Pre-service teachers and informal statistical inference

de Vetten, Arjen; Schoonenboom, Judith; Keijzer, Ronald; van Oers, Bert

published in
Topics and trends in current statistics education research
2019

DOI (link to publisher)
10.1007/978-3-030-03472-6_9

document version
Publisher's PDF, also known as Version of record

Chapter 9
Pre-service Teachers and Informal Statistical Inference: Exploring Their Reasoning During a Growing Samples Activity

Arjen de Vetten, Judith Schoonenboom, Ronald Keijzer and Bert van Oers

Abstract Researchers have recently started focusing on the development of informal statistical inference (ISI) skills by primary school students. However, primary school teachers generally lack knowledge of ISI. In the literature, the growing samples heuristic is proposed as a way to learn to reason about ISI. The aim of this study was to explore pre-service teachers' reasoning processes about ISI when they are engaged in a growing samples activity. Three classes of first-year pre-service teachers were asked to generalize to a population and to predict the graph of a larger sample during three rounds with increasing sample sizes. The content analysis revealed that most pre-service teachers described only the data and showed limited understanding of how a sample can represent the population.

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

A. de Vetten (✉) · B. van Oers
Section of Educational Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: a.j.de.vetten@fsw.leidenuniv.nl

B. van Oers
e-mail: bert.van.oers@vu.nl

J. Schoonenboom
Department of Education, University of Vienna, Vienna, Austria
e-mail: judith.schoonenboom@univie.ac.at

R. Keijzer
Academy for Teacher Education, University of Applied Sciences iPabo, Amsterdam, The Netherlands
e-mail: r.keijzer@ipabo.nl

G. Burrill and D. Ben-Zvi (eds.), *Topics and Trends in Current Statistics Education Research*, ICME-13 Monographs, https://doi.org/10.1007/978-3-030-03472-6_9

9.1 Introduction

In today's society, the ability to reason inferentially is increasingly important (Liu and Grusky 2013). One form of inferential reasoning is informal statistical inference (ISI), which is defined as "a generalized conclusion expressed with uncertainty and evidenced by, yet extending beyond, available data" (Ben-Zvi et al. 2015, p. 293) without the use of formal statistical tests based on probability theory (Harradine et al. 2011). In recent years, statistics education researchers have focused on how primary school students can be introduced to ISI. Scholars have hypothesized that if children are familiarized with the concept in primary school, they will understand the processes involved in ISI reasoning and in statistical reasoning in general (Bakker and Derry 2011; Makar et al. 2011). Evidence suggests that meaningful learning environments can render ISI accessible to primary school students (Ben-Zvi et al. 2015; Meletiou-Mavrotheris and Paparistodemou 2015).

If children are to be introduced to ISI in primary school, future teachers need to be well prepared to provide this introduction (Batanero and Díaz 2010). They must have appropriate knowledge of the field that extends beyond the students' knowledge (Burgess 2009). It has been shown, however, that pre-service teachers' knowledge of ISI is generally weak (Batanero and Díaz 2010; De Vetten et al. 2018). This points out the need to improve the ISI content knowledge of pre-service teachers.

Current research provides only scant evidence for how to support the development of pre-service teachers' knowledge of ISI (Ben-Zvi et al. 2015). In some statistics education literature, the growing samples heuristic is recommended to stimulate ISI reasoning (Joan Garfield et al. 2015). The idea of this heuristic is that samples of increasing size are used to make inferential statements about a larger sample or population. Using this heuristic to informally and coherently construct and discuss ISI has typically not been investigated in the context of teacher education. Therefore, we implemented the growing samples heuristic in three classes of first-year pre-service teachers and explored their ISI reasoning when engaged in an activity that applies this heuristic.

9.2 Theoretical Background

9.2.1 Teachers' Knowledge of ISI

Teachers need to possess thorough knowledge of the content they teach (Hill et al. 2008) that extends beyond what their students actually learn (Ball et al. 2008), since the former's content knowledge impacts the latter's learning achievements (Rivkin et al. 2005) and facilitates the development of pedagogical content knowledge (Ball et al. 2008; Shulman 1986). It has been shown that these relationships also hold for ISI (Burgess 2009; Leavy 2010).

To conceptualize the required knowledge of ISI for pre-service teachers, we used the Makar and Rubin (2009) ISI framework. The three components of this framework are broad to include various types of students (Makar and Rubin 2014). For this study among pre-service teachers, we conceptualized the components in the following way:

1. "Data as evidence": The inference needs to be based on the data and not on tradition, personal beliefs or experience. To base an inference on the sample data, the data need to be analyzed descriptively, 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).
2. "Generalization beyond the data": The inference goes beyond a description of the sample data to make a claim about a situation beyond the sample data.
3. "Probabilistic language": The inference includes a discussion of the sample characteristics, such as the sample size and sampling method, and what these characteristics imply about the representativeness of the sample and the certainty of the inference. Moreover, the inference requires understanding whether a sample is properly selected, the sample-to-sample variability is low, and this sample is representative of the population and can be used for an inference.

One of the studies that have investigated (pre-service) primary school teachers' content knowledge is De Vetten et al. (2018). In a large-scale questionnaire study, they found that about half of the pre-service teachers agreed that data can be used as reliable evidence for a generalization. The authors also showed that the respondents were able to discern that probabilistic generalizations are possible, while deterministic generalizations are not. The evidence for the Probabilistic language component suggests that many pre-service teachers have a limited understanding of sampling methods, sample size, representativeness and sampling variability (De Vetten et al. 2018; Meletiou-Mavrotheris et al. 2014; Mooney et al. 2014; Watson 2001). With respect to the knowledge of descriptive statistics more generally, (pre-service) teachers' knowledge has been shown to be typically superficial (Batanero and Díaz 2010; Garfield and Ben-Zvi 2007; Jacobbe and Carvalho 2011). More specifically, pre-service teachers tend to focus on measures of central tendency at the expense of measures of dispersion (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). De Vetten et al. (2018) asked respondents to evaluate which descriptive statistics were well suited as arguments within ISI. The respondents acknowledged that ISI can be based on global descriptive statistics, but they did not recognize that ISI based on local aspects of the sample distribution is not correct. These studies indicate that there is a need to improve pre-service teachers' ISI content knowledge.

