J Math Teacher Educ
https://doi.org/10.1007/s10857-018-9403-9

Pre-service primary school teachers' knowledge of informal statistical inference

Arjen de Vetten[1,2] · Judith Schoonenboom[3] · Ronald Keijzer[2,4] · Bert van Oers[1]

Abstract The ability to reason inferentially is increasingly important in today's society. It is hypothesized here that engaging primary school students in informal statistical reasoning (ISI), defined as making generalizations without the use of formal statistical tests, will help them acquire the foundations for inferential and statistical thinking. Teachers who engage students in ISI need to have good content knowledge of ISI (ISI-CK). However, little is known about the ISI-CK of primary education pre-service teachers. Therefore, the aim of this paper is to describe this knowledge by surveying 722 first-year pre-service teachers from seven teacher colleges across the Netherlands. The survey consisted of five tasks using open-ended questions and true/false statements. The descriptive analysis showed that most respondents understood that descriptive statistics that take the global shape of the distribution into account can be used as arguments within ISI. Although a majority agreed that random sampling is a valid sampling method, distributed sampling was the preferred strategy for selecting a sample. Moreover, when asked to make a generalization beyond the data, most pre-service teachers only described the data and did not appear to understand that a representative sample can be used to make inferences about a population. These findings suggest that it may be useful if statistics education for pre-service teachers places more emphasis on sampling and inference, thereby prompting pre-service teachers to engage in ISI.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10857-018-9403-9) contains supplementary material, which is available to authorized users.

✉ Arjen de Vetten
a.j.de.vetten@fsw.leidenuniv.nl

[1] Department Research and Theory in Education, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

[2] Academy for Teacher Education, University of Applied Sciences iPabo, Jan Tooropstraat 136, 1061 AD Amsterdam, The Netherlands

[3] Department of Education, University of Vienna, Sensengasse 3a, 1090 Vienna, Austria

[4] Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

Published online: 14 March 2018

Keywords Informal statistical inference · Informal inferential reasoning · Statistics education · Samples and sampling · Primary education · Initial teacher education

Introduction

The aim of this paper is to describe the knowledge that pre-service primary education teachers have of informal statistical inference (ISI). We define ISI as making generalizations based on sample data, expressed with uncertainty, and without the use of formal statistical tests (Ben-Zvi et al. 2015; Zieffler et al. 2008). Inferential reasoning (i.e., the process of drawing a conclusion based on evidence or reasoning), of which ISI is a form, has become more and more in demand in our technology-driven society. Its importance is likely to increase even further as, on the one hand, our societies become increasingly complex and change ever more quickly and, on the other hand, inferential reasoning is a skill not easily routinized (Liu and Grusky 2013). If children are to participate fully in society, both of today and of the future, education must play a crucial role, as one of its primary purposes is to help children become qualified citizens (Biesta 2009). Since inferential reasoning is an important skill in becoming qualified members of society, teachers have the task of fostering children's inferential reasoning ability.

One way to foster children's ability to reason inferentially is to introduce primary school students to ISI. This ISI is based on a reasoning process in which multiple statistical concepts are used as arguments to support an inference (Makar and Rubin 2009; Watson 2004). An example of this might be when someone evaluates a sample as being too small and possibly biased, with a high level of variance, and therefore expresses a high degree of uncertainty when making an inference. In this example, the concepts of sample size, bias, sampling methods, and variance are used as statistical arguments to support the inference.

Although in many countries statistical inference is not introduced until the final years of secondary education, researchers have recently started to explore the possibility of acquainting students with the concept in upper primary school and familiarizing them with the core concepts underlying ISI, such as centrality, variation, samples, and sampling (Ben-Zvi 2006; Makar 2016; Meletiou-Mavrotheris et al. 2014; Meletiou-Mavrotheris and Paparistodemou 2015; Paparistodemou and Meletiou-Mavrotheris 2008; Watson and Kelly 2005). Because ISI does not require the use of formal inferential concepts, such as hypothesis testing and probability distributions, it is assumed to be within the reach of primary school learners. Recent research evidence indeed suggests that ISI can be made accessible for primary school students (Ben-Zvi 2006; Ben-Zvi et al. 2015; Makar 2016; Meletiou-Mavrotheris and Paparistodemou 2015; Paparistodemou and Meletiou-Mavrotheris 2008).

It is hypothesized that an introduction to ISI in the upper grades of primary school has several potential benefits. First, since it has been shown that understanding the statistical ideas and concepts underpinning statistical inference is challenging (Shaughnessy et al. 1996), students require repeated exposure to these notions over an extended period of time (Watson and Moritz 2000), calling for an early introduction of ISI. Second, as ISI involves a reasoning process in which multiple statistical concepts are used as reasoning tools, when students learn statistics within the context of ISI, it is conjectured that they will gradually understand the processes involved in inferential reasoning (Bakker and Derry 2011; Makar 2016; Makar et al. 2011) and in statistical reasoning from a more general perspective (Zieffler et al. 2008).

However, if students are to be introduced to ISI in primary school, future teachers need to be well prepared to conduct this introduction (Batanero and Díaz 2010). This means that teachers need to possess a thorough knowledge of the content they teach (Hill et al. 2008), which must go beyond what their students will actually learn (Ball et al. 2008) since the content knowledge of teachers will impact the learning achievements of their students (Rivkin et al. 2005). This preparation of teachers will also facilitate the development of their pedagogical content knowledge (Ball et al. 2008; Shulman 1986), which has been shown to be very relevant to ISI (Burgess 2009; Leavy 2010).

To conceptualize the content knowledge of ISI (ISI-CK) that pre-service teachers need to acquire, we used the general ISI framework of Makar and Rubin (2009). For this study among pre-service teachers, we conceptualized the components of the framework as follows:

1. "Data as evidence": We subdivided this into two subcomponents:
    a. "Using data": The inference is based on available data and not on tradition, personal beliefs or personal experience.
    b. "Describing data": Before the data can be used as evidence within ISI, one first needs to descriptively analyze the sample data, for example by calculating the mean (Zieffler et al. 2008). The resulting descriptive statistic then functions as an evidence-based argument within ISI (Ben-Zvi 2006). We distinguish between two types of descriptive statistics. First, descriptive statistics which determination requires to take all values of the distribution into account, such as the mean, majority ["modal clump" (Konold et al. 2002)] and spread. Second, descriptive statistics that do not require to take all values of the distribution into account, such as the mode, the minimum, and the range (see Sampling variability).
2. "Generalization beyond the data": The inference goes beyond a description of the sample data to make a probabilistic claim about a situation beyond the sample data.
3. "Uncertainty in inferences": We subdivided this component into four subcomponents:
    a. "Sampling method": The inference includes a discussion of the sampling method and its implications for the representativeness of the sample.
    b. "Sample size": The inference includes a discussion of the sample size and its implications for the representativeness of the sample.
    c. "Sampling variability": The inference is based on an understanding of sampling variability; that is, the understanding that the outcomes of representative samples are similar and therefore a particular sample can be used for an inference (Saldanha and Thompson 2007). Moreover, the inference uses aspects of the sample distribution that are relatively stable, such as the mean and majority. These stable aspects can function as signals for the population distribution (Konold and Pollatsek 2002).
    d. "Uncertainty language": The inference is expressed with uncertainty and includes a discussion of what the sample characteristics, such as the sampling method employed and the sample size, imply for the certainty of the inference.

