empirisch gefundeerd leertraject om statistische inferentie te introduceren bij vwo 3-leerlingen (Grade 9).

Hoofdstuk 1: Introductie

Dit onderzoeksproject volgde een ontwerpgerichte aanpak (Bakker, 2018). Deze aanpak kenmerkt zich door een cyclisch proces waarin onderwijsmateriaal voor leeromgevingen wordt ontworpen, geïmplementeerd en geëvalueerd, voor vervolgcycli van (her)ontwerp en testen (McKenney & Reeves, 2012). In de beginfase richtten we ons vooral op de ontwikkeling van een theoretisch gefundeerd ontwerp, met daarin een specificatie van beoogde leerdoelen en de uitwerking hiervan in een—op dat moment nog hypothetisch—leertraject. Naarmate het onderzoek vorderde, werden meerdere interventies met het leertraject uitgevoerd in de lespraktijk en geëvalueerd. Deze interventies werden in iedere cyclus opgeschaald in zowel de lengte van het leertraject als in het aantal deelnemers. In dit onderzoek zijn drie cycli doorlopen: beginnend met een onderwijsexperiment in één klas, via een interventie in drie klassen, naar een interventie in 13 klassen op verschillende scholen. Daarnaast is tussen cyclus 2 en 3 een verdiepende casestudie uitgevoerd naar het leren van en met technologie. Deze verdiepende casestudie richtte zich op de samenhang tussen het leren van gebruikstechnieken voor een digitale tool en het ontwikkelen van conceptueel statistisch begrip.

Hoofdstuk 2. Herhaalde steekproeven met een black box als opstap naar statistische inferentie

Dit hoofdstuk presenteert de resultaten uit de eerste cyclus—en daarmee de eerste studie—van dit ontwerponderzoek. Succesvolle implementatie van theorie in de onderwijspraktijk impliceert het stapsgewijze ontwerp en de evaluatie in echte klaslokalen van krachtige leertrajecten die ons huidige begrip van effectief leren belichamen (De Corte, 2000). De eerste cyclus richtte zich daarom op het ontwerp, de implementatie en de evaluatie van het eerste deel van het leertraject: leerstap 1 tot en met 3.

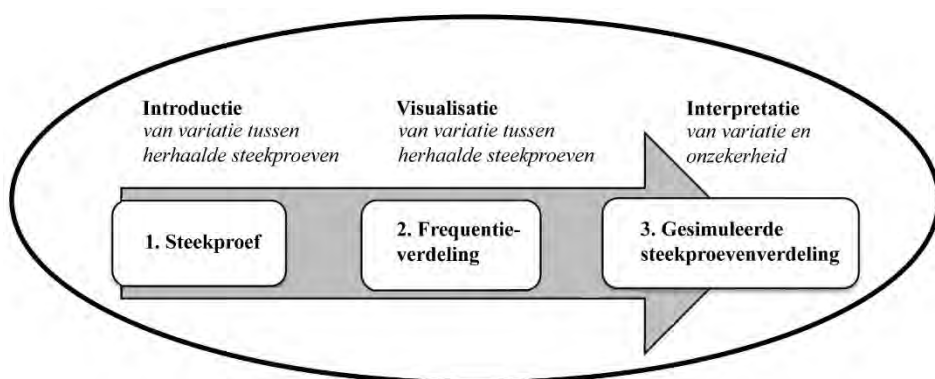


Figuur 1. Afbeelding fysieke black box gevuld met balletjes

Allereerst werd op basis van literatuurstudie een hypothetisch leertraject (Simon, 1995) ontwikkeld voor het introduceren van drie kernbegrippen voor statistische inferentie: steekproef, frequentieverdeling en gesimuleerde steekproevenverdeling (ICT). Figuur 2 toont een overzicht van de drie kernbegrippen.

Een hypothetisch leertraject bestaat uit leerdoelen voor de leerlingen, een beschrijving van leeractiviteiten met bijbehorende hulpmiddelen, materialen en taakstructuren, leerlingkenmerken, en onderwijsmethoden die leiden tot het vereiste leerproces en de beoogde leerdoelen (Sandoval, 2014; Simon, 1995). Het door ons ontworpen hypothetische leertraject werd vervolgens geïmplementeerd in één klas met 20 vwo 3-leerlingen. Voor de evaluatie van het traject werden bij elke leerstap indicatoren opgesteld over observeerbaar leergedrag van leerlingen die de hypothese van iedere stap ondersteunen.

De hypothese in leerstap 1 was dat leerlingen zich bewust zouden worden van steekproefvariatie door het uitvoeren van experimenten met een fysieke black box gevuld met balletjes. Door het uitvoeren van herhaalde experimenten met een klein en groot kijkvenster, konden ze het effect van herhaalde steekproeven en steekproefomvang op de schatting van de populatie (inhoud black box) exploreren. De resultaten toonden aan dat de met de hypothese verbonden indicatoren werden waargenomen. De eerste leerstap stelde de leerlingen in staat om in korte tijd te redeneren met steekproefdata, inclusief het (informeel) interpreteren van variatie en onzekerheid. Zie Figuur 3 voor een impressie van leerstappen 1 tot en met 3.



Figuur 2. Overzicht van de drie kernbegrippen voor statistische inferentie, zoals ingebed in stappen 1 tot en met 3 van het leertraject

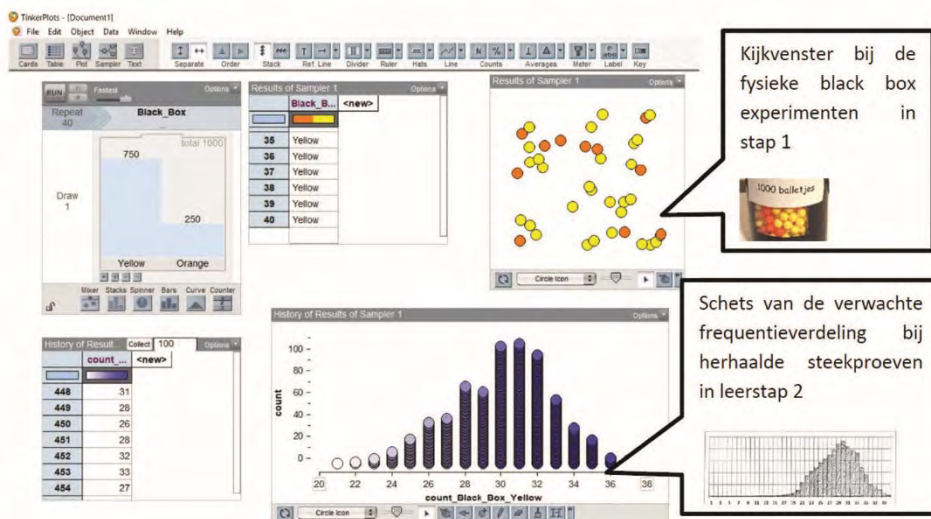


Figuur 3. Impressies van leerstappen 1 tot en met 3. Links twee foto's van leerstap 1, rechtsboven leerstap 2 en rechtsonder leerstap 3

Leerstap 2 van het leertraject was gericht op het concept van de frequentieverdeling bij herhaalde steekproeven. De hypothese was dat leerlingen deze frequentieverdeling allereerst zouden interpreteren als een (visualisatie) *model van* de verkregen resultaten bij een bepaalde black box en dat de ontworpen leeractiviteiten hen zouden stimuleren om de conceptuele overstap te maken naar het gebruik hiervan als *model voor* het interpreteren van variatie en onzekerheid—deze conceptuele overstap van een *model van* een specifieke wiskundige situatie naar een *model voor* een netwerk aan gerelateerde wiskundige situaties is bekend als *emergent modeling* (Gravemeijer, 1999). In deze leerstap stonden twee activiteiten centraal. Ten eerste het schetsen van de verwachte frequentieverdeling bij 100.000 herhaalde steekproeven met de fysieke black box uit de eerste leerstap, en ten tweede het gebruik van een dergelijke frequentieverdeling om de waarschijnlijkheid van specifieke steekproefresultaten te bepalen. Uit de resultaten van deze leerstap bleek dat de meeste indicatoren werden waargenomen. Leerlingen waren in staat om een correcte schets te maken en de kans op specifieke steekproefresultaten te bepalen in de context van de black box—bijvoorbeeld

het bepalen van de kans op een steekproefresultaat van meer dan 35 gele balletjes bij een steekproef van 40 uit een populatie met een proportie van 75%.