9.2.2 *Using the Growing Samples Heuristic to Support the Development of ISI*

Research on how to support pre-service teachers' development of ISI content knowledge is almost nonexistent (Ben-Zvi et al. 2015). Leavy (2006) intervention study examined pre-service primary school teachers' distributional reasoning when engaged in experimental investigations. She found that the pre-service teachers in the sample tended to compute measures of centrality only rather than explore datasets, for example, using graphical representations. Moreover, the teachers often neglected the role of variation in comparing distributions. However, the participants became more attentive to variation and looked more at aggregate features of the distributions. Although the tasks used were inferential, the analysis focused on distributional reasoning only. Leavy (2010) showed that final-year pre-service teachers do not reflect on the meaning of the graphs and calculations they perform. The activity at the start of the intervention involved making inferences and discussing sampling issues, but the author did not analyze the activity in depth. These studies revealed that the pre-service teachers in her sample tended to restrict their attention to descriptive statistics, rather than how these descriptive statistics can be used in ISI.

In the context of statistics education generally, the growing samples heuristic has been suggested as a promising approach to support the development of ISI reasoning (Joan Garfield and Ben-Zvi 2008; Garfield et al. 2015). The idea of this heuristic is that samples of increasing size are used to make inferential statements about a larger sample or population. Ben-Zvi et al. (2012) showed that the heuristic helps middle-grade students not only describe samples but also draw conclusions beyond the data. Moreover, these students' reasoning about uncertainty developed from either certainty only or uncertainty only to more sophisticated reasoning in which probability language was used. Bakker (2004) found that when middle-grade students use this heuristic, they develop coherent reasoning about key distributional aspects of samples, such as center, spread and density. We hypothesize that the growing samples heuristic can aid the use of data as evidence because this heuristic draws students' attention repeatedly to the data. Research suggests it is not self-evident that students see sample data as evidence from which to make generalizations and predictions (Ben-Zvi et al. 2007; Makar and Rubin 2009). The heuristic could also help students to understand sample-to-sample variability because as the sample size increases, the shape of the distribution stabilizes and more likely resembles the population distribution (Garfield and Ben-Zvi 2008; Konold and Pollatsek 2002). In our view, the heuristic may be well suited for teacher education, because the relative simplicity of the heuristic allows pre-service teachers to translate the growing samples activities to their own teaching practice in primary schools.

9.2.3 Research Aim and Question

Until now, little, if anything, is known about how the growing samples heuristic supports pre-service teachers' development of reasoning about ISI. We hypothesize that pre-service teachers may reason differently from middle school students. On the one hand, we hypothesize that pre-service teachers are better suited to reason about ISI because they have more (procedural) statistical knowledge, which they can use in reasoning about ISI. Moreover, given their older age, they may be more able to reason about an abstract population. On the other hand, their future role as teachers might hinder them in drawing inferences as they might have a class of children in mind as their natural population of interest (Schön 1983). Therefore, pre-service teachers could relate sample results to a class instead of to an abstract population.

The aim of this exploratory study was to investigate the reasoning about ISI of 35 pre-service primary school teachers divided over three classes when they were engaged in a growing samples activity (Fig. 9.1). The research question is: What reasoning about informal statistical inference do first-year pre-service primary school teachers display when they are engaged in a growing samples activity, and what is the quality of their reasoning?

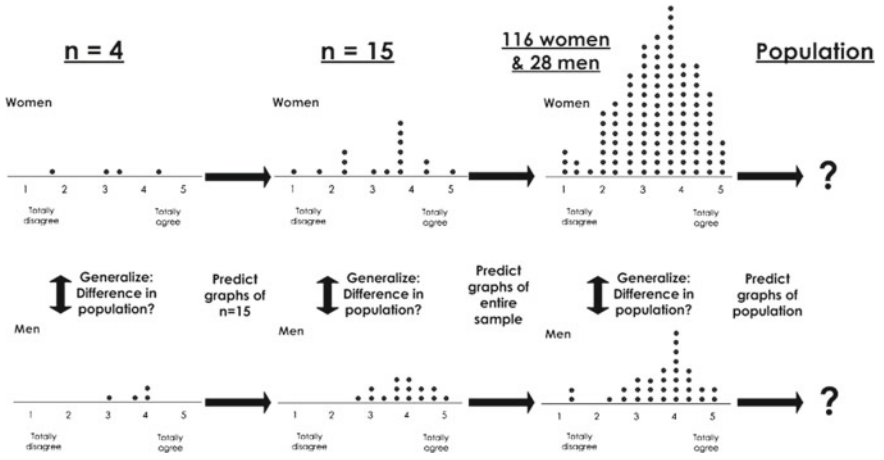

Fig. 9.1 Growing samples activity used in the current study

9.3 Methods

9.3.1 Intervention

The three components of ISI provided the framework for the pre-service teachers' ISI learning objectives. We formulated 10 learning objectives (Table 9.3, the last column), which informed the design of the growing samples activity (Fig. 9.1). The activity was inspired by the activities used by Bakker (2004) and Ben-Zvi et al. (2012) and consisted of three rounds. In each round, the participants answered the question, "Is the attitude toward mathematics of first-year male pre-service teachers in general more positive than the attitude toward mathematics of first-year female pre-service teachers?" Before the participants analyzed the data, they discussed how the data were collected and how the data could be used to answer this question, the "talking through the data creation process" (Cobb and Tzou 2009). This process was used to support the participants' confidence in the validity of their conclusions. The teacher educator stressed that the question pertained to the population of all Dutch first-year pre-service teachers and explained that the data came from a research project conducted among pre-service teachers at their teacher college the previous year. The data showed the averages of three 5-point Likert items. Therefore, the data could take on values between one and five with increments of one third. Next, the participants were provided with graphs of samples of increasing size. During each round, the participants answered the question about the difference in the population and predicted the shape of the graphs of the next round. The sheet of paper on which the participants filled in their answers to these questions also showed the graphs of the particular round. During the first round, the samples consisted of four men and four women, during the second round 15 men and 15 women, and during the third round 28 men and 116 women, which was the size of the original dataset. The sample size of four was meant to elicit responses of high or even complete uncertainty. The samples sizes of the second and third rounds were chosen to investigate whether the certainty of the participants' responses would increase. Round 3 also provided the opportunity to discuss ways to compare samples of unequal sizes. After each round, the answers were discussed in a class discussion. This discussion had similar patterns for each round in each class: The teacher educator asked for an answer to the question, asked on what grounds this answer was reached and probed for the certainty of the conclusion. Next, the prediction for the larger samples was discussed. During the last round, the comparison of the samples of unequal sizes was discussed, and the arguments used during the entire activity were summed up. Some parts of the discussion were more extensive than others, depending on the input from the pre-service teachers. The activity lasted for 80 min. All three rounds were held on one day. The participants worked in groups of two or three (Class A: 7 groups; Class B: 5 groups; Class C: 4 groups). In Class C, the first round was skipped because during the third round in classes A and B motivation seemed to decline.

