



Effects of a Learning Trajectory for statistical inference on 9th-grade students' statistical literacy

Marianne van Dijke-Droogers¹ · Paul Drijvers¹ · Arthur Bakker²

Received: 16 September 2022 / Revised: 21 February 2024 / Accepted: 4 March 2024
© The Author(s) 2024

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 inferences 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 Statistical Inference 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 ($n=267$). The results indicated that students made significant progress on Statistical Literacy and on the ability to make inferences in particular, but also on the other domains of Statistical Literacy. To further interpret the 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 confirmed a significant positive effect on all 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

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.

✉ 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

² Faculty of Social and Behavioral Sciences, University of Amsterdam, 15776, Nieuwe Achtergracht 127, 1001 NG Amsterdam, Netherlands

Statistical Literacy (SL) 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 SL. More details about the design of the LT are elaborated by van Dijke-Droogers et al. (2021).

Currently, the typical Dutch pre-10th-grade curriculum is mainly focused on the descriptive SL domains. Adding new learning trajectories on top of regular curricula is difficult, e.g., in time and effort. In this regard, the purpose of the LT is to expand the 9th-grade curriculum with SI, the more complex domain of SL, 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 SL as offered in the regular curriculum. Assessment instruments with a specific focus on SI hardly exist for our age group. As such, 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' SL, and on the SI domain in particular.

Theoretical background

Domains of statistical literacy

Statistical Literacy (SL) concerns the use of statistical information as evidence in arguments (Schild, 1999). 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—evidence-based claims—and a reasoned process for probabilistic generalizations from data—interpreting the uncertainty involved (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; Papanistodemou & 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*TM 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*TM and use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real-world situation, and to draw inferences. The details of the LT will be illustrated and discussed later in the methodology section.

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 et al. (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 and Garfield (2018). 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. A study by Novak (2014) shares content and design with ours as it involves the evaluation of a simulation-based intervention using a pre-post research design.

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—in their organization of assessment items—between the three domains of SL. Their approach allowed us to identify students' SL, and also their performance on the domain of SI in particular. Using archived data from 1993 to 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 1. A follow-up study by Callingham and Watson (2017) showed that the

Table 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

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. 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.

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. Additionally, to further 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, and with level scores 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 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*TM. 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.


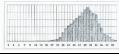
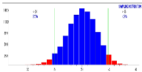
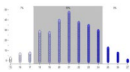


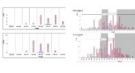
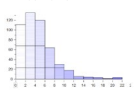
The investigative approach and learning activities in the more complex SI domain also attend to the other domains of SL. For example, the AC domain, average and chance, is addressed as students summarize their obtained sample data in measures of center and spread. As another example, the graphing part 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.

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) visualizing distributions, (3) statistical modeling using *TinkerPlots*TM, (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-min lessons, with a total of 12 lessons. More details about the design of the LT are elaborated by van Dijke-Droogers et al. (2021).

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. In line with Novak (2014), we chose a pre-post research design to measure the effects of the LT on students' proficiency, i.e., students' progress when working with the LT. For the design of the tests, following Ziegler and Garfield (2018), we used existing items from validated tests

Table 2 (van Dijke–Droogers et al., 2021) Overview of steps 1–8 of the designed 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 marbles 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	<p>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</p> <p>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</p> <p>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?</p>
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	
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 <i>TinkerPlots™</i> 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 <i>TinkerPlots™</i> 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]	
4. Modeling real-life contexts (ICT) 	Build and run a model of a real-life situation in <i>TinkerPlots™</i> 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 <i>TinkerPlots™</i> 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 <i>TinkerPlots™</i> to make inferences, within the context of a real-life problem	
<i>Numerical data</i>				
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 student's 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	<p>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?</p>
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. Students draw inferences about the population, considering distribution, mean, sample variability and probability	
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 <i>TinkerPlots™</i> 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 <i>TinkerPlots™</i> 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]	<p>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</p> <p>From Step 7 emerges the question of how to apply statistical modeling with numerical data in other contexts and situations</p>
8. Modeling real-life contexts 	Run a model of a real-life situation in <i>TinkerPlots™</i> and use this model, by simulating and interpreting the sampling distribution of repeated samples, to understand the real-life situation	Use <i>TinkerPlots™</i> 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 <i>TinkerPlots™</i> to make inferences, within the context of a real-life problem	

by Watson and Callingham (2003, 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 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 min. 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 3 provides an overview of the composition of the pre- and posttest, with reference to sources and accompanying domains of SL.