In leerstap 3 lag de focus op het conceptualiseren van de gesimuleerde steekproevenverdeling. Hiervoor gebruikten leerlingen statistisch modelleren in een digitale omgeving om herhaalde steekproeven met de bijbehorende steekproevenverdeling te simuleren. De hypothese was dat leerlingen zouden begrijpen dat deze gesimuleerde verdeling kan worden gebruikt als een model voor het interpreteren van variatie en onzekerheid. In deze leerstap neemt de computer als het ware hun handwerk uit de eerste twee leerstappen over. De resultaten van leerstap 3 toonden aan dat ook hier de indicatoren die de hypothese ondersteunen, werden waargenomen.



Figuur 4. De samenhang tussen het werken in de digitale omgeving van TinkerPlots in leerstap 3, en de fysieke black box-activiteiten in leerstappen 1 en 2

Op basis van de bevindingen in deze studie vermoedden we dat de sterke samenhang en opbouw tussen de drie leerstappen het voor leerlingen mogelijk maakte om deze probleemloos te doorlopen. Vanuit hun concrete ervaringen met steekproefvariatie in leerstap 1, gevolgd door het visualiseren van de opschaling van dit experiment in leerstap 2, konden leerlingen gemakkelijk de overgang maken naar het modelleren en interpreteren van de gesimuleerde steekproevenverdeling in leerstap 3. Zie figuur 4 voor een illustratie van deze samenhang tussen leerstap 1 tot en met 3. Deze eerste drie stappen van het

leertraject gaven leerlingen het benodigde inzicht in hoe een steekproevenverdeling ontstaat en hoe deze kan worden gebruikt als model voor het interpreteren van variatie en onzekerheid. Deze bevindingen suggereerden een veelbelovende manier om leerlingen te laten kennismaken met (informele) statistische inferentie.

Hoofdstuk 3. Statistische modelleerprocessen bekeken door de lens van instrumentele genese

Om meer inzicht te krijgen in het leren van en met technologie in leerstap 3 en verder werd een verdiepende casestudie uitgevoerd. Inzicht in leren met digitale middelen is voorwaardelijk om deze effectief te kunnen inzetten voor het bereiken van beoogde leerdoelen. Digitale leermiddelen voor statistiek, zoals TinkerPlots, bieden mogelijkheden voor statistisch modelleren via een informele aanpak. Deze digitale middelen faciliteren leerlingen om populatiemodellen te bouwen en deze modellen te gebruiken om steekproefdata te simuleren. Dit statistisch modelleren bevordert het inzicht in concepten en modellen die fundamenteel zijn voor statistische inferentie (Biehler et al., 2017; Manor & Ben-Zvi, 2017).

Vanuit wiskundeonderwijs is bekend dat het aanleren van gebruikstechnieken voor een digitale tool en het ontwikkelen van conceptueel begrip met elkaar verweven zijn. Tot nu toe heeft deze verwevenheid van gebruikstechnisch en conceptueel begrip, bekend als instrumentele genese (Artigue, 2002), weinig aandacht gekregen in onderzoek naar statistiekonderwijs met digitale middelen. Deze verdiepende casestudie richtte zich daarom op de toepasbaarheid van het theoretisch perspectief van instrumentele genese binnen statistiekonderwijs, en meer specifiek bij het statistisch modelleren in de digitale omgeving van TinkerPlots.

Een geschikte fase om de instrumentele genese van leerlingen te onderzoeken is na de introductie van de tool en de concepten, bij het toepassen van de verworven kennis in nieuwe situaties. Deze fase vindt plaats in leerstap 4 van het leertraject. De data voor dit onderzoek bestonden uit video- en audio-opnames van twee laboratoriumsessies met in totaal 28 leerlingen uit vwo 3 bij het uitvoeren van leeractiviteiten in stap 4 van het traject. In het bijzonder analyseerden we hoe de ontwikkeling van (gebruiks)technieken en conceptueel begrip verweven waren in de instrumentatieschema's die leerlingen ontwikkelden. We identificeerden zes instrumentatieschema's, A tot en met F, voor statistisch modelleren met TinkerPlots. Figuur 5 illustreert als voorbeeld instrumentatieschema C. De linkerzijde van het figuur bevat een beknopte

beschrijving van het schema, in het midden is een schermafbeelding vanuit TinkerPlots weergegeven met een duiding van de gebruikte technieken, en de rechterzijde beschrijft het conceptueel begrip dat in dit schema aan de orde is.

Instrumentaties Schema C: Visualiseer herhaalde steekproeven

Simuleer herhaalde steekproeven en visualiseer deze in een steekproevenverdeling; Onderzoek het gedrag van het model

TinkerPlots technieken



Conceptueel begrip

Bij het oplossen van een realistisch probleem kan het modelleren hiervan inzicht bieden. Door het invoeren van een (verwacht) populatiemodel, het simuleren van herhaalde steekproeven en het visualiseren hiervan in een steekproevenverdeling, kan het gedrag van het model worden onderzocht—bijvoorbeeld het verkennen van veel voorkomende, uitzonderlijke hoge en lage steekproefresultaten. Een steekproevenverdeling is een weergave van de resultaten bij veel herhaalde (gesimuleerde) steekproeven in een frequentieverdeling. Langs de horizontale as staan de mogelijke steekproefresultaten en de verticale as geeft aan hoe vaak bepaalde resultaten voorkomen. Steekproefresultaten van eenzelfde populatie variëren op basis van toeval, waarbij resultaten die dichtbij het (populatie)model liggen vaker zullen voorkomen dan sterk afwijkende resultaten.

Figuur 5. Voorbeeld van instrumentatieschema C voor statistische modellerprocessen met TinkerPlots

We observeerden een sterke verwevenheid tussen het aanleren van technieken en het ontwikkelen van conceptueel begrip. Technieken voor het gebruik van TinkerPlots hielpen de leerlingen om contextonafhankelijke technische patronen te ontdekken, die de belangrijke conceptuele overstap van een *model van* naar een *model voor* (Gravemeijer, 1999) bevorderden. Meer concreet betekende dit dat leerlingen ontdekten dat gebruikstechnieken in specifieke contexten meer algemeen, dus contextonafhankelijk, toegepast konden worden. Dit ging gepaard met meer abstracte statistische terminologie—bijvoorbeeld het invoeren van de steekproefomvang in plaats van het aantal bevraagde leerlingen. Omgekeerd leidde het conceptuele begrip van de leerlingen tot de verkenning van meer geavanceerde digitale technieken. Deze bevindingen toonden aan dat investeren in het aanleren van digitale technieken tegelijkertijd een positief effect heeft op het ontwikkelen van statistisch begrip.