For each learning objective, we formulated a conceptual mechanism that explained how the activity was hypothesized to scaffold the participants' reasoning to attain the

learning objectives (see Table 9.3, the conceptual mechanism column). We hypothesized that the repetition of the question to generalize and predict would invite the participants to use the data to draw a conclusion (the Data as evidence component). When the sample sizes increased, the averages and the global shape stabilized. We expected that the participants would notice that this and would use them as reliable signals for generalization and prediction. The repetition of the questions would also draw the participants' attention to the inferential nature of the question (component Generalization). Furthermore, we expected that presenting samples of different sizes and shapes would draw the participants' attention to differences in the sample distributions and would influence the participants to realize that other sample distributions could have resulted as well. They would, in turn, take the uncertainty of their conclusions into account. Finally, comparing how the sample data were spread about the center of the data of the various samples would encourage the participants to take uncertainty into account (the Probabilistic language component).

9.3.2 Participants

Three classes (A, B and C) for a total 35 first-year pre-service primary education teachers participated in this study. This was a convenience sample, as the first author also taught their course on mathematics education. They attended a small teacher college in a large city in the Netherlands. In the Netherlands, initial teacher education starts immediately after secondary school and leads to a bachelor's degree. For these students, mathematics teaching is usually not their main motive for becoming teachers. The mean age of the participants was 19.47 years (SD: 1.54), three were male, 20 had a background in secondary vocational education (students attending this type of course are typically between 16 and 20 years old), 13 came from senior general secondary education, and the educational background of the remaining two was either something else entirely or unknown. Table 9.1 shows the educational backgrounds for each class. Whereas descriptive statistics, probability theory and some inferential statistics are part of the mathematics curriculum of senior general education, these topics are generally not taught in secondary vocational education.

9.3.3 Data Collection and Data Analysis

Data collection consisted of the participants' answer sheets and sound recordings of the class discussions.

Content analysis in Atlas.ti was used to analyze the data. A coding scheme was developed based on the learning objectives. All answer sheets (both text and graphs) and class discussions were coded by assigning one or more codes related to the learning objectives to the data. In an iterative process, the first author and an external coder coded and discussed the coding scheme instructions until the instructions

Table 9.1 Educational background per class of pre-service teachers

Educational background	Class A	Class B	Class C	Total
Secondary vocational education	7	9	4	20
Senior general education	7	2	4	13
Something else/unknown	1	0	1	2
Total	15	11	9	35

were deemed clear enough to be applied by a second coder. First, for each round all contributions of each group or participant were put in a table, organized by learning objective. Second, the group or individual contributions were aggregated for each class, organized by learning objective. Third, to measure the quality of the participants' reasoning about ISI, for each class, the contributions per learning objective were compared to the hypothesized conceptual mechanism. This comparison was conducted for each round and separately for the answer sheets and the class discussions and for each class. Per round, an indicator (− −, −, 0, + or + +) was assigned to each learning objective, indicating to what extent the actual reasoning of a class was in line with the hypothesized reasoning (see Table 9.2). The indicators served as quality indicators of the classes' reasoning about ISI. Next, the indicators per round were combined into one indicator per learning objective for the three rounds together, separately for the answer sheets and the class discussions. Finally, the separate indicators for the answer sheets and the class discussions were combined into one indicator per learning objective. These combined indicators were used to compare the reasoning between the three classes. The assignment of the indicators was discussed with an external researcher until consensus was reached (Table 9.3).

9.4 Results

9.4.1 Answer Sheets

Table 9.4 shows the results of the answer sheets, aggregated over the three classes to give a comprehensive picture of the results. Aggregation was possible because the differences between the three classes' answer sheets were small. The left part of Table 9.4 shows how the participants answered the question to draw a conclusion about the difference between men and women in the population; the right part shows how the participants predicted the distribution of a larger sample or the population. In general, most participants used the data as evidence for their conclusions. In 36 of

Table 9.2 Meaning of indicators assigned to indicate the quality of classes' reasoning about ISI

Indicator	Meaning
+ +	The reasoning of all pre-service teachers is in line with the learning objective.
+	The reasoning of about half (Class A: 3 of 7 groups; B: 2 or 3 of 5 groups; C: 2 of 4 groups) of the pre-service teachers is in line with the learning objectives, whereas there is insufficient evidence about the other pre-service teachers; or: 1 or 2 (Class A) or 1 (B and C) groups show misconceptions regarding the learning objective, while the other groups reason in line with the learning objective.
0	About half (A: 3 groups; B: 2 or 3 groups; C: 2 groups) of the pre-service teachers show reasoning that is and that is not in line with the learning objective; or: 1 or 2 (A) or 1 (B and C) groups of pre-service teachers reason in line with the learning objectives, whereas there is insufficient evidence about the groups.
−	The reasoning of about half (A: 3 groups; B: 2 or 3 groups; C: 2 groups) of the pre-service teachers is not in line with the learning objectives, whereas there is insufficient evidence about the other pre-service teachers; or: 1 or 2 (A) or 1 (B and C) group shows reasoning that is in line with the learning objectives, while the reasoning of the other groups is in not line with the learning objective.
− −	The reasoning of all pre-service teachers is not in line with the learning objective, evidenced from misconceptions, or from no attention for the learning objective, where attention was expected.