Figure 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 Fig. 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.

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 cluster 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. The level scores of these new items were, as with the existing items,

Table 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

SL Statistical Literacy, *AC* average and chance, *GV* graphing and variation, *SI* Statistical Inference

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	

Fig. 1 Item with corresponding level description from Watson and Callingham (2004, p. 138)

based on Watson and Callingham's (2003) level descriptions, and on the exemplary items they formulated on the SI domain (2004). See Fig. 2 for an example.

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 newly 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

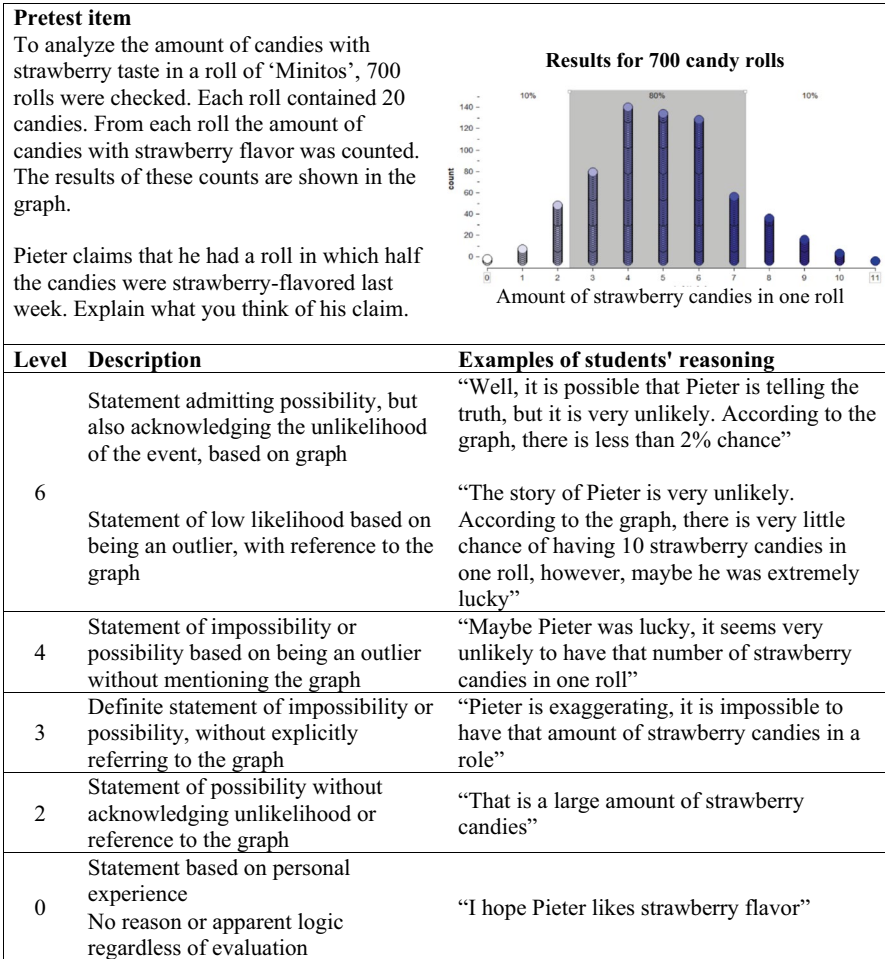


Fig. 2 Newly designed item with corresponding level description on the SI domain of Statistical Literacy (SI, Statistical Inference)

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 the same two classrooms with 25 students each. 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 posttest was piloted 4 weeks after the pretest pilots. The 25 students did not follow any statistics education in the intervening weeks. The results of the pre- and posttest pilot were non-significantly different ($-0.08, t(51) = 0.84, p = .40$). Additionally, the posttest was thoroughly examined to ensure the pre- and posttest were comparable.