Hoofdstuk 4. Introductie in statistische inferentie: Ontwerp van een theoretisch en empirisch onderbouwd leertraject

In dit hoofdstuk worden de resultaten van de derde studie—gebaseerd op onderzoekscyclus 3—gepresenteerd. Op basis van de eerste twee onderzoekcycli, de verdiepende casestudie en aanvullend literatuuronderzoek werd het (hypothetische) leertraject (her)ontworpen voor de derde cyclus. Deze cyclus omvatte het gehele traject van acht leerstappen, opgesplitst in twee vergelijkbare delen van vier: (1) experimenteren met een fysieke black box, (2) visualiseren van verdelingen, (3) onderzoeken van steekproevenverdelingen met behulp van simulatiesoftware, (4) interpreteren van steekproevenverdelingen voor inferenties in realistische contexten. De stappen 1 tot en met 4 zijn alleen gericht op categoriale data en in de stappen 5 tot en met 8 wordt gewerkt met numerieke data. Een overzicht van het gehele leertraject is weergegeven in tabel 1.

De focus van deze studie was gericht op empirisch onderzoeken *of* en *hoe* het (vanuit bestaande theorieën) ontworpen leertraject het inzicht van leerlingen in statistische inferentie stimuleert. Hiervoor werd het leertraject geïmplementeerd in een interventie onder 267 leerlingen in 13 klassen op verschillende scholen. De tijdsomvang van het leertraject bestond uit zes lesuren per deel, met een totaal van 12 lesuren. We analyseerden de posttestresultaten van de leerlingen na de interventie om te onderzoeken *of* het traject inderdaad de beoogde leerstapgerelateerde doelen voor statistische inferentie stimuleerde. Om de posttestresultaten te kunnen interpreteren werden deze vergeleken met die van een vergelijkingsgroep ($n = 217$) die het reguliere

vwo 3-curriculum gevolgd had. De reguliere aanpak bestond uit 10–12 lessen gericht op beschrijvende statistiek. Tevens analyseerden we de werkbladen van de leerlingen tijdens elke leerstap om te onderzoeken *hoe* het stapsgewijze traject het leerproces bevorderde.

De posttestresultaten toonden aan dat leerlingen die les kregen vanuit het leertraject significant hoger scoorden op *alle* specifiek leerstapgerelateerde doelen uit het leertraject dan leerlingen van een vergelijkingsgroep ($n = 217$) die het reguliere curriculum volgden. Deze leerdoelen omvatten in leerstap 1 en 5 het gebruik van steekproeven, in leerstap 2 en 6 het visualiseren van verdelingen, in leerstap 3 en 7 het effect van herhaalde steekproeven en steekproefomvang, en in leerstap 4 en 8 het interpreteren van inferenties in realistische contexten. Dit betekent dat elk onderdeel uit de opbouw in leerstappen, zoals gepresenteerd in de laatste kolom van tabel 1, van essentieel belang is voor het totale leertraject.

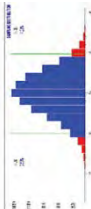
De analyse van werkbladen en notities van docenten en onderzoeker bevestigden het belang van elk onderdeel en de sterke samenhang in opbouw tussen elke leerstap uit het traject. Met name de koppeling tussen het fysieke experiment met de black box en de digitale leeromgeving (zoals weergegeven in figuur 4) bevorderde het inzicht van leerlingen in statistische modellen en modelleren. Dit inzicht maakte het voor leerlingen mogelijk om vervolgens in leerstap 4 en 8 inferenties te interpreteren in realistische contexten.


De bevindingen in deze studie toonden aan *dat* en *hoe* het ontworpen leertraject werkt. Een aanpak gebaseerd op herhaalde steekproeven met een black box gecombineerd met statistisch modelleren in de digitale omgeving van TinkerPlots, bleek vruchtbaar voor het introduceren van statistische inferentie. Beide ideeën hebben tevens potentieel voor inbedding in meer complexe vervolgactiviteiten, zoals het toetsen van hypothesen en het vergelijken van groepen. Deze bevindingen suggereren dat informele inferentiële activiteiten al in de onderbouw van het voortgezet onderwijs geïntroduceerd kunnen worden, zodat beter kan worden geanticipeerd op vervolgstappen van leerlingen in statistiekonderwijs.

Tabel 1. Overzicht van stap 1 – 8 van het leertraject

Leerstap	Beschrijving	Voorbeeld van activiteiten	Leerdoel	Opbouw in leerstappen
Categorale data				
1. Experimenteren met een fysieke black box	Fysiek black box met balletjes experiment (met klein en groot kijkvenster)	Schat het aantal gele balletjes in de black box, gevuld met 1000 balletjes, door het (herhaaldelijk) schudden en onderzoeken van het aantal zichtbare balletjes	Leerlingen formuleren inferenties en maken kennis met concepten als steekproef, steekproefomvang, steekproefvariatie, frequentie en centrum- en spreidingsmaten, in de context van herhaalde steekproeven met een fysieke black box	Leerlingen ervaren dat steekproefresultaten variëren en dat een grotere steekproefomvang en meer herhalingen leiden tot een betere schatting van de populatie. Vervolgvrraag: Wat gebeurt er als we de steekproefomvang en het aantal herhaalde steekproeven verder vergroten? Het uitvoeren van meer en grotere steekproeven is tijdrovend, daarom kan een gedachtenexperiment helpen.
2. Visualiseren van verdelingen	Grafiek als model (visualisatie) voor de frequentieverdeling bij herhaalde steekproeven met	Schets de frequentieverdeling die je verwacht als het black box-experiment met groot venster	Leerlingen kunnen een visualisatie tekenen van de verwachte steekproevenverdeling bij herhaalde steekproeven.	



een black box	100.000 keer wordt herhaald	Leerlingen interpreteren de steekproevenverdeling om inferenties te formuleren over de waarschijnlijkheid van specifieke steekproefresultaten	De steekproevenverdeling van herhaalde steekproeven kan gebruikt worden om de waarschijnlijkheid van specifieke steekproefresultaten te bepalen.
3. Modelleren van een black box (ICT)	<div><div>Simulaties van herhaalde steekproeven met een gemiddelde black box in een steekproevenverdeling, voor het bepalen en interpreteren van de waarschijnlijkheid van specifieke resultaten</div><div></div></div>	Gebruik TinkerPlots om de meest voorkomende steekproefresultaten te bepalen voor een black box gevuld met 750 gele en 250 oranje balletjes, bij een steekproefomvang van 40	
Leerlingen gebruiken statistisch modelleren in de digitale omgeving van TinkerPlots om (on)waarschijnlijke steekproefresultaten te bepalen, in de context van een black box [Statistisch modelleren omvat het invoeren van een model, simuleren van (herhaalde) steekproeven, visualiseren van de steekproevenverdeling en het interpreteren van de resultaten]			Statistisch modelleren— inclusief het interpreteren van de steekproevenverdeling bij herhaalde steekproeven— kan gebruikt worden om de waarschijnlijkheid van specifieke

4. Modelleren van realistische contexten (ICT)	Modelleren van een specifieke situatie in TinkerPlots, voor het simuleren en interpreteren van de steekproevenverdeling en het beter begrijpen van de situatie (inclusief de bijbehorende onzekerheid)		Gebruik TinkerPlots om te bepalen welke steekproefresultaten je kunt verwachten als een steekproef met 30 leerlingen wordt uitgevoerd op een school met 300 leerlingen, bij onderzoek naar het aantal leerlingen dat dagelijks ontbijt (gegeven dat gemiddeld 70% van de VO-leerlingen dagelijks ontbijt)	Leerlingen gebruiken statistisch modelleren in de digitale omgeving van TinkerPlots voor het formuleren van inferenties, in de context van realistische probleemstellingen	Vervolg vraag: Kan statistisch modelleren meer algemeen gebruikt worden, in andere situaties en contexten?	Vanuit stap 1 tot en met 4 ontstaat de vraag hoe statistisch modelleren gebruikt kan worden bij andere—niet categoriale—data.
	Numerieke data					
5. Experimenteren met een fysieke black box	Fysiek black box met briefjes experiment. (De black box is	Neem een steekproef van 40 briefjes en vat de steekproefdata	Leerlingen formuleren inferenties in de context van een fysieke black box met briefjes			



gevuld met 4000 briefjes. Elk briefje bevat gegevens over het geslacht en de lichaamslengte van een leerling uit klas 2, bijvoorbeeld: jongen—155 cm)

samen (bereken centrum- en spreidingsmaten, gebruik visualisaties). Maak een schatting van de lichaamslengte van de 4000 leerlingen uit klas 2

(geslacht en lichaamslengte van 4000 leerlingen uit klas 2) met inachtneming van steekproefomvang, steekproefvariatie en centrummaten

Leerlingen bespreken hoe numerieke data van herhaalde steekproeven gebruikt kunnen worden om inferenties over de populatie te formuleren.