Table 9.3 Conceptual mechanism and learning objectives for the growing samples activity

Component	Aspect	Aspect of the growing samples activity	Conceptual mechanism	Learning objective
Data as evidence	Data as evidence	PTs[a] are repeatedly asked to draw conclusions.	The repetition of the question creates awareness that empirical data can be used to base conclusions on.	PTs base their conclusions on evidence from the data, not on context-based claims, such as preconceived ideas.
	Center	PTs are asked to generalize to a population and to predict the shape of a larger sample or of the population.	To answer the questions to generalize or to predict, PTs use conceptual tools as data-based arguments. These tools are statistical measures of aspects of the sample distribution. One measure is the mean, which is a stable feature of the distribution. Another is the mode, which is not stable for smaller samples but can function as an indicator for the majority.	PTs compare samples by using measures of centrality that are familiar to them, such as the mean (see Sampling variability). They either estimate or calculate the mean. To predict, they use approximately the same mean, mode and global shape as in the smaller sample.
	Spread	PTs are asked to generalize and to predict based on samples with varying spreads.	When PTs are confronted with only one sample, they may not consider the spread. Presenting samples with different spreads creates awareness about the spread of the sample distributions.	PTs employ different statistical measures for the spread, such as the range, or informal statistical expressions for the spread, such as more spread out. These spread measures are used in the PTs' generalizations (see Heterogeneity) but also as a signal for the spread of the larger sample sizes.

(continued)

Table 9.3 (continued)

Component	Aspect	Aspect of the growing samples activity	Conceptual mechanism	Learning objective
	Distribution	PTs are asked to predict the shape of a larger sample or of the population.	The question to predict the shape of a larger sample helps PTs to focus on the general shape and on aggregate features of the distribution. When the samples sizes increase, PTs will notice that the shape of the distributions stabilizes (i.e. does not change much).	Based on visual inspection of the graphs, PTs describe the general shape of the sample distributions in informal statistical expressions, such as majority, minority, few, most are left, and positive. PTs reason about the general shape and about the aggregate features of the distribution. In the predictions, PTs sketch graphs with approximately the same shape as in the previous round.
Generalization beyond the data	Generalization	PTs are repeatedly asked to generalize to the population.	If the question to generalize is asked only once, PTs may describe only the sample at hand. Repetition will alert PTs to the inferential aspect of the question.	PTs make probabilistic generalizations.
	Prediction	PTs are asked to predict the shape of a larger sample or to the population.	The question itself obliges PTs to generalize beyond the data, since PTs are asked to expand the graph with additional data.	PTs generalize to a larger sample or to the population, copying the general shape of the smaller sample, with approximately the same mean and mode (see Center).
Probabilistic language	Sample size	PTs are asked to compare samples of different sizes.	When the sample sizes increases, PTs notice that the shape of the distribution does not change much from a certain sample size onward.	PTs conclude that larger sample sizes provide more certainty about their inferences, because some sample features appear to be stable.

(continued)

Table 9.3 (continued)

Component	Aspect	Aspect of the growing samples activity	Conceptual mechanism	Learning objective
	Heterogeneity	PTs are provided with samples with different spreads.	PTs compare samples where the data is spread differently about the center of the data and understand that sample distributions with large spreads provide less information about whether two population means differ.	PTs conclude that the morethe sample data are spread about the center of the data (see Spread), the more uncertain the generalization.
	Sampling variability	PTs are provided with samples of increasing size and are asked to generalize and to predict shape of a larger sample or of the population.	When the sample increases, PTs notice the s9.table features of distribution, such as the mean and mode, i.e., the shape of the sample distribution stabilizes. They will be confident that population distribution will have approximately the same characteristics as the sample distribution.	PTs identify features of the observed sample that are stable and features that are variable and use the stable features of sample distribution as arguments to generalize beyond the data.
	Certainty	PTs are provided with samples of increasing size and with different distributions.	If the question to generalize is asked only once, PTs may not consider the certainty of their conclusions. Using samples with different sizes and distributions draws their attention to variation in the sample distributions. They will notice that their conclusion depends on the sample distribution and will conclude that their inferences are inherently uncertain.	PTs include probabilistic statements in their generalizations and compare the certainty of their conclusions for different sample size s.

[a]PTs stands for pre-service teachers

Table 9.4 Results of the analysis of the answer sheets

Conclusion	Frequency
Data as evidence	
Yes	36
Other sources	1
No	8
Total	45
Descriptive statistics	
Global shape of distribution	6
Mean, median or sum	8
Spread	14
Local aspects of distribution	2
None	18
Total	48
Type of conclusion	
Descriptive	1
Unclear: Descriptive or inferential	22
Inferential	9
Probabilistic inferential	1
Refusal to generalize	1
None	11
Total	45
Uncertainty	
Probabilistic language	3
Sample size	3
None	39
Total	45
Prediction	Frequency
Type of prediction	
Shape smaller sample copied	26
Overemphasized conclusion	9
Shape smaller sample mimicked	7
Total	42

the 45 conclusions, the participants used the data as evidence, although not always explicitly. For example, they wrote down only: "men are on average more positive." We interpret the use of "average" as an indication that the average of a sample distribution was used as evidence. Only one group used another source as evidence in their conclusion. This group argued that women's decreased logical and spatial thinking ability influenced their attitude toward mathematics.