The knowledge required can be most effectively appropriated when it takes pre-service teachers' pre-existing knowledge into account (Darling-Hammond et al. 2008). However, current research provides only scant evidence of (pre-service) teachers' ISI-CK (Ben-Zvi et al. 2015). The (sub-) components using data and generalization beyond the data have hitherto not been investigated. The evidence on the uncertainty in inferences components

suggests that many pre-service teachers show a limited understanding of sampling methods, sample size, representativeness, and sources of bias in the case of self-selection (Groth and Bergner 2005; Meletiou-Mavrotheris et al. 2014; Watson 2001). Concerning the sampling variability subcomponent, Mooney et al. (2014) reported that a substantial proportion of the subjects they examined understood that sample distributions are likely to be different from the population distribution. However, an understanding of the implications of variability for the actual sample proved to be more difficult. Watson and Callingham (2013) found that only half of the teachers they interviewed could conceptualize that a smaller sample has larger variability. No studies have so far been conducted on how (pre-service) teachers use descriptive statistics in the context of ISI. However, the literature on (pre-service) teachers' understanding of descriptive statistics in general has shown this understanding to be generally superficial (Batanero and Díaz 2010; Chatzivasileiou et al. 2011; Garfield and Ben-Zvi 2007; Jacobbe and Carvalho 2011; Koleza and Kontogianni 2016). More specifically, only a minority of such pre-service teachers attend to both center and spread when comparing data sets (Canada and Ciancetta 2007); while the group's understanding of the mean, median and mode is mostly procedural (Groth and Bergner 2006; Jacobbe and Carvalho 2011).

The few small-scale studies on the uncertainty in inferences and describing data (sub-) components indicate weak or superficial knowledge, while it is unknown what knowledge pre-service teachers have of the (sub-) components generalization beyond the data and using data. Moreover, since not all components have been investigated, pre-service teachers' knowledge of all the components cannot be described and compared in relation to each other. However, societal trends emphasize the importance of future teachers being equipped to introduce children to inferential reasoning. Therefore, the aim of the current study is to describe the ISI-CK of first-year pre-service teachers. This description can be used to design teacher college education that will improve teachers' ISI-CK. To this end, we report on the findings of a study of 722 first-year pre-service teachers from the Netherlands who completed five tasks made up of open-ended questions and true/false statements. The research question addressed in this paper is: To what extent do first-year pre-service primary school teachers have appropriate content knowledge of informal statistical inference?

Method

Context

In the Netherlands, the current curricula for statistics education in primary and secondary education do not include ISI. Actual teaching practices focus primarily on statistical procedures and graphing skills, where concepts are learned without reference to the need to collect and analyze data (Meijerink 2009). When statistical inference does form part of the secondary education curriculum, the ideas of sample and population are often only dealt with on a technical level.

The target population for this study was first-year pre-service primary school teachers enrolled in a full-time study program. In contrast to many other countries where students can only opt for teacher education after the completion of a bachelor's degree, in the Netherlands, initial teacher education starts immediately after secondary school and leads to the attainment of such a degree. For the students attending one of the 45 teacher

colleges, mathematics teaching seldom seems to be their main motive for becoming teachers (Blom and Keijzer 1997).

Respondents

The sampling proceeded in two steps: First, the teacher colleges were selected. These were recruited via personal contacts and the Dutch network of mathematics teacher instructors. The seven participating colleges were diverse in terms of size and location. Second, the participating teacher educators asked their first-year students to participate in the study. In total, 826 pre-service teachers took the test. Respondents were asked to provide their informed consent. While all of them took the test (as described below) and received feedback, 93 (11%) invoked the option of having their results excluded from the analysis. At the first teacher college where data were collected, a relatively large number of the respondents made use of this prerogative, probably because the information provided at that time was not sufficiently specific. After making the information more specific, fewer pre-service teachers opted out. Eleven respondents who did not complete the test had their data removed, leaving a total of 722 participating pre-service teachers. The procedure was approved by the ethical board of the Faculty of Behavioral and Movement Sciences of Vrije Universiteit Amsterdam (Scientific and Ethical Review Board of the Faculty of Behavourial Science of Vrije Universiteit Amsterdam 2016).

The average age of the respondents was 18.43 years (SD: 2.45), 26% were male, 43% had a background in secondary vocational education (students attending this type of course are typically aged between 16 and 20), 52% came from senior general secondary education, 3% had been enrolled in university preparatory education, and the educational background of the remaining 2% was either entirely something else or unknown. Their average score on the obligatory first-year mathematics examination for Dutch pre-service teachers was 103.31 out of 200 possible points (SD: 24.21). A score of 103 equals the 80th percentile of Grade 6 primary school students in the Netherlands.

Instrument

The instrument by which the pre-service teachers' ISI-CK was measured was a digital test consisting of five tasks combining open-ended questions and true/false statements, which collectively encompassed the three ISI components. We describe the rationale and design of the instrument below.

Rationale and design of the instrument

Three requirements guided the selection of the instrument. First, as we were striving for a representative sample of first-year pre-service primary school teachers, the instrument had to accommodate the analysis of a large number of responses. Second, we wanted to allow the respondents to give their own answers without steering them into a particular direction. Third, since we suspected that, without probing, the pre-service teachers would not reveal all their ISI-CK, the instrument should allow probing for information that was not provided through the initial questioning. No instrument found in the literature on statistics education satisfied these requirements. They either contained open-ended tasks only (e.g., Watson 2001) or focused on more advanced statistical reasoning (delMas et al. 2007; Haller and

Krauss 2002). Moreover, we did not find any instruments that covered all the components of ISI.