Vervolg vraag: Hoe kan de onderliggende populatieverdeling—de lichaamslengte van de 4000 leerlingen op de briefjes in de black box— weergegeven worden op basis van de gevonden steekproefresultaten?

6. Visualiseren van verdelingen



Verwachtingen over de populatie (lengte van 4000 leerlingen) samenvatten en visualiseren op basis van de gevonden steekproefdata in leerstap 5

Schets de frequentieverdeling die je verwacht voor de totale populatie, op basis van de gevonden steekproefresultaten in leerstap 5

Leerlingen schetsen een visualisatie van de populatieverdeling die ze verwachten op basis van de gevonden steekproefresultaten. Leerlingen maken inferenties over de populatieverdeling en het populatie-gemiddelde.

Leerlingen formuleren inferenties over het populatiegemiddelde en de populatieverdeling op basis van de gevonden steekproefdata.

Vervolg vraag: Wat zijn de effecten van grotere en meer steekproeven op de schatting van het populatiegemiddelde en de populatieverdeling?

Het gebruik van technologie kan helpen bij het onderzoeken van deze effecten

Vanuit leerstap 7 ontstaat de vraag hoe het statistisch modelleren met numerieke data kan worden toegepast in andere contexten en situaties

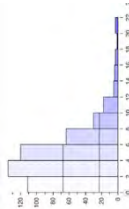
Leerlingen gebruiken statistisch modelleren in de digitale omgeving van TinkerPlots om (on)waarschijnlijke steekproefresultaten te bepalen, in de context van de black box met briefjes. [Statistisch modelleren omvat hier het simuleren van (herhaalde) steekproeven vanuit een *gegeven* model, visualiseren van de steekproefverdeling van het steekproef-gemiddelde, en het interpreteren van de resultaten]

Experimenteren met simulaties van (herhaalde) steekproeven, bij variatie in omvang en aantal herhalingen, in de context van de black box met briefjes uit leerstap 5 en 6

Gebruik TinkerPlots om de meest voorkomende steekproefresultaten te bepalen voor de (gegeven) gemiddelde black box met briefjes die hoort bij leerstap 5

7. Modelleren van een black box (ICT)



<p>8. Modelleren van realistische contexten</p>	 <p>(ICT)</p>	<p>Modelleren van een realistische situatie in TinkerPlots, voor het simuleren en interpreteren van de steekproevenverdeling, en het beter begrijpen van de situatie (inclusief de bijbehorende onzekerheid).</p>	<p>Gebruik TinkerPlots voor het simuleren van herhaalde steekproeven (omvang 200) vanuit een verborgen populatiemodel van 4000 personen om de sporttijd van deze personen te onderzoeken en bepalen</p>	<p>Leerlingen gebruiken statistisch modelleren in de digitale omgeving van TinkerPlots om inferenties te formuleren, in realistische contexten en probleemstellingen</p>
---	--	---	---	--

Hoofdstuk 5. De effecten van het leertraject voor statistische inferentie op de statistische geletterdheid van leerlingen

Als laatste studie in dit onderzoeksproject, werden de effecten van het leertraject op de *algehele* statistische geletterdheid van leerlingen onderzocht. Statistische geletterdheid kan onderverdeeld worden in drie domeinen: (1) Statistische inferentie, (2) Grafieken en variatie, (3) Kans en gemiddelde. Het ontworpen leertraject was vooral gericht op statistische inferentie, het eerste domein. We vermoedden echter dat het leertraject ook positieve invloed zou hebben op de andere twee domeinen. Het reguliere vwo 3-curriculum was enkel gericht op domein twee en drie. Voor de evaluatie werd gebruik gemaakt van een pre-post onderzoekszet met de interventiegroep ($n = 267$) uit de vorige studie—de derde onderzoekscyclus. De pre- en posttest werden ontwikkeld op basis van reeds bestaande tests om de statistische geletterdheid—op alle drie domeinen, maar met name op statistische inferentie—van leerlingen te onderzoeken (uit onderzoek van Watson & Callingham, 2003; Watson & Callingham 2004; Callingham & Watson, 2017; delMas et al., 2007). Om de significante leerwinst van leerlingen uit de interventiegroep te interpreteren, hebben we de resultaten van deze leerlingen vergeleken met een nationale baseline en internationale prestaties. Voor de nationale baseline gebruikten we de pre- en posttestresultaten van de vergelijkingsgroep ($n = 217$) uit de vorige studie die het reguliere leerplan van vwo 3 volgde, en de internationale vergelijking werd gedaan aan de hand van een Australische studie met vergelijkbare testopzet.

De nationale vergelijking van testresultaten toonde aan dat de interventiegroep significant hoger scoorde op statistische geletterdheid, en in het bijzonder op het domein van statistische inferentie. Tevens vonden we aanzienlijk positieve effecten voor de andere twee domeinen. Hoewel het leertraject niet gericht was op de andere domeinen, bleek het leertraject—bestaande uit een onderzoekgerichte aanpak met digitale middelen en meer complexe leeractiviteiten voor statistische inferentie—ook hier een positief effect te hebben. Tabel 2 toont een overzicht van de resultaten voor de interventie- en vergelijkingsgroep. De laatste kolom in het vak ‘Posttest’ geeft het verschil weer tussen de resultaten van de interventie- en vergelijkingsgroep op de posttest. De eerste kolom in het vak ‘Pre naar post’ geeft de vooruitgang weer van de interventiegroep. Bij de pretestresultaten van de vergelijkingsgroep moet vermeld worden dat deze leerlingen de reguliere statistieklessen voorafgaand aan de pretest gevolgd hadden, waardoor hun pretestresultaten hoger zijn dan die van de interventiegroep. Tussen de pre- en posttest volgde

deze vergelijkingsgroep geen statistieklussen, wat zichtbaar is in de gelijkwaardige resultaten op beide tests voor deze groep”.