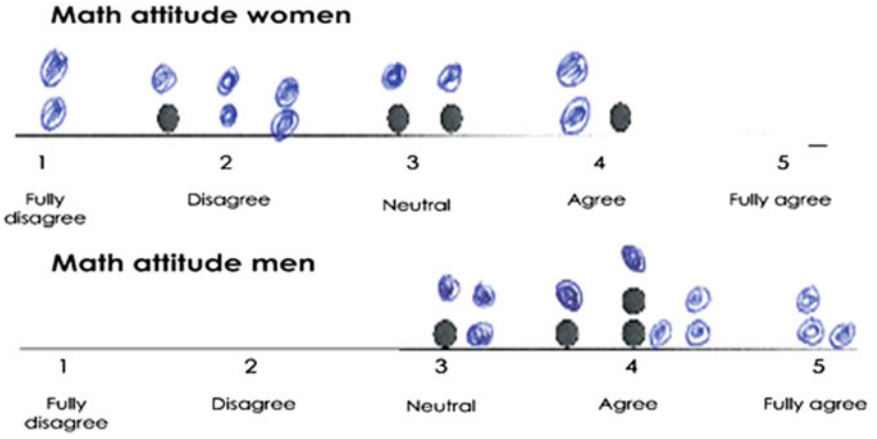

Fig. 9.2 Yasmine's and Esther's (Class B; both participants had a vocational education background) predictions of a larger sample that overemphasized the conclusion based on the small sample (pen writing constitutes the participants' predictions)

Concerning the descriptive statistics used as evidence, the participants often noticed the higher average of men and the high spread of women, but they did not connect the spread to their conclusion, as evidenced in this quote: "Many differences between men, few among women. Men love mathematics." Moreover, of the 34 answers that included conclusions only 18 were accompanied by a descriptive statistic as an argument, and only ten of these were supported by the mean or the global shape, which are suitable descriptive statistics to compare two distributions.

In the predictions, there were more indications that the participants used other sources of information. During the first round, six out of 11 groups overemphasized in their prediction the conclusion that men are more positive about mathematics than women, by moving the men's distribution to the right and the women's distribution to the left, as shown in Fig. 9.2.

Related to the learning objective Generalization, in 22 of the 34 answers that included a conclusion, it was unclear whether the conclusion pertained to the sample only or to the population. The following is a typical example: "Men are more positive about mathematics than women." Nine conclusions were coded as inferential because they included the words "in general." Only one group made a truly probabilistic generalization, by stating, "With more people, the conclusion we draw is more reliable, the attitude of men seems to be more positive."

Of the total of 42 predictions, 26 were plausible in the sense that they followed the global shape of the graph of the smaller samples. No group consistently smoothed the graphs, and only about half of the groups widened the range. Nine graphs mimicked the shape of the graphs of the smaller datasets, for example, by multiplying each frequency with the factor the sample increased, which is a very unlikely outcome (see Fig. 9.3). These results indicate that the participants did not understand that

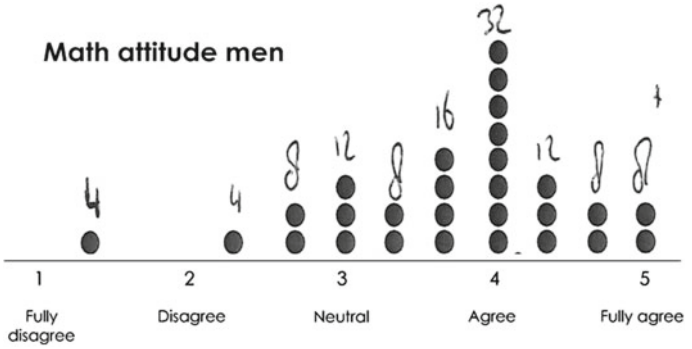

Fig. 9.3 Karel's and Nick's (Class C; other and vocational education background, respectively) predictions of a larger sample that mimics the smaller sample (pen writing constitutes the participants' predictions)

a sample distribution resembles the population distribution more when the sample increases.

On the answer sheets, uncertainty was mentioned by only one group. This group concluded that no inference could be made based on four men and four women, and they were also the only group that made a probabilistic generalization. The other groups never mentioned uncertainty, apart from one weak indication of probabilistic language in the prediction of a graph ("We expect…").

9.4.2 Class Discussions

Whereas the reasoning on the answer sheets was largely similar across the three classes, the class discussions differed in many respects. Therefore, we summarize these class discussions by class.

9.4.2.1 Discussion in Class A

In Class A, the first round started with several participants stating that male pre-service teachers are more positive about mathematics than female teachers. Cindy[1] objected, thus opening up the field to discuss ISI.

Cindy: Here they are more positive, but just for these four persons. […]
Teacher educator: So you say the sample is too small.
Cindy: It is just like you take four persons from a class and ask them what they think of the class. If you accidently pick four positive

[1] All the participants' names are pseudonyms.

people from the class, you get a very good picture [inaudible], while the four other people could not like it at all.

Teacher educator: Ok, so we agree with this: the sample is just too small to say anything sensible about.

Various: Yes.

Cindy argued that another sample could result in an entirely different outcome because of the small sample size. She thus used ideas from sampling variability to explain why she thought that generalization was not possible. At first, all participants agreed. However, Merel objected by referring to the lower spread in the men's distribution: "Well, I think, the men's graph because it is less spread out, does say something, I think." She was very tentative about her conclusion that this lower spread about the center of the data meant more certainty about the population distribution compared to the higher spread in the women's distribution, but the teacher educator confirmed the validity of her reasoning. Next, the averages for men and women were compared. After the teacher educator pointed at the high spread about the center of the data in the women's graph, Merel again stated that the women's average was less informative because of the higher spread.

Turning to the predictions of the shapes of graphs for the 15 men and women, Yanka explained her prediction and indicated she had made the women's distribution more negative than the distribution of the smaller sample (see Fig. 9.2 for a similar prediction). Stressing the difference between men and women suggests she used other sources of evidence than the sample data for her prediction. A few participants suggested adjusting the predictions to make them resemble more the global shape of the women's graph for the sample of four.

During the class discussions in the second and third rounds, most of the time was spent on discussing how one could estimate the means without calculations, leaving little time to discuss ISI. During the second round, there was some attention to ISI when the teacher educator asked whether the participants expected the range to widen in the prediction for the larger sample. The participants thought it would but without showing indications of having thought about why the range would widen. It is thus doubtful whether the participants understood that the sample distribution resembles the population distribution more closely as the sample increases.

