Educ Stud Math (2018) 99:217–234
https://doi.org/10.1007/s10649-018-9823-6

The development of informal statistical inference content knowledge of pre-service primary school teachers during a teacher college intervention

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

Published online: 8 July 2018

Abstract Teachers who engage primary school students in informal statistical inference (ISI) must themselves have good content knowledge of ISI (ISI-CK). However, little is known about how college education for pre-service teachers can contribute to the development of their ISI-CK. To address this shortcoming, we used a case study to investigate ISI-CK development in a class of 21 pre-service primary school teachers who participated in a short intervention (180 min). Based on qualitative and quantitative analyses of the pretest, posttest and intervention data, the results suggest that most participants acknowledged it is possible to make uncertain inferences. An assignment to search the media for inferential claims seemed to create awareness regarding inference and the need to distinguish between a sample and a population. A simulation involving random sampling and varied sample size probably increased the participants' knowledge of sampling variability and random sampling. No development was seen in the participants' knowledge about sufficient sample sizes. The statistical investigation conducted by the participants during a model lesson may have strengthened their awareness of ISI, but it also revealed that many participants continued to favour distributed sampling over random sampling. Further research on belief formation with regard to data as evidence, sampling methods and the expression of uncertainty in the context of ISI is needed.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10649-018-9823-6) contains supplementary material, which is available to authorized users.

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

[1] Vrije Universiteit Amsterdam, Section of Educational Sciences, LEARN! Research Institute, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

[2] Faculty of Social Sciences, Institute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

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

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

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

1 Introduction

In today's society, it is increasingly important to be able to reason inferentially (Liu & Grusky, 2013). One form of inferential reasoning is statistical inference, defined as "a generalized conclusion expressed with uncertainty and evidenced by, yet extending beyond, available data" (Ben-Zvi, Bakker, & Makar, 2015, p. 293). Two types of statistical inference can be distinguished. The first is formal statistical inference, which uses formal statistical tests based on probability theory. This type is usually considered out of reach for primary school students. The second is informal statistical inference (ISI). The statistical reasoning involved in ISI is of lower complexity than in formal statistical inference. For example, ISI allows for qualitative instead of quantitative expressions of uncertainty and for inferences based on simulations instead of on closed-form formulas (Makar & Rubin, 2018). Evidence suggests ISI can be made accessible to primary school students (Meletiou-Mavrotheris & Paparistodemou, 2015; Watson & English, 2016). Presumably, if students are familiarized with ISI in primary school, they may understand the processes involved in ISI reasoning and in statistical reasoning in general (Bakker & Derry, 2011; Makar, Bakker, & Ben-Zvi, 2011).

If primary school students are to be introduced to ISI, their future teachers must be well prepared to conduct this introduction (Batanero & Díaz, 2010). This requires them to have appropriate content knowledge (CK) of ISI (Groth & Meletiou-Mavrotheris, 2018) that must extend beyond what their students will learn (Ball, Thames, & Phelps, 2008; Fennema & Franke, 1992; Pfannkuch & Ben-Zvi, 2011). However, many students enter tertiary education in general (Chance, delMas, & Garfield, 2004) and teacher college in particular (De Vetten, Schoonenboom, Keijzer, & Van Oers, 2018b) with a shallow, isolated understanding of the concepts underlying statistical inference. Many pre-service teachers have difficulty making inferences and lack understanding of representativeness and sampling variability (De Vetten, Schoonenboom, Keijzer, & Van Oers, 2018a; De Vetten et al., 2018b; Groth & Meletiou-Mavrotheris, 2018).

In many countries, including the Netherlands, primary education teacher college curricula spent usually little time on statistics. The research literature does not provide examples of interventional studies that show how to foster in a limited time frame the ISI-CK of pre-service teachers with limited knowledge of this topic at the onset of the intervention (De Vetten et al., 2018a; Groth & Meletiou-Mavrotheris, 2018; Leavy, 2010). Therefore, the aim of the present work is to study the development of the ISI-CK of pre-service primary school teachers with limited ISI-CK in an intervention of limited length. The research question is: "In what respect does the ISI-CK of pre-service primary school teachers develop during a teacher college intervention, and what role do the activities used during the intervention play in this development?" The intervention specifically aimed to make the pre-service teachers' attentive to the issue of inference and to provide them with sufficient ISI-CK to introduce primary school students to ISI.