The requirements for both probing and large-scale administration were achieved by designing an instrument consisting of five tasks each of which combined an open-ended question with a number of true/false statements (see Online Resource for the instrument). In total, 27 statements were presented. For each task, the respondents answered an open-ended question by typing in their answer with an explanation. Next, the responses of fictional pre-service teachers to the same question were shown. These responses were an explanation, an argument or a method for solving the task. Some were complete responses, while others were only a fragment of a full response. Respondents were asked to evaluate the correctness of these fictional answers. It has been argued that true/false items can be used to assess complex issues (Burton 2005; Ebel 1972). The true/false statements jointly embodied all relevant aspects of the three components of ISI. The instrument was designed by the authors who have a mixture of backgrounds, including research and teaching expertise in statistics education, mathematics education, and mixed-methods and general pedagogy. So, by using open-ended questions to elicit respondents' own responses, by subdividing complete and complex inferential reasoning into small parts, and by probing all relevant aspects of ISI, we were able to investigate the level of ISI-CK in great detail, thereby strengthening the content validity. The instrument's validity was further strengthened by triangulating the results of the open-ended responses with the evaluations of the true/false statements.

The instrument's reliability was checked by calculating the proportion of inconsistencies between the open-ended responses and the statement evaluation. Although these two question types are not pure parallel forms, a low percentage of inconsistencies provides some evidence of sufficient inter-method reliability. An inconsistency would be present if a respondent first suggests a strategy in an open-ended response, for example suggests to use random sampling, and subsequently, when the same strategy is presented in a statement, the respondent evaluates the strategy as incorrect. Because in the open-ended response a respondent is not likely to come up with all possible strategies herself, the opposite response pattern (i.e., not proposing a particular strategy, but subsequently evaluating that strategy as correct) is not illogical and therefore did not count as an inconsistency. Approximately 13% of the cases were inconsistent. This low percentage provides some evidence of sufficient inter-method reliability.

Description of the tasks

In order to design the tasks, the level of ISI-CK relevant for pre-service teachers was first established by using curriculum guidelines, recommendations from the literature, and our own experience. Reasoning considered to be within the reach of primary school students includes the reasoning underlying the following (sub-)components: Using data, generalization beyond the data, sampling methods, sample size, and sampling variability (Ben-Zvi 2006; Paparistodemou and Meletiou-Mavrotheris 2008). It also includes knowledge of the suitability of the mean, majority and spread as evidence-based arguments for ISI (Franklin et al. 2007; Meijerink 2009), and reasoning about ISI within the context of a dot plot (SLO 2008). As stated, in order to teach ISI, teachers need knowledge of aspects of it that transcends the understanding of their students. This includes knowledge of appropriate sample size and sampling methods, a thorough grasp of sampling variability (Saldanha and Thompson 2007), an understanding of which descriptive statistics are sufficient arguments within ISI, and reasoning about ISI within the context of a scatter plot.

In all five tasks, the context of primary school students' enjoyment of math was used. We believed this to be a familiar context for the pre-service teachers, either from their personal experience or from their teaching internships. An example showing an abbreviated task and a true/false statement is shown in Fig. 1.

In Task 1, respondents were asked whether they agreed with senior pre-service teachers who based their conclusions on a large research project they had conducted, or with their fellow classmates who formed their conclusions on the basis of their personal experiences. The task thus investigated whether the pre-service teachers relied on research data to underpin their conclusions. Furthermore, one statement operationalized an aspect of the component generalization beyond the data.

In Task 2, respondents were asked how they would select a representative sample. The task thus investigated the respondents' knowledge of sampling methods and sample size. In addition, one statement covered an aspect of sampling variability.

In Task 3, inspired by Bakker (2004), respondents were shown a dot plot with data for 20 boys and were asked to predict the shape of the plot when the sample was enlarged to 40 boys. The pre-service teachers had to identify stable aspects of the distribution that could function as signals for the larger sample. The task operationalized the subcomponent sampling variability.

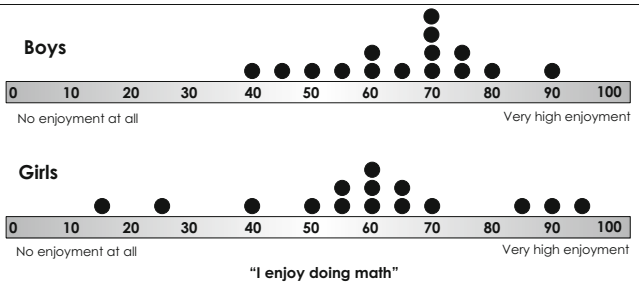
Tasks 4 and 5 measured the extent to which the respondents were able to use appropriate descriptive statistics as arguments within ISI. In Task 4, inspired by Zieffler et al. (2008), respondents were asked to compare samples of 15 boys and 15 girls and to make a generalization about whether the genders, in general, differed in their enjoyment of math. The task was designed to check whether the pre-service teachers understood which descriptive statistics can be used as evidence-based arguments within ISI in the context of comparing two dot plots (SLO 2008). In addition, a statement on the small sample size used operationalized an aspect of the subcomponent sample size, while one statement operationalized an aspect of the component generalization beyond the data.

Task 5, which was adapted from Cobb et al. (2003), presented the respondents with a scatterplot and requested them to predict the score of an individual with a given x-value. The pre-service teachers were asked to evaluate the suitability of a number of descriptive statistics as evidence-based arguments. One statement operationalized an aspect of the component generalization beyond the data.

Procedure

Cognitive interviews (Willis 2004) with three pre-service teachers and one teacher educator were conducted in order to assess whether the format of the test was easily understood, whether the true/false statements were interpreted as intended, and whether the list of statements was exhaustive. After each interview, the test was adjusted. The interviewees understood the format of the test with no issues; during the third and fourth interview, all statements were interpreted as expected. Next, the instrument was tested in one class of pre-service teachers in a similar environment to the final test setting, which led to some minor adjustments to the test.

The test was administered during the respondents' second or third month of study at their teacher college for the purpose of ensuring comparability across institutions. The teacher educators who administered the test received detailed instructions in order to guarantee comparable test circumstances. The pre-service teachers were randomly assigned to one of two versions of the test. Each version consisted of three of the five tasks. In this way, test completion time was acceptable, while the test still covered all aspects of



Based on the information about these 15 boys and 15 girls, answer the following question about boys and girls in general in grades 3 to 6 in the Netherlands: In general, who enjoy doing math more: boys or girls? Or is there no difference? Explain your answer.