The third round, with samples of 28 men and 116 women, was concluded by summarizing the arguments used. The teacher educator asked about the role of spread regarding the conclusions. Initially, the participants struggled to answer this question, but finally, one said less spread meant more certainty about the conclusion. However, Merel, who introduced this claim during the first round, and Nicky appeared to doubt whether there were any stable signals in the sample distribution. The latter indicated that another sample (sampled the next week) may result in entirely different distribution:

Merel: We just don't know.

Nicky: [inaudible] next week it is on 1 again. [i.e., they are very negative about mathematics.]

In summary, at the start of the class discussion of the first round, Cindy's remark that a sample of four is too small for generalizations provided the opportunity to discuss ISI. This discussion yielded the insight that a sample of four is too small for generalizations. In addition, the group seemed to generally agree that the data could be used as evidence. Apart from these insights, little attention was paid to ISI. Moreover, understanding of how a sample can be used to generalize seemed to be absent.

9.4.2.2 Discussion in Class B

In Class B, the class discussions during the first and second rounds were short; moreover, only half of the participants participated. It was also unclear whether the conclusions were meant to be descriptive or inferential.

During the first round, the teacher educator asked for the participants' opinion about the reliability of the generalization. Manon responded that four teachers "never reveal the opinion of all pre-service teachers. One needs many more subjects." None of the participants supported or objected to this statement. Next, Manon described the graph of the larger sample of women, which she made more negative, using her own ideas about men's and women's attitudes about mathematics: "There are women who find it very difficult […], and therefore, we made the women lower in the end." Again, the other participants did not participate in the discussion. During the second round, the group discussed the conclusion and the prediction only briefly.

During the third round, the discussion about the comparison of samples of unequal size yielded insights into the participants' conceptions of sampling variability. To compare the samples of 28 men and 116 women, some participants suggested multiplying the 28 men by 4 to make the sample sizes approximately equal (see Fig. 9.3). This deterministic approach, which neglects sampling variability, was challenged by other participants. For example, Rebecca stated, "Now you are estimating. You don't exactly know how men think and what they will fill in […]. But in the end you can't know anything about it."

Later on, Yente explained why in her view this strategy is permissible:

Yente: Suppose at once there are all men who score 5 points. What we look at is, if it would go like this. How we think that it went exactly, then it is no problem. But we never know how other people think about it.

As Rebecca, Yente did not understand that a sample can be representative of the population. However, to solve the problem of comparing samples of unequal sizes, Yente simply *assumed* that if other men were sampled, they would have exactly the same sample distribution. She seemed to be primarily concerned with how to compare samples of unequal sizes, not with the issue of generalization:

Yente: If you just ask the Netherlands, then you can simply compare a number of men and women, but because among pre-service teachers there are not so many men, then you need to kind of estimate, I think.

Doubts about whether a sample can be representative of a population were also visible in the remarks of Marleen: "Yes, we can estimate. There are 116 women. […] If there are suddenly ten more who all totally agree, then it is quite different from how it is now. It will always be estimation." When challenged by the teacher educator, Marleen changed her mind:

Teacher educator: If there are ten more who all totally agree, you said. [Draws imaginary bar of ten women who all totally agree.] Is that likely?
Marleen: No.

Later on, she returned to her original idea:

Teacher educator: And Marleen says that suppose there are 15 women very positive, or ten, we immediately say that is very unlikely.
Marleen: It could still be.

Whereas Yente switched from complete uncertainty to complete certainty, by assuming that other men had the same attitude as the men sampled, Marleen remained doubtful about what result one could expect in a different sample. Only Rebecca seemed to believe in the possibility of making generalizations: "That is the way they always do it, right? If they want to know something, they don't ask [inaudible; probably: everyone]."

Overall, except for Rebecca, the Class B participants were reluctant to accept the claim that a sample can be representative of a population. The class discussions showed that this reluctance was probably caused by their lack of understanding of sampling variability.

9.4.2.3 Discussion in Class C

In Class C, the first round was skipped because during the third round in Class A and B motivation seemed to decline. The second round started with the conclusion that men are more positive about mathematics than women. Whether this conclusion was meant to be descriptive or inferential initially was unclear. In an extensive discussion about the use of measures of centrality as arguments, the group concluded that neither the midpoint of the range nor the mode for non-normal distributions is useful in comparing distributions. Next, seven of the nine participants participated in a lively discussion about whether the conclusion would hold for the population. Initially, various participants denied this.

Teacher educator: If we look at these men… I am curious to know, can we say something about those 500 based on 15 men [500 was the assumed size of the men's population]?
Khloud: But this is not a good sample, is it?
Teacher educator: Why is this sample not good?
Khloud: Because there are way too few.

Later on, Khloud related the size of the sample to sampling variability: "Yes, at random, I understand, but maybe you by chance picked the best 15." Two other participants agreed with this statement, and one concluded that "in the end, one can never say something about it," which multiple participants agreed with. Then the teacher educator asked about the relationship between spread and certainty.

Teacher educator: About whose attitude do you have more certainty: about men or women?
(Almost) all: Men.

[…]

Karel: The chance is higher that they are all over there [points at the positive part of the graph] than they are not over there.

Karel claimed that it was more likely that another sample of men would also be predominantly positive about mathematics. Various participants agreed with Karel. The teacher educator confronted the participants with their conflicting opinions: Earlier they had said that nothing could be known, but then they indicated they had more certainty about men. Khloud's response illustrates how the participants appeared to combine these opinions: "We know the chance, but we are not sure." Laura concluded, "Because you did such a small sample, [inaudible] you never know for sure [inaudible] I still don't think it is a good sample." The participants wanted not only to have a larger sample but also seemed to want a sample that would give them complete certainty about their generalizations. Making generalizations with a certain degree of uncertainty appeared to be problematic for them.

During the third round, ways to compare samples of unequal sizes were discussed. Although, as in Class B, some participants suggested multiplying the 28 men by four (see Fig. 9.3), or, alternatively, dividing the 116 women by 4, three participants argued that one could use the mean to compare samples of unequal sizes because the shape of a sample is expected to remain approximately the same as the sample increases. In this discussion, Khloud showed a remarkably good understanding of the effect of sample size on sample-to-sample variability: "In any case, if you would take 15 people, then the chance is smaller it remains the same […] than if you have 28 people." Overall, a small majority of the participants indicated that they expected approximately the same results for a different sample, while none of the participants expressed the opposite.