Tabel 2. Gemiddelde leerlingenscore op de domeinen van statistische geletterdheid bij de pre- en posttest voor de interventie- en vergelijkingsgroep, inclusief vooruitgang van pre naar post

		Interventie- groep (<i>n</i> = 267) M (SD)	Vergelijkingsgroep (<i>n</i> = 217) M (SD)	Interventie – Vergelijking M(I) – M(C)
Pretest	SG	2.60 (0.61)	2.97 (0.68)	–0.37***
	SI	2.45 (0.65)	2.72 (0.71)	–0.27***
	GV	2.07 (0.63)	2.29 (0.58)	–0.22***
	KG	3.29 (1.38)	3.92 (1.31)	–0.63***
Posttest	SG	3.28 (0.69)	2.95 (0.78)	+ 0.33***
	SI	3.34 (0.84)	2.67 (0.84)	+ 0.67***
	GV	2.59 (0.81)	2.38 (0.88)	+ 0.21*
	KG	3.92 (0.88)	3.80 (1.06)	+ 0.12
Pre naar post	SG	+ 0.68 (0.86)***	–0.02 (0.73)	0.70***
	SI	+ 0.89 (0.92)***	–0.04 (0.71)	0.93***
	GV	+ 0.52 (0.98)***	+ 0.09 (0.94)	0.43***
	KG	+ 0.63 (1.53)*	–0.11 (1.45)	0.74***

* $p < .05$, ** $p < .005$, en *** $p < .0005$

Noot. SG = statistische geletterdheid; SI, GV, KG zijn de drie domeinen binnen SG; SI = statistische inferentie; GV = grafieken en variatie; KG = kans en gemiddelde.

De vergelijking met de internationale studie toonde aan dat de posttestresultaten van de interventiegroep met 14–15-jarigen op statistische geletterdheid het meest overeenkwamen met die van Australische leerlingen in Grade 7–8 met een leeftijd van ongeveer 13 jaar. De resultaten van de vergelijkingsgroep met 14–15-jarigen waren het meest vergelijkbaar met die van Australische leerlingen in Grade 6–7 met een leeftijd van ongeveer 12 jaar. Uit deze internationale vergelijking kunnen we opnieuw concluderen dat de interventiegroep aanzienlijk hoger scoorde, met ongeveer één leerjaar verschil, dan de vergelijkingsgroep. Opvallend is dat de resultaten van beide groepen met

leerlingen in de leeftijd van 14–15 jaar overeenkwamen met die van aanzienlijk jongere Australische leerlingen. Vermoedelijk komt dit doordat het statistiekaanbod in Australië uitgebreider is dan in Nederland.

De bevindingen vanuit zowel de nationale als internationale vergelijking toonden aan dat het leertraject een significant positief effect had op de statistische geletterdheid van de leerlingen, en in het bijzonder op het domein van statistische inferenties. Tevens signaleerden we positieve effecten voor de andere domeinen. Op basis hiervan kunnen we constateren dat huidige statistiecurricula met een sterk beschrijvende focus verrijkt kunnen worden met een inferentiële focus—in ieder geval voor de onderbouw van het vwo. Het voordeel hiervan is dat leerlingen meer leren over statistische inferenties en niet minder over de andere domeinen van statistische geletterdheid, om zo beter te anticiperen op vervolgstappen van de leerling binnen statistiekonderwijs.

Hoofdstuk 6. Algemene discussie

Dit onderzoeksproject heeft kennis opgeleverd over essentiële vernieuwingen in statistiekonderwijs. Theoretische inzichten werden ontwikkeld in nauwe samenhang met een praktisch onderwijsontwerp. Deze inzichten waren zowel inhoudelijk als methodologisch van aard.

Inhoudelijke bijdrage

Op inhoudelijk gebied draagt dit onderzoek bij aan inzicht in de samenhang tussen de ontwikkeling van statistische inferentie en statistische geletterdheid. Statistische inferentie wordt beschouwd als een complex domein van statistische geletterdheid, wat vaak pas op latere leeftijd wordt aangeboden. De resultaten in dit onderzoek toonden aan dat het ontworpen leertraject met (informele) inferentiële activiteiten een significant positief effect had op het domein van statistische inferentie, en eveneens op de andere twee domeinen van statistische geletterdheid—de domeinen grafieken en variatie, en gemiddelde en kans, beiden met een beschrijvende focus. Dit positieve effect van (informele) inferentiële activiteiten op de andere domeinen van statistische geletterdheid pleit voor het eerder introduceren hiervan. Leerlingen ontwikkelen dan al op vroege leeftijd statistische concepten die noodzakelijk zijn voor statistische inferentie en voor statistische geletterdheid. Het integreren van (informele) statistische inferentie bij de huidige aanpak voor statistische geletterdheid kan zo leiden tot een duurzame verandering in het leren van leerlingen. Het grote voordeel hiervan is dat leerlingen meer leren over inferenties, en hierdoor beter worden voorbereid op hun vervolgstappen in statistiekonderwijs.

Een tweede inhoudelijke bijdrage van het onderzoek betreft het netwerken van theorieën. Het statistiekonderwijs wordt steeds meer gezien als onderscheidend van het wiskundeonderwijs, met eigen perspectieven op onderwijzen en leren (Groth, 2015). Het integreren van onderwijsperspectieven vanuit verschillende disciplines is wenselijk om tot nieuwe kennis en inzichten te komen. Dit onderzoek draagt bij door theoretische perspectieven uit onderzoek naar wiskundeonderwijs te integreren in onderzoek naar statistiekonderwijs. Het theoretisch perspectief van Realistisch Wiskundeonderwijs (Freudenthal, 1983) werd gebruikt bij het ontwerp van dit leertraject voor statistiek. Op basis van de ontwerpheuristieken vanuit deze theorie werd het black box-paradigma uitgewerkt in concrete leeractiviteiten. Het black box-paradigma bleek effectief als leidende activiteit binnen de leerstappen van het traject. Het theoretisch perspectief van Instrumentele Genese werd gebruikt voor onderzoek naar het leren van en met technologie. Het toepassen van dit perspectief leidde tot inzicht in hoe leerlingen uit vwo 3 concepten ontwikkelen bij het statistisch modelleren in TinkerPlots. Vanuit deze bevindingen lijkt het theoretisch perspectief van instrumentele genese breder inzetbaar binnen onderzoek naar statistiekonderwijs, zoals bij de inzet van andere digitale middelen en in andere onderwijsleerjaren en niveaus.

Methodologische bijdrage

Op methodologisch gebied draag dit onderzoek bij door te laten zien hoe de complexiteit die gepaard gaat bij het experimenteren met innovatief onderwijsmateriaal, overwonnen kan worden door gebruik te maken van ontwerpgericht onderzoek (Bakker, 2018). Een ontwerpgerichte aanpak met een cyclische opschaling in zowel het aantal deelnemers als in de lengte van het leertraject bleek effectief voor het ontwerp en de evaluatie van het innovatieve leertraject. De start met een kleinschalige interventie in de eigen klas van de docent-onderzoeker maakte het mogelijk om de leerdoelen voor het traject te expliciteren en de haalbaarheid ervan te beproeven. De evaluatie was hier vooral gericht op de eerste drie stappen van het leertraject. In deze stappen werd het fundament gelegd van het leertraject en de resultaten uit deze cyclus werden dan ook als uitgangspunt gebruikt voor het ontwerp van de vervolgstappen. In cyclus 2 werd opgeschaald naar drie klassen met 60 leerlingen. De evaluatie was hier voornamelijk gericht op leerstap 4. Aangezien leerstap 5 tot en met 8 een vergelijkbare aanpak en benadering hadden als leerstap 1 tot en met 4, konden we door deze stapsgewijze opschaling een constructief ontwerp realiseren.