Concerning the reliability of the tests, Cronbach's alpha values were 0.84 and 0.85 on the pre- and posttest, respectively, indicating a good reliability (Taber, 2018). 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 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 > 0.30 are indicated as good items, scores between 0.20 and 0.30 as medium, and scores < 0.20 as poor items (Ebel & Frisbie, 1991). The pretest Rit values indicated 5 poor, 12 moderate, and 22 good items, and the Rir scores indicated 8 poor, 16 moderate, and 15 good items. For the posttest, we observed moderately difficult items with 4 easy items and no difficult items. The Rit values indicated 1 poor, 9 moderate, and 24 good items, and the Rir scores indicated 2 poor, 13 moderate, and 19 good items. We considered these item scores on the pre- and posttest to be most acceptable. The pre- and posttest can be found in Appendix A and B.

Participants

For participants in 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 13 classes in 5 different schools. Two teachers participated with 2 of their classes. The teachers were instructed for the LT during 2 similar 3-h 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 eliminated all the regular 9th-grade statistics lessons to save time for the LT. The participating students from the intervention group were in the pre-university stream and thus belonged to the 15% best-performing students in our educational system.

Table 4 Reliability and item characteristics the pre- and posttest

	Pretest		Posttest	
	Average measure	Rating	Average measure	Rating
p value	0.54	Moderately difficult	0.62	Moderately difficult
Rit value	0.35	Good	0.42	Good
Rir value	0.30	Medium/good	0.36	Good
Cronbach's α	0.84	Good	0.85	Good

Due to practical constraints—that is, the number of participants in the intervention group—it was not possible to set a randomized trial with a control group. To be able to indicate the effect of the LT in comparison with the regular curriculum, we established a “Dutch baseline” from a comparison group. For the participants in the comparison group, through a national call, we invited teachers who were interested in administering a test to identify the SL of their students. Six teachers applied with a total of 217 students in 10 classrooms. When the comparison group was registered, all participants had recently completed the 9th-grade regular statistics education, consisting of 10–16 lessons. The regular curriculum focused on the AC and GV domains of SL, as described earlier in the section on the domains of SL. To identify these students’ SL, and also to compare well with the intervention group, the average results on two tests were administered, with a 4-week interval between them. The two tests consisted of items from the pre- and posttest of the intervention group. In retrospect, the results of the comparison group on the two tests were found to be similar, for both SL as a whole (-0.02 , $t(216)=0.4$, $p=.65$), and for the domains SI (-0.04 , $t(216)=0.9$, $p=.40$), GV ($+0.09$, $t(216)=1.4$, $p=.18$), and AC (-0.11 , $t(216)=1.1$, $p=.26$). This could be expected, since no statistics education was given during the intervening 4 weeks. The average results on both tests were used as Dutch baseline achievements for the SL of 9th graders.

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 for the pre-university stream 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 of the intervention group consisted of pre- and posttests. The pretest was taken in months 7–8 of the school year 2019–2020. The participating teachers administered the test, according to a clear instruction for testing, from their own students during their regular 45-min mathematics lessons. The posttest was administered in months 9–10 of the school year, after completing the LT, in a similar way, by the teachers during their regular lessons in their own school.

The data of the comparison group consisted of two tests, taken in months 8 and 9 of the school year. The average of both tests was used to identify the Dutch baseline. The tests were administered by the participating teachers, according to a clear instruction, during their regular mathematics lessons.

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. Second, we compared the scores of the intervention group with Dutch baseline achievements from the comparison group, and with the scores of Australian students (Callingham & Watson, 2017).

First, for assessing students' proficiency on the domains of SL, the pre- and posttest data for the intervention group 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, we compared changes in students' pre- and posttest scores. 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. 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 finding with a Dutch baseline. For this baseline, the average scores of the comparison group on two tests were used. The test data were coded with the level scores 0–6, in a similar way as for the intervention group. For significance, we used independent samples *t* tests for comparing the intervention group scores with the Dutch baseline. Additionally, 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 levels 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 significance, we used independent samples *t* tests and chi-squared tests for the distribution over 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 second 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.

Results

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

Students' level scores for SL

Table 5 displays students' proficiency on the domains of SL in level scores at the pre- and posttest for the intervention group, including their progress from pre to post.

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 ($+0.68$, $t(266)=13.0$, $p < .0005$). Students' results on SL confirmed our conjecture that following the LT had a clear positive effect on students' SL.

With regard to the SI domain of SL, a paired t test between the pre- and posttest indicated that students' average level score on the posttest was considerably higher than on the pretest ($+0.89$, $t(266)=15.8$, $p < .0005$). These results 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.