[Answer field]

You will now read statements from other students who have also answered this question. Indicate for each statement whether you find their reasoning correct.

4.2 Max: "In general, boys enjoy math slightly more than girls, because only two boys were negative about math (with scores lower than 50), while there were three girls with a negative attitude."

- ❍ Correct reasoning
- ❍ Incorrect reasoning

Fig. 1 Task 4 is presented without its introductory text, and including one true/false statement. The pre-service teachers were asked to provide an open-ended response and to evaluate the correctness of Max' reasoning. See the Online resource for the full test

ISI. The Version A consisted of Tasks 1, 2, and 5 and was completed by 365 respondents; while Version B consisted of Tasks 1, 3, and 4 and was completed by 357 respondents. So Task 1 was completed by all 722 respondents.

Data analysis

Two true/false statements (1.2 and 2.8) were excluded from the analysis, as upon further reflection we concluded that the formulation of the statements might have been confusing. To check for the relationship between ISI-CK and the background variables educational background, mathematical content knowledge, sex, and age, linear regression analyses were performed for the true/false statements with the number of correctly evaluated statements per respondent as dependent variable, with separate analyses for Version A and Version B. The data were analyzed using Atlas.ti, Excel and R (R Core Team 2014).

The analysis of the open-ended questions was, for practical reasons, performed on a random sample of 100 open-ended responses per task. The coding of these responses, conducted by the first author of this paper and discussed with the other authors, consisted of two cycles. The first cycle consisted of a thematic analysis during which the open-ended responses were labeled with extended codes that captured the meaning of the particular response (Saldaña 2015). Codes were categorized into the components of ISI. The second round consisted of merging codes with similar meanings or covering closely related issues, for example, codes relating to the background variables of schools or children to be controlled for in sampling were merged into one code "quota factor school or child." The

resulting codes were used to summarize the main strategies and arguments used in the open-ended responses.

The analysis of the true/false statements was performed on all data, so it was on 722 respondents for Task 1, on 365 respondents for Tasks 2 and 5, and on 357 respondents for Tasks 3 and 4. For each statement, the percentage of respondents who evaluated the statement correctly was calculated.

The results of the open-ended responses and the true/false statements were combined to describe the ICI-CK for each component. Combining both data sources yielded a comprehensive picture both of active knowledge, as revealed by the open-ended responses, and of passive knowledge, as revealed by the true/false statements. It also hinted at potential explanations for why particular responses were given. Finally, it showed the extent to which there were strategies with which respondents agreed, but which they did not propose themselves in the open-ended responses.

Results

Relationship between ISI-CK and background variables

The regression analyses to check for the relationship between the results for the true/false statements and the background variables showed that the number of correctly evaluated statements was not significantly related to the variables educational background, mathematical content knowledge, sex, or age, $F(5, 188) = 1.05$, $p = .39$ for Version A, and $F(5, 185) = 1.60$, $p = .16$ for Version B (Table 1). The regression analyses satisfied the assumptions of the normality, independence and linearity of the residuals. The only effect found was for respondents with university preparatory education who scored significantly higher on ISI-CK than respondents with vocational education, $t(187) = 2.02$, $p = .045$ for Version A, and $t(184) = 2.26$, $p = .025$ for Version B, and higher than respondents with

Table 1 Results of regression analyses between ISI-CK and background variables

Variable	Number of correctly evaluated statements	
	Version A[a] B (se)	Version B[b] B (se)
Constant	8.62 (2.14)	10.91 (2.21)
Mathematical content knowledge	0.00 (0.01)	0.00 (0.01)
Educational background		
Senior general secondary education	0.29 (0.39)	0.02 (0.36)
University preparatory education	1.57 (0.78)*	2.06 (0.91)*
R^2	0.03	0.04
F	1.05	1.60
n	193	190

Educational background: base group is vocational education. Control variables included: sex and age

[a]Version A includes Task 1, 2 and 5

[b]Version B includes Task 1, 3 and 4

*$p < .05$

senior general secondary education in version B, $t(187) = 2.25$, $p = .026$, but not in Version A, $t(184) = 1.67$, $p = .097$. So, while the regression models as a whole were not significant, there is some evidence that pre-service teachers with university preparatory education scored better than pre-service teachers with lower level educational backgrounds.

Data as evidence

Table 2 provides an overview of the results of the data as evidence component.

Using data

Open-ended questions In Task 1, 31% of the respondents expressed their trust in the conclusions of the fourth-year pre-service teachers who had used data from a large-scale research project. Another 10% combined such data-based arguments with personal experience or personal opinions. Examples of data-based arguments were references to the size or representativeness of the sample, or claims that the research was well-conducted and thus yielded reliable information. In total, 38% of the respondents were led by their personal opinion, personal experience or a combination of both. A relationship between math ability and math enjoyment was mentioned by 22% of the respondents. In Task 4, 91% made use of the data to draw a conclusion, as evidenced from their use of one or more descriptive statistics. In Task 5, 72% of the respondents used the data as evidence, while 14% used their personal opinion to make or refute an argument or to make a prediction. So, while in Task 1 only around 40% of the pre-service teachers regarded research data as valuable evidence, in Tasks 4 and 5 a (large) majority used the data as evidence.

True/false statements Both Statement 1.1, expressing agreement with the fourth-year pre-service teachers who used research data as evidence for their conclusions, and Statement 1.3, expressing agreement with classmates who used their personal experience, were evaluated as correct by around 40%. Agreement with Statement 1.4, expressing a supposedly generally held belief, was endorsed by a smaller figure (17.5%).

Conclusion For Task 1, the results showed that the main sources of evidence (data, personal opinion and personal experience) were all valued by substantial minorities of respondents, where sometimes sources of evidence were combined. Both in the open-ended responses and in the true/false statements, around 40% of the respondents showed that they appreciated the data as evidence. With respect to personal experience as the source of evidence, only 12% used this source as an argument in the open-ended question, while 42.7% agreed with the statement that this is a valid source of evidence. Respondents did not use supposed generally held beliefs as sources of evidence in their open-ended responses and only 17.5% agreed that such a belief is a valid source. However, using one's *personal* opinion as evidence was present in 31% of the open-ended question answers. When, in Tasks 4 and 5, respondents were actually asked to use the data provided as evidence, a (large) majority did use the data as evidence.