In cyclus 3 werd een kwantitatieve benadering gebruikt om de effecten van het leertraject te onderzoeken. Een kwantitatieve aanpak wordt zelden gecombineerd met ontwerpgericht onderzoek. Het kwantificeren, en daarmee samenhangend het opschalen naar een grote groep deelnemers, is een intensief proces. Bij het kwantificeren van de effecten van een leertraject is het van belang dat alle materialen eenduidig, compleet en haalbaar zijn, zodat het traject op de beoogde wijze door docenten kan worden uitgevoerd. Voor de evaluatie van het traject werd een pre-posttestaanpak met een interventie- en vergelijkingsgroep gebruikt. Bij de analyse van de testresultaten werd gekeken naar de prestaties van de leerlingen voor statistische geletterdheid, en tevens naar hun score op leerstapgerichte items. Daarnaast werden in cyclus 3 de werkbladen van leerlingen uit de interventiegroep geanalyseerd. Deze aanpak maakte het mogelijk om empirisch aan te tonen *dat* het leertraject werkt, en tevens *hoe* het leertraject werkt. Dit onderzoek toont aan hoe het werken met simulaties in een digitale omgeving van meerwaarde kan zijn op een zuiver fysieke onderwijsaanpak

Beperkingen van het onderzoek

Zoals elke studie heeft ook dit onderzoek uiteraard beperkingen. In dit onderzoek hebben we aangetoond dat het leertraject voor het introduceren van statistische geletterdheid een positief effect heeft op het leren van statistische inferentie bij vwo 3-leerlingen. Het is echter mogelijk dat andere aanpakken ook werken, waardoor niet zeker is of dit leertraject ook de meest effectieve manier is. Het onderzoek toont echter wel aan dat het ontworpen leertraject werkt. Bij het evalueren van de effecten van het leertraject is het moeilijk om de generaliseerbaarheid en causaliteit te waarborgen. Door te werken met een grote groep leerlingen met verschillende docenten op diverse scholen- bieden de resultaten een sterke indicatie dat het doorlopen van het leertraject (bij uitvoering zoals beoogd) een positief effect heeft op het leren van leerlingen.

Aanbevelingen voor vervolgonderzoek en de onderwijspraktijk

Vanuit deze studie doen we een aantal aanbevelingen voor vervolgonderzoek en de lespraktijk. Het aanvullen van bestaande statistiecurricula met (informele) statistische inferentie lijkt haalbaar en wenselijk. Het veranderen van bestaande curricula is echter complex. Meer onderzoek is nodig voor succesvolle implementatie. Het paradigma van de black box lijkt tevens toepasbaar voor andere onderwijsniveaus en ook voor meer complexe vervolgactiviteiten zoals hypothese toetsen. Het ontwikkelen van efficiënte leertrajecten hiervoor vereist nader onderzoek. De inzet van technologie is onmisbaar voor statistiek (onderwijs), en voor statistische inferentie in het bijzonder. Wiskundedocenten

zijn vaak onervaren in het gebruik van digitale leermiddelen in de les. Tevens worden statistieklessen in de onderbouw van het voortgezet onderwijs vaak verzorgd door tweedegraads wiskundedocenten die onervaren zijn in het doceren van inferentiële statistiek. Onderzoek naar hoe docenten toegerust kunnen worden voor het doceren van inferentiële statistiek met behulp van technologie is wenselijk. Daarnaast kampen veel scholen nog met praktische beperkingen bij de inzet van computers en het installeren van software. Onderzoek naar mogelijkheden om deze praktische obstakels te beperken kan het gebruik van technologie in (statistiek)onderwijs bevorderen. Tot slot veroorzaakte de COVID-pandemie en bijbehorende schoolsluiting een overweldigende toename van technologie in de onderwijspraktijk. Deze actuele ontwikkeling vraagt om onderzoek naar duurzame onderwijsvernieuwingen waarin het gebruik van technologie geïntegreerd kan worden in het reguliere onderwijssysteem.

Persoonlijke reflectie als docent-onderzoeker

Dit onderzoeksproject heeft een rijke bijdrage geleverd aan mijn professionele ontwikkeling als docent en als onderzoeker. Dit onderzoek heeft mijn docentschap op zowel micro-, meso- als macroniveau (Akkerman & Bruining, 2016) versterkt. Op microniveau in mijn eigen lespraktijk als docent heeft dit traject inzicht gegeven in leerprocessen van leerlingen en hoe deze bij het lesgeven gepromoot kunnen worden. Op mesoniveau als docent in de school heeft dit traject geleid tot een meer analytische blik op het schoolsysteem, en op vernieuwende (inter)nationale onderwijsaanpakken en methoden. Op macroniveau van de (regionale en landelijke) onderwijswereld zijn de onderzoeksresultaten via verschillende docentworkshops en artikelen in vaktijdschriften voor wiskundedocenten gedeeld. Diverse docenten zijn vervolgens zelf aan de slag gegaan met het ontwerpen leertraject in allerlei onderwijsniveaus—zoals in de vwo-bovenbouw en in het hbo. Deze ervaringen vormden een waardevol vervolg en aanvulling op dit onderzoeksproject. Als onderzoeker heb ik mijn competenties in het doen van onderzoek kunnen ontwikkelen. Het functioneren in een wetenschappelijke omgeving heeft mijn kijk op onderwijsonderzoek verdiept en verbreed. Tevens heeft de intensieve samenwerking met internationale collega's mijn visie op onderzoek in allerlei opzichten verruimd. Samenvattend heeft dit traject mijn brede professionele functioneren versterkt—zowel in de klas, binnen de school als binnen de (inter)nationale onderwijs- en onderzoekswereld.



Dankwoord

Dit onderzoek was in veel opzichten een intensief project. Hierbij was de steun van anderen onmisbaar om het traject te doorlopen. De eerste fase was voor mij, als ervaren wiskundedocente, een onzekere zoektocht in een onbekende wetenschappelijke wereld. Het vinden van een passende onderzoeksrichting, het verwerven van benodigde kennis en onderzoeksvaardigheden, en het balanceren tussen onderwijs en onderzoek, vergden in het begin veel tijd, inspanning en doorzettingsvermogen. Mede door de betrokkenheid en begeleiding van directe collega's op het Freudenthal Instituut voelde ik me gaandeweg steeds bekwaamer als onderzoeker en werd mijn passie voor het doen van onderzoek vergroot. Het lesgeven in de onderwijspraktijk gecombineerd met onderzoeken hoe leerlingen leren heeft geleid tot een prachtige verbinding tussen mijn affiniteit voor onderwijs, onderzoek en wiskunde.

Allereerst wil ik Paul en Arthur bedanken. Jullie intensieve, deskundige en betrokken begeleiding was voor mij essentieel. Bij de start van dit project kenden we elkaar nauwelijks, maar al snel was er een vertrouwd contact. Paul, jij bent vanaf het begin mijn houvast en tevens motor geweest. Vrijwel wekelijks hadden we contact, waarbij je heel gericht steeds nieuwe impulsen of andere invalshoeken aangaf. Je was enorm betrokken bij de inhoud, maar had tevens oog voor de persoonlijke kant. Zo was je standaard openingsvraag "Hoe gaat het met je?", en was jouw doel tijdens een overleg om mij "nog meer en dieper te laten nadenken". Dat laatste wist je iedere keer weer voor elkaar te krijgen. Tevens was er ruimte om te sparren over andere zijdelings gerelateerde ontwikkelingen, zoals opgedane ervaringen in de onderwijspraktijk of binnen

het FI, en de gedeelde liefde voor muziek. Deze balans tussen persoonlijke en inhoudelijke begeleiding maakte de samenwerking vertrouwd en effectief. Arthur, jouw gedrevenheid en waardevolle feedback, met name op het gebied van methodologie, heeft mij veel geleerd. Vanuit jouw analytisch perspectief gaf je steeds gerichte feedback over welke onderdelen in het geheel verder uitgediept konden worden. Naast detailopmerkingen, over bijvoorbeeld het gebruik van de En Dash, raakte jouw feedback de essentie van het onderzoek. Dit maakte dat deze vaak verregaande gevolgen had voor de hele onderzoeksopzet. Je zette dan als het ware het fundament recht, waardoor een betere cohesie en structuur ontstond. Dit leidde regelmatig tot waardevolle, inhoudelijke discussies, waarin je met weinig woorden een convergerende oplossingsrichting wist aan te geven. Arthur en Paul, de combinatie van jullie als begeleiders is een sterk concept. Jullie eigen professionele perspectieven die elkaar mooi aanvullen, gecombineerd met een sterke onderlinge relatie, resulteerden in een ijzersterke begeleiding.