In sum, the majority of the Class C participants seemed to understand that it is possible to make generalizations because the shape of a sample is expected to remain approximately the same as the sample increases. Comparing the spread and the certainty about the two distributions led the participants to express correct ideas about sampling variability. However, they displayed an inclination to demand complete certainty about the generalizations.

Table 9.5 Quality of reasoning about ISI per class

ISI component	Aspect	Class		
		A	B	C
Data as evidence	Data as evidence	+ +	+	+ +
	Center	+	−	+
	Spread	+	−	0
	Distribution	0	0	0
Generalization beyond the data	Generalization	0	− −	+
	Prediction	0	−	−
Probabilistic language	Sample size	−	− −	−
	Heterogeneity	−	− −	+
	Sampling variabilit y	−	− −	0
	Certainty	0	− −	+

Note Quality of the reasoning about ISI, ranging from − − (*reasoning not at all in line with learning objective*) to ++ (*reasoning entirely in line with learning objective*). See Table 9.2 for a detailed explanation of the indicators

9.4.3 *Quality of Reasoning About ISI*

Table 9.5 shows the quality of reasoning about ISI for each class. Indicators were assigned to the learning objectives that showed to what extent the pre-service teachers attained the learning objective (see Table 9.2). For most aspects, Class C's reasoning about ISI was the most sophisticated, although the first round was skipped in this class. Class B's reasoning was poor overall.

In all three classes, elaborate discussions about estimating the mean took place. To most participants, it seemed clear that a comparison of the means or the global shape of the sample distributions was a valid way to compare the distributions. However, although many participants noticed the high spread about the center of the data in the women's distribution, few understood that this made a generalization more uncertain. In Class A, two participants discussed the effect of the spread on the choice of the measure of centrality.

On the answer sheets, there was little evidence that the participants intentionally generalized to the population. Questioning by the teacher educator during the class discussions made generalization a topic of discussion. To accept the feasibility of making generalizations, an understanding of sampling variability proved vital. In Class B, all but one of the participants thought nothing could be known about the population as a whole, since another sample could differ significantly from the current sample. In Classes A and C, most participants seemed to understand that making uncertain generalizations is possible. Class A's reasoning was superficial. Most of the participants acknowledged the uncertainty of generalizations and that a sample of

four is too small for any generalization, but what could be stated about a population based on a sample was not discussed. Class C's understanding of generalizations was more sophisticated. They agreed that less spread means more certainty. However, they demanded complete certainty about the generalizations, dismissing samples that could not provide this complete certainty.

The participants' predictions of the distributions of larger samples and the population revealed many did not understand that sample distributions would resemble the population distribution when the sample increases. Classes B and C made many predictions that exactly copied the shape of the distribution of the smaller sample. In Class C, discussing the predictions was only a small part of the discussions, which might partly explain the low quality of reasoning about predictions in this class.

Understanding of the learning objectives of the Probabilistic language component differed among the three classes. Even in the best-performing class, Class C, the ISI reasoning was not as sophisticated as expected. In all classes, the answer sheets almost completely lacked attention to uncertainty. Class B's reasoning about uncertainty was the least developed. Although there was broad consensus that for the sample of four generalization was impossible, other uncertainty aspects were not discussed. The participants did not use probabilistic language, except for the opposite, certainty language, for example, in statements, such as "It can never be the case…" In addition, the Class A participants agreed that for the sample of four, generalization is impossible. Moreover, the majority of this class seemed to understand the possibility of making uncertain generalizations. Only Class C had an intense class discussion about the extent to which a sample may provide information about other people not interviewed. Even there, however, the majority was not convinced that a larger sample would look like the smaller sample. Only a minority could express the idea that the shape of a sample is likely to remain approximately the same when the sample increases, provided the sample is sufficiently large. The participants in Class C regularly used uncertainty language during the class discussions.

9.5 Conclusions and Discussion

This study explored the growing samples heuristic in the context of teacher education. We investigated how three classes of first-year pre-service primary education teachers reasoned about ISI when engaged in a growing samples activity. The results show that in two classes most seemed to agree that making (uncertain) generalizations based on a sample is possible. However, overall, the majority was unable to link the possibility of making generalizations to an understanding of how a well-selected sample can be representative of the population.

Concerning the way descriptive statistics were used as arguments in ISI, the class discussions revealed that most pre-service teachers implicitly used suitable descriptive statistics to compare two distributions. On the answer sheets, however, only a third of the conclusions were supported by suitable descriptive statistics. In particular in Rounds 1 and 2 it could easily be seen, without calculation, that men were on aver-

age more positive about mathematics than women. The difference may have been too obvious to induce the pre-service teachers to write down descriptive statistics.

On the answer sheets, most conclusions seemed to describe the sample data only, rather than generalize beyond the sample data. Inferential statements used at best the colloquial term 'in general,' which could have been copied from the question without the intention to generalize. The first explanation for this finding may be that the need to generalize was not compelling enough. Another explanation could be that the participants, in their role as future teachers, had a class of primary school students in mind as their population of interest. When the class is the population of interest, description suffices, and there is less inclination to generalize beyond this class. The questioning by the teacher educator during the class discussions was necessary to draw the pre-service teachers' attention to the inferential nature of the questions.

Attention to uncertainty and sample size was virtually absent on the participants' answer sheets. This underlines our conclusion that most of the pre-service teachers described only the sample data. Description does not require uncertainty and sample size to be taken into account. During the class discussion in Class B, the majority of the pre-service teachers concluded that generalization is impossible because they accepted the claim that nothing can be known about people who are not in the sample. This resembles the instance found by Ben-Zvi et al. (2012) of students uttering complete uncertainty. In Classes A and C, in contrast, the majority of the participants seemed to acknowledge that making uncertain generalizations based on a sample is possible. This finding is similar to the findings in De Vetten et al. (2018) and the reasoning displayed by high-ability middle-grade students studied by Ben-Zvi et al. (2012).

