International Journal of Science and Mathematics Education (2022) 20:1743–1766
https://doi.org/10.1007/s10763-021-10208-8

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

Marianne van Dijke-Droogers[1] · Paul Drijvers[1] · Arthur Bakker[1]

Received: 22 January 2021 / Accepted: 30 July 2021 / Published online: 13 November 2022

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

✉ Marianne van Dijke-Droogers
m.j.s.vandijke-droogers@uu.nl

[1] Freudenthal Institute, Utrecht University, PO Box 85.170, 3508 AD Utrecht, The Netherlands

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 et al., 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.[1] 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.

[1] https://nzcurriculum.tki.org.nz/The-New-Zealand-Curriculum/Mathematics-and-statistics/Achievement-objectives

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 et al., 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 toward 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 et al., 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:

1. 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?
2. How do the designed LT steps foster students' learning processes?

Methods

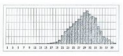
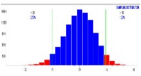
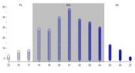
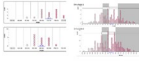
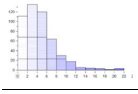
The designed LT aims at introducing students to key concepts of statistical inference by using theories of informal statistical inference. We first outline the design of the LT. We incorporated two main ideas: repeated sampling with a black box and statistical modeling with a digital tool. Second, we 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 min each. In the last step of sequences I and II, three different real-life contexts were offered with a time duration of 45 min 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 1. A more detailed description can be found in the Supplementary Material A.

Table 1 Overview of steps 1–8 of the learning trajectory

LT Step	Description	Example of activities	Learning Goal	Construction of LT steps
Categorical data				
1. Experimenting with physical black box 	Physical black box *with marbles* experiment (with small and large viewing window)	Estimate the number of yellow marbles in a black box filled with 1,000 marbles (*balletjes* 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				
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
7. Modeling a black box 	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	Use TinkerPlots to determine most common sample results—and extraordinary high/low results—from the (given) modeled black box of step 5	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]	From step 7 emerges the question of how to apply statistical modeling with numerical data in other contexts and situations
8. Modeling real-life contexts 	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	Use TinkerPlots to simulate repeated samples (size 200) from a hidden dataset of 4,000 students to determine the time students spent on sport	Students use statistical modeling within the digital environment of TinkerPlots to make inferences, within the context of a real-life problem	

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 toward 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 et al., 2021). 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.

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-h 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. 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 ($M_{new} = 2.49$, $SD_{new} = 0.71$, $M_{ex} = 2.78$, $SD_{ex} = 1.38$, $n = 50$, $t(49) = -1.6$; $p = 0.11$). 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 0.81, 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 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 and 5, steps 2 and 6, steps 3 and 7, steps 4 and 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. Figs. 2 and 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

Table 2 Defined 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

until agreement was reached. Adjustments in the coding were also applied to the rest of the data.

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 3 displays the indicators and corresponding learning activities on students' worksheet, for each LT step (see Table 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 steps 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.

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, $F(1, 482) = 75.0$, $p < 0.0005$). 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. Although we conjectured a similar pretest score for both groups, the results showed that the

Table 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	a. 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	b. 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	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)

Table 3 (continued)

LT step	Indicators	Task description per indicator (Worksheet Task)
7	Using statistical modeling in TinkerPlots (*given* model) for	
	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 *given* model for varying sample sizes (W7.1; W7.8)
	b. Interpreting effect of sample size on expected population distribution	b. Students mention that a larger sample size better reflects the population distribution (7.15)
	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 h per week on sports, using statistical modeling with a *hidden* model of the population (size 4,000) and sample size 500 (W8.5; W8.6)