Als inspirator voor het doen van onderzoek wil ik Jos Tolboom bedanken. Jos, jij bent degene die mij jaren geleden aanspoorde om wetenschappelijk onderzoek te gaan doen. Dit was nog tijdens mijn masteronderzoek en resulteerde in een eenjarig NRO-onderzoekstraject. Gedurende dit kortlopende traject maakte jij mij wegwijs in de onderzoekswereld en deelde je tal van inspirerende en vernieuwende ideeën. Vanuit dit kortlopende traject ontkiemde mijn passie voor het doen van onderzoek, met als vervolg de uitvoering van dit promotieonderzoek. Tijdens dit promotietraject was jij ook degene die de klankbordgroep aanstuurde. Met name in de eerste jaren was het waardevol om met de personen in deze groep, ieder met hun eigen perspectief, te reflecteren op het onderzoeksproces. Bij deze wil ik naast Jos ook Swier Garst, Theo van den Bogaart, Peter Kop, Rijk Verkerk en Karma Dajani heel hartelijk bedanken voor de prettige klankbordgroepbijeenkomsten waarbij verschillende perspectieven op mijn onderzoek verdiept en bediscussieerd werden. Ook een woord van dank aan Walter Steenhagen voor zijn gedreven inzet en constructieve bijdrage aan het ontwerp en de implementatie van de pre- en posttests in onderzoekscyclus 3. Met betrekking tot internationale contacten wil ik met name Rolf Biehler, Katie Makar en Dani Ben-Zvi bedanken voor de fijne en leerzame samenwerking, die mij geholpen heeft om dit onderzoek in een breder internationaal perspectief te plaatsen.

De directie van de CSG Prins Maurits ben ik ontzettend dankbaar voor hun flexibiliteit, steun en meelevens bij het uitvoeren van dit onderzoekstraject.

In goed overleg was het steeds mogelijk om een werkbare balans te vinden tussen docenttaken en onderzoekstaken. Ook bij het integreren van beide, tijdens het uitvoeren van interventies stonden jullie open voor vernieuwing. Bij deze wil ik ook mijn wiskundecollega's bedanken, met name Arjan van Wijk, Rijk Verkerk en Ellis Peekstok, die direct betrokken waren bij het onderzoek. Bedankt voor jullie inzet, praktische aanvullingen en inhoudelijke ideeën. Hierbij wil ik ook Swier Garst nogmaals noemen, die als wiskundecollega 'van de overkant' meerdere keren als interventiedocent heeft meegewerkt. Ook een woord van dank aan de interventiedocenten in de laatste onderzoekscyclus voor hun inzet en deelname aan het onderzoekproject. Mijn teamleider, Andre Knulst, wil ik persoonlijk bedanken voor zijn interesse en stimulans. Met name wil ik hierbij de waardevolle gesprekken benoemen over de balans tussen wetenschappelijk onderzoek en de onderwijspraktijk, en over mogelijke onderwijsvernieuwingen binnen de school en daarbuiten.

Dan een woord van dank aan mijn kamergenoten en collega's op het Freudenthal Instituut, met name Lonneke Boels en Annemiek van Leendert, voor hun steun en toeverlaat. Lonneke, jij was gedurende het hele traject mijn voorbeeld en houvast. Je maakte me wegwijs in het gebouw, introduceerde me bij personen en deelde handige tips en inhoudelijke ideeën over onderzoek doen. Je oprechte belangstelling, en de tijd die je nam om je te verdiepen in mijn onderzoek, zorgde voor waardevolle positief-kritische feedback. Ook was jij degene met wie ik de soms lastige balans tussen onderzoek, onderwijs en gezin kon delen en bespreken. Je nam altijd de tijd voor overleg. Tijdens onze gezamenlijke conferenties, met name ICOTS in Japan en de pre-SRTL in De Bilt, heb ik veel van je geleerd en genoten van de gezellige momenten. Annemiek, ook jou wil ik bedanken voor je interesse, de openhartige gesprekken en je gezelligheid. Tevens dank aan alle collega's op het FI voor de fijne gesprekken in de wandelgangen, bij de koffieautomaat, tijdens de lunchmeetings en de NWD. Tot slot wil ik Nathalie Kuijpers bedanken voor haar onmisbare hulp bij het zetten van de 'puntjes op de i', qua taalcheck en layout, in de manuscripten van deze thesis.

Mijn mede-Dudoc-ers wil ik bedanken voor hun openheid in het delen van onderzoek ervaringen, de waardevolle inhoudelijke feedback en de gezelligheid tijdens de Dudoc-bijeenkomsten. Marie-Jetta, Sathyam, Gerben, Tim B., Melde, Tore, Tim van D., Farran, Jacqueline, Koen, Stefan, Pier en Kirsten, het was fijn om met 'lotgenoten' te kunnen sparren. Ook een woord van dank aan de Dudoc-programmaraad, Erik Barendsen, Wouter van

Joolingen, Martin Goedhart, Birgit Pepin en Marc de Vries, voor jullie openheid, interesse en inhoudelijke expertise.

Naast de onderwijs- en onderzoekwereld waren vriendinnen van onschatbare waarde om te ontspannen en te relativieren. Marian, met jou kon ik alle mooie en moeilijke momenten delen. Bedankt voor je grenzeloze vertrouwen en onvoorwaardelijke liefdevolle steun: je schouder om op te huilen en je gezelligheid om van te genieten. Bianca, en ook Robert-Jan, bedankt voor je hartelijke meelevens en de plezierige activiteiten met onze gezinnen. Marleen, Marjan R. en Sonja bedankt voor jullie warme vriendschap, en Christa, Tabitha en Annette, bedankt voor de mooie muzikale momenten.

Dan wil ik mijn lieve moeder bedanken die altijd voor me klaarstond. De aangename thee-momentjes waarin we allerlei zaken bespraken, je heerlijke appeltaart bij verjaardagen, je interesse en betrokkenheid bij alle activiteiten rondom ons gezin. Je liefdevolle adviezen en ook de zorgen die je deelde rondom mijn onderzoek en welzijn, waren waardevol om de juiste keuzes te maken. Ook mijn schoonouders wil ik bedanken voor hun steun op allerlei manieren, zoals de ontspannen koffie-uurtjes op de zondagmorgen, de zelf geteelde groente en fruit, en de praktische klusjes in en om het huis.

Tot slot, mijn man en kinderen. Mark, bedankt dat je er altijd voor me was. Joerie, Bas en Jesse, wat ben ik trots op wie jullie (geworden) zijn. Bedankt voor al het moois dat jullie toevoeg(d)en aan ons gezin: voor jullie humor, verhalen en eigen kijk op de (onderwijs)wereld, die gedurende het onderzoekstraject—maar hopelijk ook nog heel lang daarna—zorg(d)en voor ontspanning en liefdevolle momenten van geluk.