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 ($+0.52$, $t(266)=8.7$, $p < .0005$). 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.

Concerning students' progress on the AC domain, 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 ($+0.63$, $t(266)=6.7$, $p < .05$). 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.

Table 5 Students' mean level scores on the domains of SL at the pre- and posttest for the intervention group, including their progress from pre to post

	Intervention group ($n=267$)		
	Pretest	Posttest	Pre to post
	M (SD)	M (SD)	M(SD)
SL	2.60 (0.61)	3.28 (0.69)	+0.68 (0.86)**
SI	2.45 (0.65)	3.34 (0.84)	+0.89 (0.92)**
GV	2.07 (0.63)	2.59 (0.81)	+0.52 (0.98)**
AC	3.29 (1.38)	3.92 (0.88)	+0.63 (1.53)*

SL Statistical Literacy, SI, GV, and C are domains of SL SI sampling and inference, GV graphing and variation, AC, average and chance

* $p < 0.05$; ** $p < 0.0005$

Students' level scores in comparison with the Dutch baseline

Table 6 displays the Dutch baseline achievements from the comparison group on the domains of SL in level scores, including a comparison with the pre- and posttest for the intervention group.

When comparing the results for SL on the posttest, an independent samples t test between the intervention group and the baseline 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 (+0.32, $t(482)=4.9$, $p<.0005$). Students' results on SL confirmed our conjecture that following the LT had a clear positive effect on students' SL.

With regard to the SI domain of SL, on the posttest, an independent samples t test 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 (+0.65, $t(482)=8.7$, $p<.0005$). Concerning the GV domain of SL, an independent samples t test indicated that the posttest score for the intervention group was slightly, but significantly, higher than the Dutch baseline (+0.26, $t(482)=3.7$, $p<.05$). 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. For the AC domain, an independent samples t test indicated that the posttest score for the intervention group that followed the LT was comparable with the Dutch baseline achievements (+0.06, $t(482)=0.6$, $p=.52$). These findings confirmed our conjecture that the LT with a focus on SI also stimulated the other domains of SL.

When comparing the results for SL on the pretest, an independent samples t test indicated that the average level score for the intervention group on SL was significantly lower than the Dutch baseline (-0.36 , $t(482)=5.9$, $p<.0005$). With regard to the SI domain, on the pretest, the score for the intervention group was slightly, but significantly, lower than the Dutch baseline (-0.24 , $t(482)=3.7$, $p<.05$). We did not expect this lower score. Although the students of the comparison group, the Dutch baseline,

Table 6 Dutch baseline level scores on the domains of SL, including a comparison with the pre- and posttest for the intervention group

	Dutch baseline ($n=217$)	Pretest intervention minus Dutch baseline	Posttest intervention minus Dutch baseline
	M (SD)	M(I) – M(D)	M(I) – M(D)
SL	2.96 (0.73)	-0.36^{**}	$+0.32^{**}$
SI	2.69 (0.78)	-0.24^*	$+0.65^{**}$
GV	2.33 (0.73)	-0.26^*	$+0.26^*$
AC	3.86 (1.18)	-0.57^{**}	$+0.06$

SL Statistical Literacy, SI, GV, and AC are domains of SL SI sampling and inference, GV graphing and variation, AC average and chance

* $p<0.05$; ** $p<0.0005$

followed the regular statistics curriculum before the test, the SI domain was not offered in the regular lessons, so we expected a similar score for both groups. Regarding the GV domain, the score for the intervention group was, as expected, significantly lower than the Dutch baseline level score (-0.26 , $t(482)=4.2$, $p < .05$). 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, the score for the intervention group was, as expected, significantly lower than the baseline (-0.57 , $t(482)=4.8$, $p < .0005$). 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.

The lower scores on the pretest for the intervention group, in comparison with the Dutch baseline, were to be expected, as the intervention group did not have 9th-grade statistics lessons prior to the pretest. Furthermore, the lower level score of -0.36 on SL for the intervention group on the pretest relative to the Dutch baseline turned out to be almost equal in size to their higher level score of $+0.32$ on the posttest. Since the intervention group had an educational disadvantage of about one school year relative to the baseline at the pretest, their score on the posttest could be interpreted as almost one school year advantage. Students' results on SL confirmed our conjecture that following the LT had a clear positive effect on students' SL.