Table 2 Pre-service teachers' knowledge of the data as evidence component of ISI

Question type	Statement or task	Strategy[a]	Percentage[b]
Using data			
Open-ended questions	1	Evidence source for conclusion:	
		Data	31%
		Data + personal experience or opinion	10
		Personal opinion	26
		Personal experience	7
		Personal opinion + personal experience	5
		No or other sources	21
	4	Use data as evidence when generalizing from 2 distributions	91
	5	Use data as evidence when predicting from scatterplot	72
	5	Use personal opinion as evidence when predicting from scatterplot	14
True/false statements	1.1	Use data as evidence for conclusion	40.2
	1.3	Use personal experience as evidence for conclusion	42.7
	1.4	Use tradition/prejudice as evidence for conclusion	17.5
Describing data			
Open-ended questions	4	When generalizing from 2 distributions, compare or use…	
		Means	33
		Spread	27
		Positive values	25
		Negative values	22
		Majorities	13
		Minimum values	10
		Global shape	9
	5	When predicting from scatterplot, use…	
		Majority	40
		Mean	25
		Other descriptive statistics	7

Table 2 continued

Question type	Statement or task	Strategy[a]	Percentage[b]
True/false statements		When generalizing from 2 distributions:	
	4.4	Compare means	88.4
	4.5	Consider spread	63.2
	4.3	Compare overlap of data	57.3
	4.2	Compare proportions negative	46.8
	4.1	Compare modes	39.1
		When predicting from scatterplot, use…	
	5.3	Majority	87.2
	5.5	Mean	72.6
	5.1	Range	55.7
	5.2	Midpoint of range	27.9

[a]See Online resource for full statements

[b]Percentage of respondents that proposed a particular strategy in an open-ended question or percentage of respondents that agreed with a particular statement

Describing data

Open-ended questions The percentage of respondents who used correct descriptive statistics was lower in Task 4 than in Task 5. In Task 4, of those who described the data, 49.5% of the respondents backed up their answer with one or more correct descriptive statistics (mean: 36.3% of respondents; majority: 14.3%; global shape: 9.9%), while 50.5% of the respondents used strategies that were incomplete when not used in conjunction with a correct strategy, such as focusing on the spread of the distribution (27%), or focusing on negative (22%) or positive (25%) values of the distributions. In Task 5, of those who described the data, 90.3% backed up their prediction with a correct descriptive statistic (majority: 55.6%; mean: 34.7%).

True/false statements The respondents evaluated statements in which the inference was based on descriptive statistics which determination requires to take all values of the distribution into account (mean, majority, spread) better than on descriptive statistics that do not require to take all values of the distribution into account (mode, overlap of data, range and midpoint of the range): 77.8% on average versus 54.6 on average.

Conclusion In evaluating the true/false statements, around three quarters of the respondents acknowledged the validity of using descriptive statistics which determination requires to take all values of the distribution into account as arguments in ISI. In the open-ended responses, most respondents also used these global statistics in Task 5, whereas only half did in Task 4. Although only a quarter of the respondents used which determination does not require to take all values of the distribution into account incorrectly, half had problems in identifying that these are incorrect statistics when used in isolation.

Generalization beyond the data

Table 3 provides an overview of the results of the Generalization beyond the data component.

Open-ended questions

In Task 2 and Task 4, in the majority of the answers, it was unclear whether the correct population was held in mind. In Task 2, from only 17% of the answers could it be deduced that the answer pertained to the population of all Dutch children; for example, because respondents proposed sampling from multiple schools. In 76% of the answers, the population remained implicit: Many respondents suggested sampling from "each class" or "every group," which could imply that the intention was to sample from each grade level and from every school in the population, but it could also imply sampling from each class of students *in one particular school*. This vagueness about the population was even more evident in Task 4, where 86.2% drew a conclusion that was neither explicitly descriptive

Table 3 Pre-service teachers' knowledge of the generalization beyond the data component of ISI

Question type	Statement or task	Strategy[a]	Percentage[b]
Open-ended questions	1	(Large) sample yields reliable information	23%
	2	Proposed sample yields good picture	26
	2	Population referred when proposing sample selection method	
		Population implicit	76
		Population is the Netherlands	17
		Population is class or school	7
	4	Type of conclusion	
		Descriptive conclusion	7
		Inferential conclusion	3
		Unclear whether conclusion is descriptive or inferential	83
	1	Every child or school is different, so aggregation is impossible	4
	3	Every child is different, so graph can take any shape	4
	4	Every child is different, so generalization is impossible	4
	5	Every child is different, so prediction is impossible	6
True/false statements	4.7	When generalizing from 2 distributions, make an completely certain generalization	12.9
	5.4	In predicting from scatterplot, make a probabilistic generalization	92.1
	1.2	Making generalizations is impossible, due to the uniqueness of the elements in the population	Excluded from analysis

[a]See Online resource for full statements

[b]Percentage of respondents that proposed a particular strategy in an open-ended question or percentage of respondents that agreed with a particular statement

nor explicitly inferential. A typical example of such a conclusion is that "boys are more positive about math than girls, because their mean is higher." In this example, it remains obscure whether the children referred to are the children in the sample only or in the population in general. A final result is that in Tasks 1, 3, 4 and 5, a constant and very small minority of around 4% remarked that every child is different, and that therefore aggregation, generalization, or prediction are impossible.

True/false statements

Both statements of the component Generalization beyond the data were evaluated correctly by a large majority of the pre-service teachers (87.3 and 92.1%, respectively). So most of the respondents recognized that making a probabilistic generalization is possible, while a completely certain generalization is not.

Conclusion

Overall, the evidence suggests that a very small minority refused to generalize at all, while probably up to 20% had the correct population in mind in Tasks 2 and 4. Although around 90% of the respondents recognized that generalizations are inherently uncertain, at least three quarters of the respondents did not make explicit what population they had in mind when answering the questions.

Uncertainty regarding inferences

Tables 4 and 5 provide an overview of the results of the uncertainty in inferences component.

Sampling methods

Open-ended questions In Task 2, 73% of the respondents proposed a distributed sampling method in which the quota for one or more variables needed to be established—math ability was mentioned most often as a variable to account for. Random sampling was proposed by 13; 2% proposed sampling "all children"; 5% would restrict the sample to one value of an external factor, for instance only sampling children with an average math ability; in 7% of responses the sampling method was missing or unclear.

True/false statements A majority (62.0%) of the respondents correctly evaluated the statement that a random sample is a valid sampling method, and a vast majority (91.8%) understood that one's own class of students is very unlikely to be representative of all students in a country. However, 88.3% incorrectly judged distributed sampling to be a correct sampling method.