Curriculum Vitae

Marianne van Dijke-Droogers was born on April 18, 1975, in Tholen, the Netherlands. After completing her secondary education at pre-university level, she enrolled in the bachelor program for primary education. She obtained her bachelor's degree in Education in 1997. After more than ten years as a teacher in primary education, while also becoming a mother of three children, she started a bachelor's program in mathematics alongside her teaching position. In 2011, she transitioned as a teacher from primary to secondary education, where she combined a parttime position as a mathematics teacher with her parttime bachelor program. She obtained her bachelor's degree in Mathematics Education in 2013 and subsequently her master's degree in Mathematics Education, cum laude, in 2015. She continued to work as a mathematics teacher at the secondary school where she is still employed today. During the 2012–2013 school year she worked for one year as a teacher trainer in mathematics at Rotterdam University of Applied Sciences.



In the 2016–2017 school year, she received a grant from NRO—the Netherlands Initiative for Education Research—for a short-term scientific research project on promoting statistical literacy. This research was supervised by Paul Drijvers from Utrecht University and Jos Tolboom from SLO—the Netherlands Institute for Curriculum Development. Out of this research collaboration, a Dudoc grant—initiated by the Ministry of Education, Culture and Science—was applied for in 2017 to conduct an advanced follow-up study as a PhD project. This research specifically focused on the design of a learning trajectory for statistical inference, with Paul Drijvers as supervisor and Arthur Bakker as co-supervisor. This fellowship was awarded from 2017–2021. Marianne combined this parttime PhD project with her position as a mathematics teacher in secondary education.

Besides her passion for mathematics and education, she has a musical love for playing the piano. She lives with her husband Mark and three sons Joerie, Bas and Jesse. She also enjoys spending time with family and friends.

Publications related to this thesis

- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted). *Effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy*.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (submitted). *Introducing Statistical Inference: Design of a Theoretically and Empirically Based Learning Trajectory*.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2021). Statistical modeling processes through the lens of instrumental genesis. *Educational Studies in Mathematics*. <https://doi.org/10.1007/s10649-020-10023-y>
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2020). Repeated sampling with a black box to make informal statistical inference accessible. *Mathematical Thinking and Learning*, 22(2), 116–138.
- Van Dijke-Droogers, M.J.S., Drijvers, P.H.M. & Bakker, A. (2019). Repeated Sampling in a Digital Environment: a Remix of Data and Chance. In Jankvist, U. T., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education (CERME11, February 6 – 10, 2019)*. Utrecht, the Netherlands: Freudenthal Group & Freudenthal Institute, Utrecht University and ERME.
- Van Dijke-Droogers, M.J.S., Drijvers, P.H.M. & Bakker, A. (2019). Een digitale remix van data en kans. *Euclides*, 94(5), 16-17.
- Van Dijke-Droogers, M.J.S., Drijvers, P.H.M. & Bakker, A. (2018). From sample to population - A hypothetical learning trajectory for informal statistical inference. In Verônica Gitirana, Takeshi Miyakawa, Maryna Rafalska, Sophie Soury-Lavergne & Luc Trouche (Eds.), *Proceedings of the Re(s)ources 2018 International conference* (pp. 348-351). Lyon: École Normale Supérieure de Lyon.
- Van Dijke-Droogers, M.J.S., Drijvers, P.H.M., & Bakker, A. (2018). Repeated Sampling as a step towards Informal Statistical Inference. In M.A Sotos, A. White & L. Guyot (Eds.), Looking back, looking forward. *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 8-13, 2018)*. Voorburg: International Statistics Institute.
- Van Dijke-Droogers, M., Drijvers, P., & Tolboom, J. (2017). Enhancing statistical literacy. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10, February 1–5, 2017)* (pp. 860–867). Dublin, Ireland: DCU Institute of Education and ERME.
- Van Dijke-Droogers, M.J.S., Drijvers, P.H.M. & Tolboom, J. (2017). Statistical literacy.... hoe dan? Onderzoekend en nieuwsgierig omgaan met data lijkt essentieel. *Euclides*, 92(5), 7-11.

Presentations related to this thesis

- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2019). *Van Steekproef naar Populatie*. Poster presented at Onderwijs meets Onderzoek 2018, Utrecht: Nederland, October 11.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2019). *Statistical Modeling Processes through the Lens of Instrumental Genesis*. Paper presented at Eleventh International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL), Los Angeles, California, United States of America. July 14-20.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2019). *Repeated Sampling in a Digital Environment: a Remix of Data and Chance*. Paper presented at Eleventh Congress of the European Society for Research in Mathematics Education (CERME), Utrecht, Nederland, February 6-10.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2019). *Viewing Statistical Modeling Processes through the Theoretical Lens of Instrumental Genesis*. Paper presented at the Pre CERME Conference (Small conference with 8 researchers from Israël, New-Zeeland and the Netherlands), Utrecht: the Netherlands. February 3–6.
- Van Dijke-Droogers, M. J. S. (2019). *Symposium Wiskundedocenten in Onderzoek*. Presentation at the Nationale Wiskunde Dagen (25th Edition), Veldhoven, the Netherlands, February 1–2.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2018). *Herhaalde Steekproeven met een Black Box*. Poster presented at Onderwijs meets Onderzoek 2018, Utrecht, Nederland, October 8.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2018). *Repeated Sampling as a step towards Informal Statistical Inference*. Paper presented at the International Conference on Teaching Statistics (ICOTS), Kyoto, Japan, July 8-13.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2018). *From sample to population - A hypothetical learning trajectory for informal statistical inference*. Paper presented at Re(s)ources: Understanding Teachers' Work through their Interactions with Resources for Teaching. Lyon, May 28–30.
- Van Dijke-Droogers, M. J. S. (2018). *Onderwijs en Onderzoek: Van Steekproef naar Populatie*. Presentation at the Wiskunde Dialoog: Studiedag voor wiskundedocenten, Nijmegen, the Netherlands, April 10.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2018). *Van Steekproef naar Populatie*. Presentation at the Nationale Wiskunde Dagen (24th Edition), Noordwijkerhout, the Netherlands, February 2–3.

- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Bakker, A. (2017). *Steekproef en Populatie*. Poster presented at Onderwijs meets Onderzoek 2017, Utrecht, Nederland, June 12.
- Van Dijke-Droogers, M. J. S., Drijvers, P. H. M., & Tolboom, J. (2017). *Enhancing Statistical Literacy*. Paper presented at the Tenth Congress of the European Society for Research in Mathematics Education, Dublin, Ireland, February 1–5.

FI Scientific Library

(formerly published as CD-b Scientific Library)

108. Wijnker, F. (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. 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 Hằng (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 VIIIth 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 getalengebied 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 voorbeeldsmateriaal 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. Vogelesang, 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.*



The increasing amount of data in media over the last year—think of COVID—illustrates the necessity for students to become statistically literate—including interpreting inferences. Drawing inferences involves making data-based claims under uncertainty when only partial data are available. However, inferences are challenging for students in Grade 10 and higher. This thesis focused on the question: How can a theoretically and empirically based learning trajectory introduce 9th-grade students to statistical inference? To answer this question, we used a design-based research approach, complemented with a case study into learning statistics from and with technology. The design of the trajectory was informed by theories on repeated sampling and statistical modeling using a black box paradigmatic context. The learning trajectory was implemented in teaching practice during three interventions. A pre- and posttest were designed to evaluate the trajectory's effects in the large-scale final cycle. A national and international comparison of student results showed that students who took part in the learning trajectory ($N = 267$) scored significantly higher on statistical literacy than the comparison group that followed the regular curriculum ($N = 217$), in particular, on the domain of statistical inference. We also observed positive effects on other domains of statistical literacy. These findings suggest that current statistics curricula for grades 6–9, usually with a strong descriptive focus, can be enriched with an inferential focus—at least for pre-university education (VWO). The benefit of this early introduction is that students learn more about inference and not less about the other domains of statistical literacy, to anticipate for subsequent steps in students' statistics education.