Table 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[a]	2.45 (0.65)	2.72 (0.71)	−0.27***
	Steps 1 and 5	2.10 (1.34)	2.43 (1.41)	−0.33**
	Steps 2 and 6	2.54 (0.91)	2.77 (0.96)	−0.23 **
	Steps 3 and 7	2.48 (0.68)	2.75 (0.66)	−0.27***
	Steps 4 and 8	2.62 (0.94)	2.83 (0.92)	−0.21*
Posttest	SI[a]	3.34 (0.84)	2.67 (0.84)	+0.67***
	Steps 1 and 5	3.52 (1.26)	2.94 (1.26)	+0.58***
	Steps 2 and 6	3.44 (1.31)	2.84 (1.42)	+0.60***
	Steps 3 and 7	2.39 (1.04)	1.85 (0.97)	+0.54***
	Steps 4 and 8	3.65 (0.97)	2.91 (1.00)	+0.74***
Progress Pre to post	SI[a]	+0.89 (0.92)***	−0.04 (0.71)	+0.93***
	Step 1 and 5	+1.42 (1.71)***	+0.52 (1.57) ***	+0.90***
	Steps 2 and 6	+0.91 (1.50) ***	+0.06 (1.48)	+0.85***
	Steps 3 and 7	−0.09 (1.15)	−0.89 (1.00) ***	+0.80***
	Steps 4 and 8	+1.04 (1.18)***	+0.08 (1.05)	+0.96***

***$p<.0005$; **$p<.005$; *$p<.05$

[a]Main results for SI (Van Dijke-Droogers et al., submitted)

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.

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 3 (see Fig. 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 (χ^2 (2) = 80.84, $p<0.0005$). 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 toward statistical inference.

Second, we present the results on posttest item 8, a newly designed item (see Fig. 2). Most students from both groups noted one specific value as their

Mrs. Jones wants to buy a new car, either a Honda or Toyota. She wants whichever car will breakdown 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%

Fig. 1 Students' achievements on posttest item 3, taken from Watson and Callingham (2004)

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 7, 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 (χ^2 (2) = 6.57, $p < 0.05$).

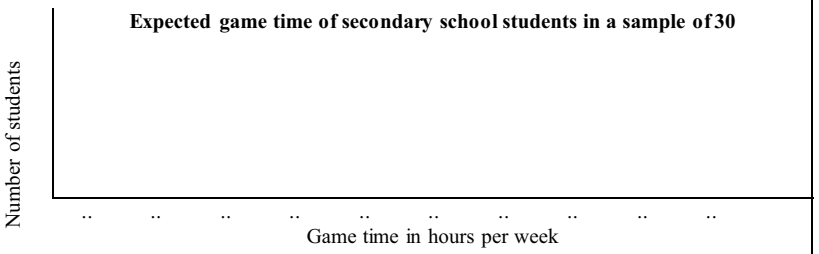
Third, we regard the results for posttest item 9, a newly designed posttest item related to items 7 and 8 (see Figs. 2 and 3). Most students in the comparison group (63.9% for levels 0–1) focused on the context, without referring to the data from their sketched graph in posttest item 7 or their average in item 8, 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 (χ^2 (5) = 28.19, $p < 0.0005$). As such, the results for items 8 and 9 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 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, which provided us with

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.

7. Draw in the graph below the sample results you expect.



8. 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 7)
		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%

Fig. 2 Students' achievements on posttest item 8, newly designed item

9. Explain your answers to Items 7 and 8. ……

Level	Code	Description
5	5	Statement considering the effect of small sample sizes on ***variability*** to explain the shape of the graph and/or the *range of values* for the average at respectively Item 7 and Item 8; the statement is bedded in the ***context***
4	4	Statement considering ***variability*** to explain the shape of the graph and/or the *range of values* for the average at respectively Item 7 and Item 8; the statement is bedded in the ***context***
3	3	Statement considering ***variability*** to explain the shape/peak of the graph and/or the average (one value or a range) at respectively Item 7 and Item 8; the statement is bedded in the ***context***
2	2	Statement ***without variability*** to explain the shape/peak of the graph and/or the average (one value or a range) at respectively Item 7 and Item 8; the statement is bedded in the ***context***
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	Level 1	Level 0
Comparison group	1.19	0.4%	2.1%	13.0%	20.6%	28.2%	35.7%
Intervention group	1.97	0.4%	9.7%	20.6%	31.8%	30.3%	7.1%

Fig. 3 Students' achievements on posttest item 9, newly designed item

insight into how each of these LT steps fostered or hindered the students' learning process.

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 Fig. 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.

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.