We found little evidence that the heuristics helped the pre-service teachers to understand the concept of sampling variability, contrary to the ideas formulated by Joan Garfield and Ben-Zvi (2008). Only one participant attempted to explain the stability of sample distributions when the sample increases by referring to probability theory. Her explanation did not convince the other pre-service teachers. The predictions of the distribution of larger samples and the population provided extra evidence for this finding, since often these predictions too strictly followed the global shape of the sample at hand. Not understanding sampling variability is problematic because it seemed to make the participants reluctant to accept the possibility of making generalizations. One reason why the activity did not foster an understanding of sampling variability could be that for a given sample size, all groups received the same data set. Understanding why the sample distributions become stable when the sample increases might require a repeated samples approach, where each group receives a different data set and compares their conclusions with other groups.

In Class C, only the second and third rounds were used, but their generalization and sampling variability were the most sophisticated. The questioning by the teacher educator appeared a more effective way to foster reasoning about these topics than repeatedly asking the participants to generalize and make predictions. The results also raise questions related to the optimal number of rounds, the sample size of the first round, and the effect.

These results show some benefits of the growing samples heuristic in general and our operationalization in particular. First, the heuristic helped to initiate discussions about the role of sample size in certainty and sampling variability, which are key concepts in ISI. In addition, using sample distributions with different variabilities seemed to have helped the participants to gain insight into the certainty of generalizations. Second, the activity was useful for discussing many distributional aspects, such as measures of centrality and the effect of spread on measures of centrality, as was the case in Bakker's (2004) study. However, discussing the calculation and estimation of descriptive statistics took considerable class time, which could have been spent more productively on how one can use descriptive statistics as arguments in ISI, for example, what the spread of the sample distributions implies for the conclusions. Third, the use of samples of unequal sizes during the third round initiated discussions about using the measure of centrality in comparing different sample sizes and about sampling variability.

The participants' educational background could have played a role in the different quality of reasoning about ISI between the classes. This background is clearly different for Class B, compared to the other two classes. In Class B, nine of the 11 participants have a background in secondary vocational education but only seven out of 15 in Class A and four out of eight in Class C. Since statistics is not part of most secondary vocational education curricula but is part of most senior general secondary education curricula, the pre-service teachers with a background in senior general secondary education may have had the vocabulary and statistical tools to further the reasoning about ISI during the class discussions. We found some evidence for this explanation. One pre-service teacher in Class C explicitly stated, "I had something about this in secondary school." Moreover, one pre-service teacher with a background in senior general secondary education introduced the term "probability theory", after which probability and chance became terms used to reason about sampling theory. For a fruitful ISI discussion, a fair number of pre-service teachers with appropriate background knowledge in statistics and probability theory may need to be present.

Some issues warrant a cautious interpretation of the results. First, this was a small-scale and explorative study, and the context was the Dutch educational system where students enter teacher college immediately after secondary education. The results are, accordingly, not readily generalizable to other contexts. However, similar processes may occur in countries where students enter teacher college with similar backgrounds and where the statistics curriculum in primary and secondary education is comparable to the Dutch system. Second, the design of the activity likely influenced the reasoning. For example, the sample distributions may have influenced the results. In particular, the data did not result in conflicting conclusions. During each round, it was quite obvious that men were more positive about mathematics than women. Third, sound recordings of the pre-service teachers when working in groups were not available. Issues spoken about but not written down could have provided a more complete picture of the participants' reasoning, in particular about whether they had spoken about generalization and uncertainty but had not written down these issues. Nonetheless, the extensive class discussions of generalization and uncertainty

probably provide a reliable general impression of the pre-service teachers' reasoning about ISI.

In conclusion, this study informs how the effectiveness of the heuristic can be further strengthened. First, the pre-service teachers seemed to use correct descriptive statistics as arguments in ISI. This finding indicates less focus might be given to descriptive statistics and by using simple descriptive statistics, more on ISI itself. Second, some pre-service teachers were reluctant to accept the possibility of making generalizations beyond the data. Comprehension of this fundamental idea may be fostered if each group uses a different data set. When the sample sizes increase, the different data sets typically will begin to resemble each other, leading to confidence on the learner's behalf that from a certain sample size onward, a sample provides reliable information about the population. Finally, because the pre-service teachers tended to describe the data only, the need to make generalizations beyond the data was not sufficiently compelling. Therefore, we recommend designing activities and contexts in which description is clearly insufficient and where generalization beyond the data is natural and inevitable. These changes to the growing samples heuristic may help to provide pre-service teachers the knowledge to demonstrate to primary school students the feasibility of making generalizations beyond the data.

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*(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., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM—Mathematics Education, 44*(7), 913–925.

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

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Young students reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Proceedings of the Fifth International Forum for Research on Statistical Reasoning, Thinking and Literacy (SRTL-5)*. Warwick, UK: University of Warwick.

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.

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.

Cobb, P., & Tzou, C. (2009). Supporting students' learning about data creation. In W.-M. Roth (Ed.), *Mathematical representation at the interface of body and culture* (pp. 135–171). Charlotte, NC: IAP.

De Vetten, A., Schoonenboom, J., Keijzer, R., & Van Oers, B. (2018). Pre-service primary school teachers' knowledge of informal statistical inference. *Journal of Mathematics Teacher Education*. https://doi.org/10.1007/s10857-018-9403-9.

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.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer.

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics, 88*(3), 327–342.

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.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. 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. 235–246). Dordrecht, The Netherlands: Springer.

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.

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.

Leavy, A. M. (2006). Using data comparison to support a focus on distribution: Examining preservice teacher's understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal, 5*(2), 89–114.

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

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

Makar, K., & Rubin, A. (2014). *Informal statistical inference revisited.* Paper presented at the Ninth International Conference on Teaching Statistics (ICOTS 9), Flagstaff, AZ.

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. In K. Makar, B. De Sousa, & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. Voorburg, The Netherlands: International Statistical Institute.

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

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. London, UK: Temple Smith.

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

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.

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.