2 Theoretical background

The Mathematical Knowledge for Teaching model by Ball et al. (2008), which is used in mathematics education in Dutch primary teacher education colleges (Van Zanten, 2010),

distinguishes between CK and pedagogical content knowledge (PCK), which is knowledge of how to teach specific content (Shulman, 1986). Because teachers' CK impacts their students' learning achievements (Fennema & Franke, 1992; Rivkin, Hanushek, & Kain, 2005) and may facilitate the development of their PCK (Groth, 2013), teachers need to possess a thorough knowledge of the content they teach, and this must extend beyond what their pupils will learn (Ball et al., 2008). These relationships are also shown to hold in the context of ISI (Burgess, 2009; Leavy, 2010).

For our study among pre-service teachers, we used the Makar and Rubin (2009) ISI framework and conceptualized ISI-CK, as follows:

1. Data as evidence: The inference is based on available data and not on tradition, personal beliefs or personal experience.
2. Generalization beyond the data: The inference goes beyond a description of the sample data by making a probabilistic claim about a population or a mechanism that produced the sample data.
3. Probabilistic language: Due to sampling variability and the degree of sample representativeness, the inference is inherently uncertain and requires using probabilistic language. For the correct usage of probabilistic language, the origins of uncertainty in inferences must be understood. Therefore, we divided this component into four subcomponents:

    a. Sampling variability: The inference is based on an understanding of sampling variability; it is expressed from an understanding that the outcomes of representative samples are similar and that therefore under certain circumstances, a sample can be used for an inference (De Vetten et al., 2018a; Saldanha & Thompson, 2007).
    b. Sampling method: The inference includes a discussion of the sampling method and the implications for the sample representativeness.
    c. Sample size: The inference includes a discussion of the sample size and the implications for the sample representativeness.
    d. Uncertainty: 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.

Previous research suggests there is a need to develop (pre-service) primary school teachers' ISI-CK, as many pre-service teachers have limited knowledge of sampling variability, sampling methods, sample size and representativeness (Canada, 2006; De Vetten et al., 2018a, 2018b; Meletiou-Mavrotheris, Kleanthous, & Paparistodemou, 2014; Mooney, Duni, VanMeenen, & Langrall, 2014; Watson, 2001). Furthermore, they lack awareness that ISI tasks require an inference over and above a descriptive analysis of the data (De Vetten et al., 2018a, 2018b). Mixed results were found regarding the extent to which pre-service teachers acknowledge the value of data as evidence and the possibility of using a sample to make (probabilistic) inferences (De Vetten et al., 2018a, 2018b).

Very few studies have investigated how to foster pre-service teachers' ISI-CK (Ben-Zvi et al., 2015). Leavy (2010) worked with pre-service teachers whose ISI-CK was already high at the onset of the intervention. De Vetten et al. (2018a) found that when engaged in a growing samples activity (Bakker, 2004), Dutch first-year pre-service teachers did not develop their understanding of sampling variability and restricted their attention to descriptive statistics rather than using these descriptive statistics as arguments in making inferences. The authors

recommend designing activities that stimulate an awareness of the inference required in the activities. Although working with high school students, Chance et al. (2004) and Saldanha and Thompson (2002) recommend having learners repeat the sampling process and compare multiple sample results to foster an understanding of the sample process. Other recommendations for the design of the intervention would be to take an approach that integrates CK and PCK (Groth, 2017), have learners conduct statistical investigations themselves (Garfield & Ben-Zvi, 2008), use hands-on activities (Canada, 2006; Pratt & Kazak, 2018), use simulation activities to illustrate sampling variability (Garfield & Ben-Zvi, 2008; Mills, 2002) and keep descriptive analyses as simple as possible to facilitate focusing on the relevant concepts to be learned (Arnold, Pfannkuch, Wild, Regan, & Budgett, 2011).

3 Method

3.1 Context

The intervention was part of a mathematics education course for pre-service primary school teachers. During this course, the pre-service teachers worked in grade 3 to 6 classes in work placement schools. In contrast to other countries where students enrol in teacher education after completing a bachelor's degree, in the Netherlands, teacher education starts immediately after secondary school and leads to the attainment of a vocational bachelor's degree in primary education. For these students, teaching mathematics is seldom their main motive for becoming teachers. The intervention was intended to have ecological validity for, and thus be useful in, the regular Dutch teacher college mathematics curriculum. As little time is usually spent on statistics in this curriculum, we designed a relatively short intervention (see the Intervention subsection). The intervention occurred in the second year of the teacher college curriculum to prepare pre-service teachers for the mathematics knowledge base test in their third year of study, while at the same time taking advantage of the participants' pedagogical skills acquired during their first year of study. The study design was approved by the ethical board of the Faculty of Behavioural and Movement Sciences of Vrije Universiteit Amsterdam.