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 Dutch students over the levels of SL on the pre- and posttest is presented in Table 7 as well as the distribution of Australian students.

The comparison of students' distribution over the levels is displayed in Fig. 3. The first two graphs compare the pre- and a posttest scores for the intervention group with those of Australian students. The third graph compares the results of the Dutch

Table 7 Distribution over Levels of SL, for Dutch and Australian students

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Dutch students						
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%	-
Dutch baseline ($n=217$)	5.3%	18.9%	50.0%	25.4%	-	-
Australian students						
Grade 6 ($n=992$)	17.4%	24.1%	43.6%	14.7%	0.2%	-
Grade 7 ($n=1788$)	8.2%	14.2%	38.7%	36.4%	2.5%	-
Grade 8 ($n=2154$)	7.2%	11.0%	32.3%	41.6%	7.6%	0.4%
Grade 9 ($n=1054$)	4.3%	7.5%	24.9%	47.9%	14.6%	0.8%

baseline with those of Australian students. For the Dutch baseline, the results on the pre- and a posttest for the comparison group were aggregated, since these results were highly similar. First, from Fig. 3, the graphs illustrate that for the lower levels; the percentage of post-intervention students was lower than the percentage of the Dutch

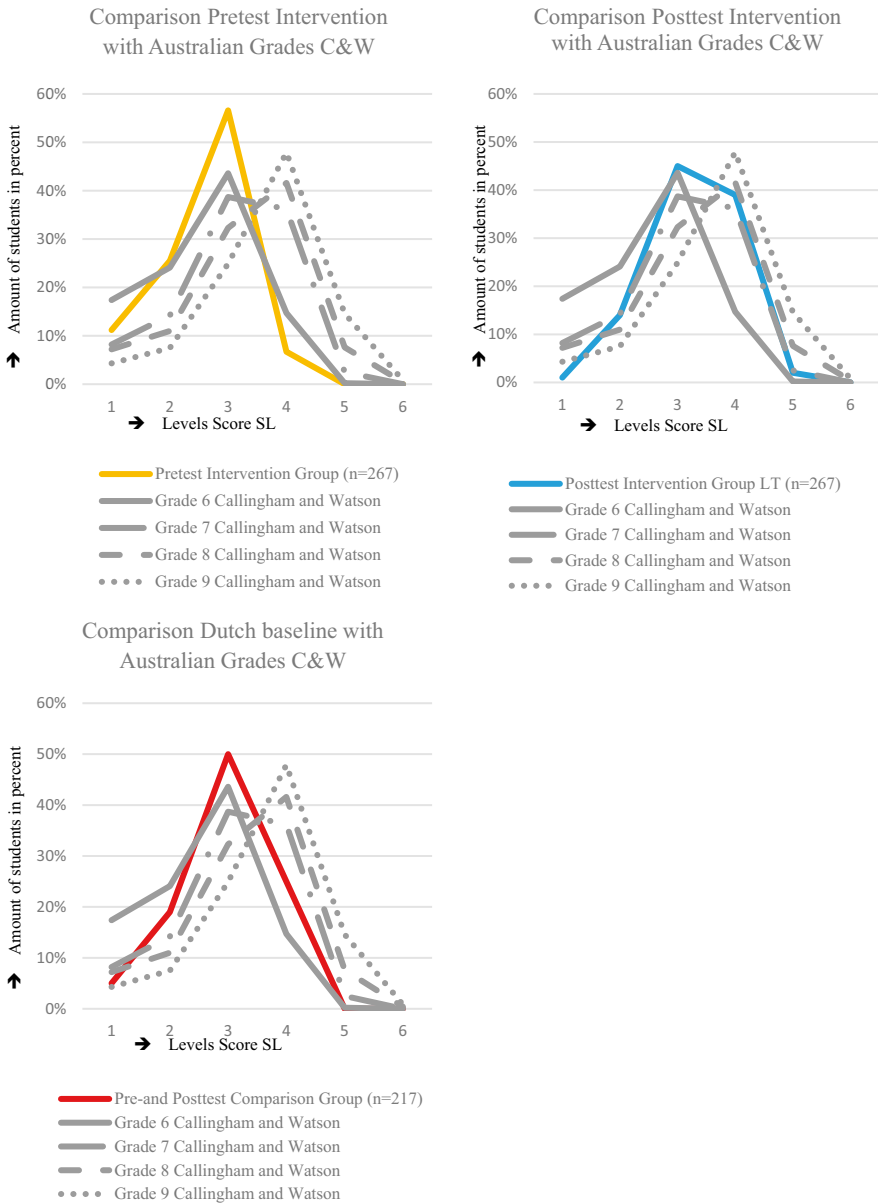


Fig. 3 Comparison of students' level scores on Statistical Literacy (SL) with Australian grade results from findings by Callingham and Watson (2017)