Conclusion The dominant sampling method proposed was distributed sampling. While only 13% proposed random sampling, 62% agreed that it is a correct sampling strategy. Most respondents regarded one class as unrepresentative. In the open-ended responses, no respondent proposed this strategy.

Table 4 Pre-service teachers' knowledge of the uncertainty in inferences subcomponents sampling method and sample size

Question type	Statement or task	Strategy[a]	Percentage[b]
Sampling methods			
Open-ended questions	2	Proposed sampling method	
		Distributed sampling	73%
		Random sampling	13
		Restrict sample to one or few values of other variable	5
		Sample entire population	2
		Unclear or absent	7
True/false statements	2.6	Distributed sampling yields a representative sample	88.3
	2.2	Random sampling yields a representative sample	62.0
	2.1	A sample of one class of students is a representative sample	8.2
Sample size			
Open-ended questions	2	Sample as many boys as girls	28
	2	Proposed sample size	
		Between 8 and 19 children	3
		Between 20 and 39 children	6
		Between 40 and 80 children	9
		Large sample or as large as possible	4
		Multiple or as many as possible schools	15
True/false statements	2.3	200 is a sufficiently large sample size	48.4
	2.4	A sample of 100,000 is better than a sample of 3000	54.9
	2.5	Comparison of 2 distributions is only possible when sample sizes are about equal	91.8

Sample size

Open-ended questions In Task 2, 35% of the respondents paid attention to the size of the proposed sample, excluding the suggestion that as many boys as girls should be sampled. This latter idea was proposed by 28% of the respondents. The requirement to sample multiple schools, or as many schools as possible, was mentioned by 15%. Only about a quarter of the respondents referred to the number of children to be sampled. Of these, 9% proposed rather small samples (sizes below 40).

True/false statements The two statements that concerned the optimal sample size for making a generalization were evaluated correctly by about half of respondents (45.1 and 48.4%), while only 8.2% accurately inferred that the comparison of two unequally sized groups is possible (Statement 2.5).

Conclusion Only around one third of the respondents paid attention to the sample size in the open-ended response, and of those, a substantial minority proposed using a very small sample. The statements revealed that about half had a correct idea of appropriate sample sizes. More than 90% of the respondents thought that unequally sized samples cannot be

Table 5 Pre-service teachers' knowledge of the uncertainty in inferences subcomponents sampling variability and uncertainty language

Question type	Statement or task	Strategy[a]	Percentage[b]
Sampling variability			
Open-ended questions	3	General strategy to predict of graph of $n = 40$ based on $n = 20$	
		Copy shape of smaller sample	45
		Double distribution	27
		Decrease spread	10
		Increase spread	5
		Other	2
		No prediction	11
	3	Widen range in prediction of graph of $n = 40$?	
		Yes	53
		No	37
	3	Fill in gaps in prediction of graph of $n = 40$?	
		Yes	49
		No	39
True/false statements	2.7	Another, equally well-selected sample may give entirely different result. Therefore, generalization is impossible	60.4
		When sample doubles:	
	3.1	Make distribution symmetric	45.8
	3.2	Double distribution	39.5
	3.3	Keep mean constant	81.2
	3.4	Widen range	62.7
	3.5	Smoothen distribution	46.5
Uncertainty language			
Open-ended questions	4	Use uncertainty language when generalizing from 2 distributions	11
	5	Type of uncertainty language used when predicting from scatterplot	
		"Think"	30
		"Probably" or similar	10
		"Can"	8
		"Other outcome is possible because there are exceptions" or similar	7
		"One has no certainty" or similar	6
		Chance language	5
Statements	4.6	$n = 15$ is too small for any generalization	71.2

[a]See Online resource for full statements

[b]Percentage of respondents that proposed a particular strategy in an open-ended question or percentage of respondents that agreed with a particular statement

compared: a result confirmed by the open-ended responses, where almost half of the references to sample size concerned the proposal to sample as many boys as girls.

Sampling variability

Open-ended questions 35% of the respondents made a completely sensible prediction by copying the general shape of the small sample, widening the range and filling in one or more gaps in the distribution of the small sample—events that are likely to happen when a sample grows. Another 10% copied the general shape, but did not fill in the gaps and/or broaden the range. 27% made a deterministic prediction by exactly doubling the smaller sample. Only 3% argued that the results would be completely uncertain; for example, because every child is different. These 3% still made a prediction.

True/false statements Statement 3.3, which states that the mean remains approximately constant when a sample grows, was well evaluated (81.2%), compared to the other statements about the change in the distribution (percentages of correct answers varied between 46.5 and 62.7%). The results were less positive for the fundamental statement about sampling variability. Only 39.6% of the respondents correctly evaluated Statement 2.7 that states: "Whatever groups we select, in the end we cannot say anything about boys and girls in general. If we would have selected other boys and girls, the result could have been entirely different."

Conclusion Although almost half of the respondents could make an (almost) correct prediction and around 80% of the respondents correctly predicted that the mean of the larger sample would be similar to the mean of the smaller sample, a substantial minority of around 40% made a prediction that was too deterministic by copying the small sample distribution exactly. Moreover, only around 40% understood that generalization is possible, because sample-to-sample variability is low for a deliberately selected sample (i.e,. a sample where an appropriate sampling method and sample size is used).

Uncertainty language

Open-ended questions In Task 5, more than half of the respondents employed probability language, ranging from weak indications of uncertainty, such as "I think," to stronger utterances, such as "probably." In Task 4, in contrast, only 11% used uncertainty language.

True/false statements A majority (71.2%) incorrectly agreed that based on a sample size of 15 *nothing at all* can be said about the population (Statement 4.6).

Conclusion In Task 5, more uncertainty language was found than in Task 4. In Task 4, where 93.1% drew a conclusion, almost three quarters of the respondents nevertheless agreed that based on the small sample size *nothing at all* could be said about the population.