3.2 Participants

One class of 21 second-year pre-service teachers participated in this study. This class was part of a small teacher college for primary education in a large city in the Netherlands. All participants had encountered some basic descriptive statistics during their first year of study. The average age of the participants was 20.95 years (SD 2.19); six were male. Ten participants had a background in secondary vocational education, where statistics is usually not part of the curriculum. Nine participants had senior general secondary or university preparatory education, and about half of these studied descriptive statistics as part of their mathematics courses. The educational background of the remaining two participants was something else or was unknown. The first author was the teacher educator. A second observer was present during the sessions, and all analyses were discussed with an external researcher. The teacher educator had four years' experience as a mathematics teacher educator, experience as a university statistics lecturer and had taught most of the participants during their first year of study.

3.3 Intervention

The intervention described in this paper was part of a larger intervention that consisted of five of the 16 sessions of the entire mathematics education course. The focus here is on the first two and a half sessions during which the emphasis was on ISI-CK. During the last two and a half sessions, the emphasis was on preparing to teach an ISI lesson in the participants' placement schools and on evaluating these lessons. The analysis of these sessions is beyond the scope of this paper, although occasional evidence relating to CK found in the second part of the intervention has been included. Below and in Table 1, the activities that were intended to support the development of the participants' ISI-CK are described. Dialogic classroom talk was used throughout the intervention to encourage dialogic inquiry and to scaffold the participants' learning (Wells, 1999). Based on our ISI framework, 12 learning objectives were formulated (see Table 2).

3.3.1 Homework assignment: samples in the media

The first activity was aimed at creating awareness of the use of inferential reasoning in the media and of the distinction between sample and population and at initiating discussions about uncertainty, sampling methods and sample size. Before the first session, the participants completed a homework assignment. They were asked to search for a news item that made a claim about a population based on a sample, to describe how the conclusions were reached and to write a critical evaluation of the quality of the research. During the first session, the participants discussed in groups of three or four any errors they had identified in the news items and answered questions about appropriate sampling methods, sample sizes and the certainty of inferences. This was followed by a class discussion.

3.3.2 Simulation

During the second half of the first session, the teacher educator used a real-time computer simulation (Van Blokland & Van de Giessen, 2016) to explain that when random sampling is used, the law of large numbers applies. By simulating samples of increasing size (100, 1000 and 10,000), it was shown that the sample distributions of multiple samples become more similar with increasing size. This simulation also aimed to foster a focus on comparing the

Table 1 Overview of the activities

	Activity	Setting	Related to learning objectives
1	Homework assignment samples in the media	Homework Session 1 (60 min): discussion of homework	1–3, 6–8, 10–11
2	Simulation	Session 1 (20 min): real-time computer simulation Session 2 (10 min): reiteration of learning points	3, 5–7, 9–12
3	Model lesson	Session 2 (70 min)	All, except 4 and 9
4	Car choice activity	Session 3 (20 min)	1, 4, 11

Table 2 ISI learning objectives for the intervention

ISI component		Learning objectives—the pre-service teachers	
Data as evidence		1	Use the data as evidence for a conclusion instead of other sources, such as their own experience, own beliefs or general opinion.
Generalization beyond the data		2	Are aware when a task requires an inference.
		3	Know that it is possible to use a sample to make general claims about the population.
		4	Know that generally not each possible outcome of a random process has equal probability of occurring (equiprobability bias, Lecoutre, 1992).
Probabilistic language	Sampling variability	5	Understand that when a sample is relatively large (e.g. 1000) and randomly selected, the probability is small that another similar sample will give an entirely different result.
	Sampling method	6	Know that random sampling is an appropriate method to obtain a sample.
		7	Prefer random sampling over distributed sampling (i.e. purposefully selecting individuals to obtain a distributed sample across critical population characteristics (Watson & Moritz, 2000a)).
		8	Know that convenience sampling, such as sampling one's own class, is an inappropriate sampling method to obtain a representative sample.
		9	Understand why an appropriate sampling method yields a sample in which aggregate characteristics are close approximates of the population characteristics.
	Sample size	10	Know what sufficient sample sizes are in different contexts and understand why this is the case (e.g. understand why a sample size of 1000 is a sufficient sample size for the entire Dutch population of 17 million people).
	Uncertainty	11	Acknowledge the uncertainty of inferences and the impossibility of making absolutely certain inferences.
		12	Know that larger samples are more likely to yield precise estimates of the population parameters.

various sample distributions (Saldanha & Thompson, 2002). At the start of the next session, the learning points from the simulation and the main ISI concepts were discussed.