Table 5 Overview of results for LT steps 1 to 8

LT step in Sequence I *Categorical data*	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 *Numerical data*	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 (*given* 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%

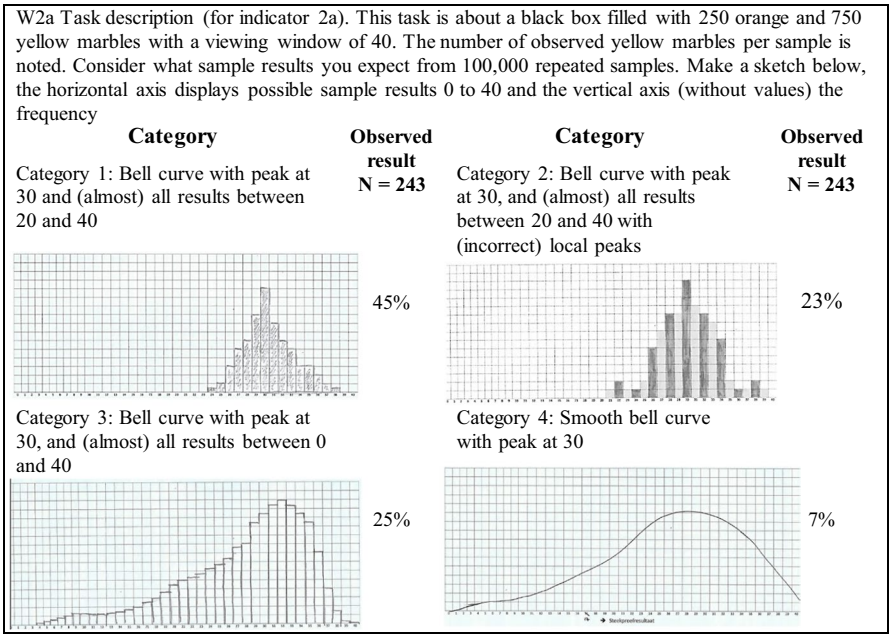


Fig. 4 Four types of correct student drawings ($N = 243$) of the expected results of repeated sampling (100,000 repetitions) with sample size 40 in a sampling distribution, with percent per type of elements to foster students' learning progress

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 2 and 3 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.

For indicators 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 Fig. 5).

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.

Specific Task	Category of answers	Examples of students' work	Observed result N = 267
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?	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%

Fig. 5 Percent of students per category of answers on worksheet 4 task 18

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 toward 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, and 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 Fig. 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.

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

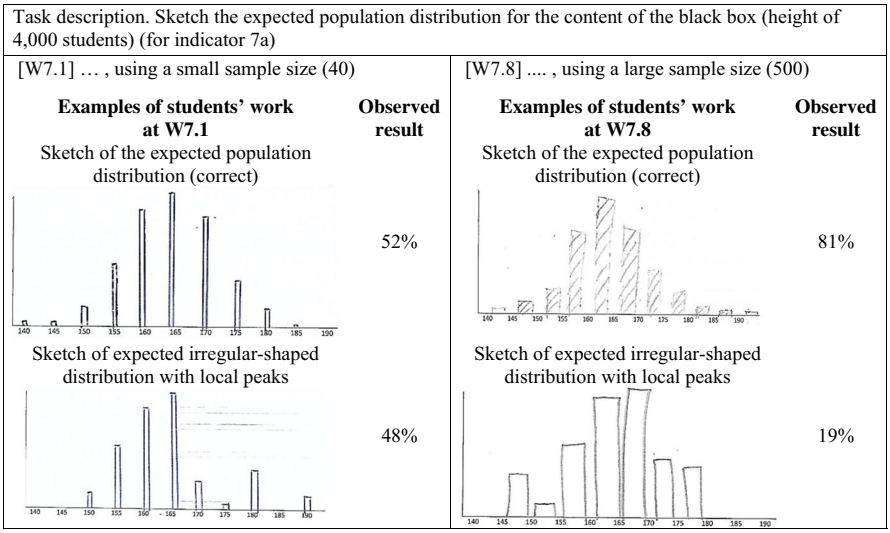


Fig. 6 Percent of students per category of answers on worksheet 7 tasks 1 and 8

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 was quite similar.

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, the 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 LTs 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 et al., 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 iteration 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 toward 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10763-021-10208-8.

References

Bakker, A. (2004). *Design research in statistics education*. Utrecht University.

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.

Biehler, R., & Ben-ZviMaker, D. 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). Springer.

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.

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). Kluwer Academic Publishers.

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). Springer.

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.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education, 47*(2), 123–182.

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Reidel.

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.

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). NCTM.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education, 2*(1), Article 1.

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.

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). 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). Catalyst Press.

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.

Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5–19.

Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257–270.

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.

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*. Kluwer.

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.

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. (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, 107*, 235–260.

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., & 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.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730.