Discussion and conclusion

Main results

The aim of this study was to describe the ISI-CK of first-year pre-service teachers. The results showed that, although a substantial minority of the respondents valued the data as evidence, other groups regarded personal opinion and personal experience as appropriate sources of evidence from which to draw conclusions. However, when provided with sample data, a large majority did use the data as evidence, rather than relying on other sources of evidence. When using these sample data, most respondents showed a good propensity to use correct descriptive statistics that take global aspects of the sample distribution into account, while around half of the respondents had difficulty in evaluating the incorrectness of descriptive statistics that which determination does not require to take all values of the distribution into account. In proposing a sampling method and in comparing two sample distributions, the population remained implicit in approximately three quarters of the responses. In the case where two sample distributions were compared, there was negligence in expressing uncertainty when making inferences. Around 90% of the respondents understood that making complete and certain generalizations is impossible, but at the same time, almost three quarters of them thought that a small sample of 15 does not permit any generalization. Around three quarters of the respondents proposed distributed sampling to select a sample, although around 60% of the respondents agreed that random sampling is a valid sampling method. Statements about appropriate sample sizes were correctly evaluated by around half of the respondents, but only about a third of the respondents paid attention to the sample size in their open-ended responses. More than 90% believed that in order to compare two groups, those groups need to be of the same size. Finally, less than 40% understood the underlying tenet of sampling variability, i.e., that generalization is possible because sample-to-sample variability will be low for a sample where an appropriate sampling method and sample size is used.

Discussion

We found that the pre-service teachers preferred distributed sampling to random sampling, which is consistent with our previous findings (De Vetten et al. 2018b) where we argued that this preference for distributed sampling could be due to a sense of loss of control when using random sampling. We called for more research to investigate this sense of control in choosing a sampling method. While the present findings are also consistent with Meletiou-Mavrotheris et al. (2014), who showed that generally pre-service teachers dismiss biased sampling methods and understand that random sampling is a valid sampling method, it has not been described previously that, although many pre-service teachers acknowledge the validness of random sampling, they still prefer distributed sampling.

Our finding that 91.8% of the respondents knew that one class of students is not representative of all students in a country stands in contrast with the finding of Watson (2001) that only about a third of primary school teachers discerned that a claim made in a journal article was based on a non-representative sample. An initial explanation for these divergent results could be that the context used in our study was more familiar to the students. A second explanation may be that the statement we used explicitly delineated that one class of students was representative, while the idea of selective sampling was more implicitly questioned in the task used by Watson.

Our result showing that many pre-service teachers understand that global descriptive statistics can be used, matches the findings of Konold and Pollatsek (2002), who showed that students intuitively summarize data around the middle 50% of the distribution, which can be interpreted as a global view of data. Unreported so far in the literature is the finding that pre-service teachers scored lower on statements involving descriptive statistics which determination does not require to take all values of the distribution into account, which may be explained by their unfamiliarity with such descriptive statistics. A second explanation may be that they lack the confidence to conclude that a certain descriptive statistic is not a sufficient argument within ISI.

We did not find significant relationships between educational background and mathematical content knowledge on the one hand and ISI-CK on the other hand. With respect to mathematical content knowledge, an explanation might be that the mathematics test focused primarily on applying mathematical procedures, while the ISI instrument demanded conceptual understanding of ISI. Moreover, the statistics covered in the mathematics tests dealt primarily with reading off tables and graphs. Further work is required to account for our finding that pre-service teachers with a university preparatory education tended to score better on the test than pre-service teachers with other educational backgrounds. It may be the case that the former group of pre-service teachers has a higher ability to reason abstractly, which helped them to perform better on the test (Korpershoek et al. 2006) or the content of their university preparatory education differed from the content of the preparatory education of other respondents.

Because, in many responses on Task 4, it was unclear whether the conclusion was descriptive or inferential, it is doubtful whether the respondents were actually making inferences, or were merely describing the sample data. Since no reference was given to uncertainty or sample characteristics, most respondents had probably not consciously made any generalization beyond the data. We found a similar result in a previous study (De Vetten et al. 2018a); there we conjectured that the need to generalize may not have been compelling enough, or that the pre-service teachers may not be inclined to generalize beyond the sample because, in their role as future teachers, they had a class of primary school students in mind as their population of interest, in which case, a description would suffice. This highlights the need to increase pre-service teachers' awareness of the inferential nature of research questions before they can actually be involved in discussing other aspects of ISI, such as uncertainty and representativeness.

A final important point concerns pre-service teachers' understanding of the possibility of making any generalization based on sample data. Most respondents did not seem to understand that a representative sample can be used to draw useful, albeit uncertain, conclusions about an unknown population. So, apart from the unfamiliarity of pre-service teachers with ISI, it could be that many of them are reluctant to make generalizations beyond the data collected because they do not understand this underlying logic of sampling (Watson 2004). In De Vetten et al. (2018b), we describe a first attempt at fostering pre-service teachers' understanding of this logic.

Limitations

A number of issues warrant a cautious interpretation of the results. First, the context of our study is the Dutch educational system where students enter teacher college immediately after secondary education with the teaching of mathematics probably not being uppermost in their minds. The results are, accordingly, not readily generalizable to other international contexts. In future research, the study could be replicated in different settings. Second,

since the instrument used in this exploratory study was the first to investigate teachers' knowledge of the entire concept of ISI, in future work, the instrument could be further developed. There is some evidence for the reliability of the instrument: There were very few inconsistencies between the open-ended responses and the true/false statements. Also, the cognitive interviews showed that the pilot respondents interpreted the statements as intended. Moreover, in De Vetten et al. (2018b), where we used a similar instrument but with more qualitative data available, the data also showed that the statements were interpreted as intended. Still, there is a need for further research to test the reliability of the instrument, for example using test–retest methods. Also, attention could be paid to the effect of the formulation of the statements as either correct or incorrect because in the current study, pre-service teachers were better able to evaluate correct than incorrect statements (respectively, 67.7 vs. 52.5% were evaluated correctly).

Implications for teacher education

In conclusion, many of the pre-service teachers did not seem to be really engaged in making inferences and did not understand the logic of sampling. The challenge for teacher education, therefore, is to develop instructional heuristics that evoke the need to engage in ISI. To this end, when pre-service teachers conduct ISI investigations, research questions need to be chosen in such a way that the question posed allows the pre-service teacher to feel that description is insufficient and makes generalization natural and inevitable. One way of doing this is to have a situation in which the sample and the population are concrete and visible, such as taking a sample of pages from all books in a library to answer the question which word is most frequently used in the library (see De Vetten et al. 2018b). A second way to evoke the need to engage in ISI is to place less emphasis on the data analysis phase of the empirical inquiry cycle. Going through the entire cycle is a commonly proposed heuristic for learning statistics (Garfield and Ben-Zvi 2007); however, Leavy (2010) reported that attention to data analysis comes at the expense of engaging in ISI and critically reflecting on the sample. Our study shows that the pre-service teachers seem to be reasonably good at analyzing sample data with commonly used descriptive statistics that requires to take all values of the distribution into account, such as the mean. This may allow reducing the focus on data analysis, so that pre-service teachers do not end the investigative process after having analyzed the sample data, but rather proceed to interpretation and inferential reasoning, and reflect critically on what the sample results say about the population. These recommendations would, hopefully, help to prompt pre-service teachers to engage in informal statistical inference, which in turn, would support them in developing primary school students' inferential reasoning skills.