baseline or Australian students in grades 6, 7, and 8. Second, the graphs show that for the higher levels, the percentage of post-intervention students was actually higher. 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. 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 8 summarizes the comparison of the intervention group and the Dutch baseline 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). According to the findings by Callingham and Watson, the Dutch baseline 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 Dutch baseline 2.96 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 Dutch baseline score from 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 Dutch baseline 2.96 (0.69), the difference of 0.32 again corresponds to a level difference of more than one school year.

Table 8 Students' proficiency compared 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 and W
Pretest intervention group ($n=267$)	No 9th-grade statistics lessons	Grade 6
Posttest intervention group ($n=267$)	Learning Trajectory	Grade 7–8
Dutch baseline ($n=217$)	Regular 9th-grade curriculum	Grade 6–7

Conclusion and discussion

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, both among researchers and teachers (Ben-Zvi et al., 2018; Biehler et al., 2018). The aim of the study presented here was to evaluate the effects of a Learning Trajectory for Statistical Inference on 9th-grade students' Statistical Literacy, and on making inferences in particular. Theories of informal Statistical Inference complemented by ideas of growing samples and repeated sampling guided the design of the Learning Trajectory.

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, and analyzing data, to answer the question posed. This is consistent with previous studies and theories that advocate a holistic approach (Ainley et al., 2006; Franklin et al., 2007; Lehrer & English, 2017; Van Dijke-Droogers et al., 2017).

In discussing these conclusions, there are a few points to consider. The first involves the low level of proficiency of Dutch students on Statistical Literacy relative to Australian students (Callingham & Watson, 2017). We expected Dutch students to score at the posttest on grade 9 level, and not on grades 6–7 and grades 7–8, for the Dutch baseline and the intervention 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 Watson's research (<https://www.australiancurriculum.edu.au/>). Another issue in this respect is that the average maximum attainable score on the graphing and variation items on both tests was lower (about 3.7) than for the other domains (about 5.5), which negatively affected students' overall Statistical Literacy scores. When we compensate for the lower graphing and variation item scores, the Statistical Literacy average level scores of participating students increase by about 0.3. When we then compare the adjusted literacy scores with the Australian grade results, the grade results for

Dutch students increase with almost one school year and, as such, were closer to our expectations. “Regarding graphing and variation, a related issue is whether working by hand or with digital tools affects students’ learning. In our study, the intervention students mainly worked with digital tools, while the Dutch baseline students mainly worked by hand. The posttest scores for the intervention students on the GV domain were significantly higher than the scores for baseline students. As such, working with digital tools for graphing and variation seems to promote students’ understanding of the GV domain.

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 Learning Trajectory on students’ Statistical Literacy and on the Statistical Inference 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 other two domains.

Limitations of our study are the following. First, we worked with students from the pre-university stream, the 15% best performing students of our educational system, for both the intervention and the comparison group. As such, the results in this research are not generalizable to regular classrooms without further research. Second, due to practical constraints—that is, the number of participants in the intervention group—it was not possible to set a randomized control group. To be able to indicate the effect of the Learning Trajectory in comparison with the regular curriculum, we established a Dutch baseline from a comparison group. This comparison group had already completed the regular 9th-grade statistics curriculum. To determine the level of Statistical Literacy of this group, and to be able to directly compare their results with the intervention group, we administered two tests that consisted of items from the pre- and posttest of the intervention group. Although a classical randomized test with a control group has added value—for example, as it enables to determine the initial and final level for the regular 9th-grade curriculum—using the Dutch baseline helped us to indicate the effect of the Learning Trajectory. Third, the items of the pre- and posttest were not identical. Despite careful alignment—through posttest pilots and expert consultation—differences in context, question wording, and visualizations may affect the result. However, the results of the comparison group ($n=217$) on both tests, administered at an interval of only 4 weeks, were found to be non-significantly different, both on Statistical Literacy and on all three domains. This finding supports our assumption that both tests were comparable. Fourth, 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 Learning Trajectory. The development of a pre- and posttest, consisting of items from validated tests—mainly from Watson and Callingham—supplemented by equivalent newly designed items on Statistical Inference, enabled us to assess students' Statistical Literacy, and their Statistical Inference in particular. Both newly designed and existing test items were found appropriate, with a Cronbach's alpha greater than 0.84 on the pre- and posttest. In analyzing the results, the levels of Statistical Literacy 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 Learning Trajectory. Therefore, we recommend researchers and educators who intend to investigate the Statistical Literacy of secondary school students to use the levels by Watson and Callingham for assessing and evaluating students' results.