Acknowledgements We thank Niels Smits for his methodological advice and the teacher educators of the participating colleges for their cooperation.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht, the Netherlands: CD-ß Press, Center for Science and Mathematics Education.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*(1–2), 5–26.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

Batanero, C., & Díaz, C. (2010). Training teachers to teach statistics: What can we learn from research? *Statistique et enseignement, 1*(1), 5–20.

Ben-Zvi, D. (2006). *Scaffolding students' informal inference and argumentation.* Paper presented at the seventh international conference on teaching statistics, Salvador, Brazil.

Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics, 88*(3), 291–303.

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability, 21*(1), 33–46.

Blom, N., & Keijzer, R. (1997). Het rekenverleden: Doe er wat mee! [Your math past: work with it!]. *Willem Bartjens, 17*(2), 20–25.

Burgess, T. (2009). Teacher knowledge and statistics: What types of knowledge are used in the primary classroom? *Montana Mathematics Enthusiast, 6*(1&2), 3–24.

Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65–72.

Canada, D., & Ciancetta, M. (2007). *Elementary preservice teachers' informal conceptions of distribution.* Paper presented at the 29th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Stateline, NV.

Chatzivasileiou, E., Michalis, I., Tsaliki, C., & Sakellariou, I. (2011). *Service elementary school teachers' conceptions of arithmetic mean.* Paper presented at the 58th World Statistical Congress, Dublin.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1–78. https://doi.org/10.1207/S1532690xci2101_1.

Darling-Hammond, L., Barron, B., Pearson, P. D., Schoenfeld, A. H., Stage, E. K., Zimmerman, T. D., et al. (2008). *Powerful learning: What we know about teaching for understanding*. San Francisco, CA: Jossey-Bass.

delMas, B., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28–58.

De Vetten, A. J., Schoonenboom, J., Keijzer, R., & van Oers, B. (2018a). The growing samples heuristic: Exploring pre-service teachers' reasoning about informal statistical inference when generalizing from samples of increasing size (submitted).

De Vetten, A. J., Schoonenboom, J., Keijzer, R., & van Oers, B. (2018b). The development of informal statistical inference content knowledge of pre-service primary school teachers during a teacher college intervention (submitted).

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, R., et al. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre K-12 curriculum framework.* Alexandria, VA: American Statistical Association.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396.

Groth, R. E., & Bergner, J. A. (2005). Pre-service elementary school teachers' metaphors for the concept of statistical sample. *Statistics Education Research Journal, 4*(2), 27–42.

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning, 8*(1), 37–63.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research, 7*(1), 1–20.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511.

Jacobbe, T., & Carvalho, C. (2011). Teachers' understanding of averages. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI study 18 and 2008 IASE round table conference* (pp. 199–209). Dordrecht, The Netherlands: Springer.

Koleza, E., & Kontogianni, A. (2016). Statistics in primary education in Greece: How ready are primary teachers? In D. Ben-Zvi & K. Makar (Eds.), *The teaching and learning of statistics: International perspectives* (pp. 289–298). Cham: Springer.

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., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R. et al. (2002). *Students' use of modal clumps to summarize data*. Paper presented at the sixth international conference on teaching statistics (ICOTS 6), Cape Town, South Africa.

Korpershoek, H., Kuyper, H., & Werf, M. P. C. (2006). *Havo-5 en vwo-5 en de tweede fase: De bovenbouwstudie van VOCL'99*. Groningen: GION.

Leavy, A. M. (2010). The challenge of preparing preservice teachers to teach informal inferential reasoning. *Statistics Education Research Journal, 9*(1), 46–67.

Liu, Y., & Grusky, D. B. (2013). The payoff to skill in the third industrial revolution. *American Journal of Sociology, 118*(5), 1330–1374.

Makar, K. (2016). Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning, 18*(1), 1–24.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1–2), 152–173.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105.

Meijerink, H. (2009). *Referentiekader taal en rekenen: De referentieniveaus. [Reference levels language and mathematics]*. Enschede: Doorlopende Leerlijnen Taal en Rekenen.

Meletiou-Mavrotheris, M., Kleanthous, I., & Paparistodemou, E. (2014). *Developing pre-service teachers' technological pedagogical content knowledge (TPACK) of sampling*. Paper presented at the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, AZ.

Meletiou-Mavrotheris, M., & Paparistodemou, E. (2015). Developing students' reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics, 88*(3), 385–404.

Mooney, E., Duni, D., VanMeenen, E., & Langrall, C. (2014). *Preservice teachers' awareness of variability*. Paper presented at the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, AZ.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal, 7*(2), 83–106.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.

Saldanha, L., & Thompson, P. (2007). Exploring connections between sampling distributions and statistical inference: An analysis of students' engagement and thinking in the context of instruction involving repeated sampling. *International Electronic Journal of Mathematics Education, 2*(3), 270–297.

Scientific and Ethical Review Board of the Faculty of Behavourial Science of Vrije Universiteit Amsterdam. (2016). *Code of ethics for research in the social and behavourial sciences involving human participants*. Retrieved from https://www.fgb.vu.nl/en/Images/ethiek-reglement-adh-landelijk-nov-2016_tcm264-810069.pdf.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 205–237). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

SLO. (2008). *TULE—Rekenen/wiskunde. Inhouden en activiteiten bij de kerndoelen. [TULE—mathematics. Content and activities adjoining core objectives.]*. Enschede, The Netherlands: SLO, nationaal expertisecentrum voor leerplanontwikkeling.

Watson, J. M. (2001). Profiling teachers' competence and confidence to teach particular mathematics topics: The case of chance and data. *Journal of Mathematics Teacher Education, 4*(4), 305–337.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277–294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watson, J. M., & Callingham, R. (2013). Likelihood and sample size: The understandings of students and their teachers. *The Journal of Mathematical Behavior, 32*(3), 660–672.

Watson, J. M., & Kelly, B. (2005). Cognition and Instruction: Reasoning about bias in sampling. *Mathematics Education Research Journal, 17*(1), 24–57.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for research in mathematics education, 31*(1), 44–70.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

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.