Second, for the participating teachers of the intervention group, implementing the Learning Trajectory required considerable effort. In our study, 11 teachers from five different schools were willing to invest in the trajectory. The load for teachers from the comparison group was limited to administering two tests, making it easier for teachers to participate. Using a Dutch baseline from a comparison group appeared 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 Learning Trajectory highly affected students' performance on Statistical Literacy and Statistical Inference, and we also indicated significant positive effects for the other domains. Although the Learning Trajectory was not focused on the latter two, the investigative approach and more complex learning activities for Statistical Inference as embedded in the trajectory appeared to have a positive effect here as well. These findings indicate that the Learning Trajectory can be used to expand the 9th-grade curriculum with the Statistical Inference domain, without neglecting the current educational goals on the other domains of Statistical Literacy.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s13394-024-00487-z>.

Acknowledgements We thank Walter Stevenhagen for his contribution to the design and implementation of the assessment tool.

Funding This research was funded by the Dutch Ministry of Education, Culture and Science under the Dudoc program.

Declarations

Ethical approval The study was conducted according to the FI Data Management Protocol. This contains guidelines for the data collection (e.g., informing participants, consent statements from participants (including parents for participants under 16)), for data storage (e.g., ensuring privacy, making backups), and the use of a secure system to store data. More information about the FI Data Management Protocol can be found at <https://www.uu.nl/sites/default/files/FI%20Data%20Management%20Protocol-dec2020.pdf>.

Informed consent Results only include data from participants who provided written consent via informed consent (for participants under 16 through their parents).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ainley, J., Pratt, D., & Hansen, A. (2006). Connecting engagement and focus in pedagogic task design. *British Educational Research Journal*, 32(1), 23–38.
- Bakker, A. (2004). *Design research in statistics education*. Utrecht University.
- 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.
- 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). Springer.
- 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.
- 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). Springer.
- 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). 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). 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.

- Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, 6(2), 81–89.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.
- 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.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.
- 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.
- 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. (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. *Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course*, 44(7), 883–898.
- 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.
- Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, 12(3), 217–230.
- Lehrer, R., & English, L. D. (2017). Introducing children to modeling variability. In D. Ben-Zvi, J. Garfield, & K. Makar (Eds.), *International handbook of research in statistics education* (pp. 229–260). Springer.
- 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. (2017). Students' emergent articulations of statistical models and modeling in making informal statistical inferences. *Statistics Education Research Journal*, 16(2), 116–143.
- 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.
- Papariotodemou, 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*, 50(7), 1197–1212.
- Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modelling to connect data, chance and context. *ZDM*, 50(7), 1113–1123.
- 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.
- 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, Milo (1999). Statistical Literacy: Thinking critically about statistics as evidence. *Of Significance*, 1(1).
- Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, 32(4), 450–466.
- 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.
- 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 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). Introducing statistical inference: Design of a theoretically and empirically based learning trajectory. *International Journal of Science and Mathematics Education*.
- van Streun, A., & van de Giessen, C. (2007). Een vernieuwd statistiekprogramma: Deel 1 [A renewed statistical program, Part 1]. *Euclides*, 82(5), 176–179.
- 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.
- Whitaker, D., Foti, S., & Jacobbe, T. (2015). The levels of conceptual understanding in statistics (LOCUS) project: Results of the pilot study. *Numeracy*, 8(2). <https://doi.org/10.5038/1936-4660.8.2.3>
- 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.
- 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.
- Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal*, 17(2), 161–178.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.