3.3.3 Model lesson

During the second session, the teacher educator taught a model lesson that involved a statistical investigation with hands-on activities that the participants could use in their placement schools. The lesson centred on a large pile of Dutch children's novels, and the driving question was which word would be used most frequently in the pile of books. The enormity and visibility of the population was expected to elicit the need to draw a sample and to make inferences. The analysis of the sample data was kept simple, as only frequencies needed to be counted. In this way, ample time was left for discussing ISI. The participants worked in groups of three or four to form a hypothesis about the most frequently used words based on logical thinking and their own experience. After the class reached a consensus about the top five most likely words, the

same groups designed a sampling method to be used employing the knowledge about sampling methods gained from the previous activities. A class discussion was used to reach consensus about the preferred sampling method so that sample data of separate groups could be pooled into one large sample. Next, the groups conducted an investigation using the agreed sampling method. They wrote down their answer to the driving question, their level of certainty and ideas regarding ways to increase the certainty of their inference. By pooling the sample data and comparing group results, a discussion about sampling variability was elicited to foster a distributional view on sampling (Saldanha & Thompson, 2002).

3.3.4 Car choice activity

During the third session, the equiprobability bias was discussed to increase the participants' own understanding of this bias and to explain its prevalence among primary school students using an adaptation of the car choice task by Watson and Moritz (2000b) (see Fig. 1).

3.4 Data collection and analysis

Development in ISI-CK was defined as observable behaviour becoming more in line with the learning objectives. A thematic analysis (Braun & Clarke, 2006) was used to measure the participants' ISI-CK development. In this analysis, the results from a pretest, an identical posttest and the intervention data were first analysed separately and then combined. The tests consisted of open-ended questions and statements; both data sources were used to provide quantitative overviews of the strategies employed. To gain a deeper understanding of the participants' ISI-CK during the intervention, the qualitative intervention data (video, audio, written work and notes) were analysed and summarized. The intervention data were also used to evaluate the possible role of the activities by identifying critical moments where a change was evident in the participants' ISI-CK before and after this moment.

3.4.1 Data collection

The pretest and posttest (see the online Supplementary material) piloted two tasks, adapted from the questionnaire used in De Vetten et al. (2018b). Both tasks started with an open-ended

Mrs. El Yakoubi wants to buy a new car, either a Peugeot or a Citroën. But first she wants to know which car will break down the least. First, she reads on the internet a research report by the Dutch Motorway Association, which has tested 400 cars of each type. In this report, the Citroën had more breakdowns than the Peugeot. Then she talks to three friends. Two are Citroën owners, neither of whom has experienced major breakdowns. The other friend used to own a Peugeot, but it had many breakdowns and so she sold it. She says she'd never buy another Peugeot.

Which car should Mrs. El Yakoubi buy?

a. Mrs. El Yakoubi should buy the Citroën because her friend had so much trouble with her Peugeot, while her other friends have had no trouble with their Citroëns.
b. She should buy the Peugeot because the Dutch Motorway Association has looked at many cars, not just one or two.
c. It doesn't matter which car she buys. Whichever type she gets, she could still be unlucky and get stuck with a particular car that needs a lot of repairs.

Fig. 1 Car choice activity

question. Next, participants were asked to evaluate the correctness of fictitious statements of primary school students concerning the same task and to explain how these students might have reasoned to probe for additional knowledge. Task 1 investigated the selection of a representative sample. In task 2, inspired by Zieffler, Garfield, delMas, and Reading (2008), participants were asked to compare two sample distributions and to generalize from these samples.

The test was conducted using cognitive interviews with four pre-service teachers from other classes. The pretest was then administered digitally during the session preceding the first intervention session; the posttest was administered during the session after the last intervention session. All 21 participants took the pretest; the posttest was completed by 16 participants. The pretest results of the five participants who, due to absence or lack of motivation, did not complete the posttest were excluded from the analysis.

During the sessions, whole class interactions (145 min) were video- and audio-recorded, while most of the group interactions were audio-recorded (35 min per group). Written work was also collected. One of the co-authors was present as observer. The observer's and the teacher educator's notes were used to triangulate the findings.

3.4.2 Data analysis

All video and audio data were transcribed literally. The transcripts and open-ended responses from the tests were coded using a process consisting of deductive and inductive elements. On the deductive side, the ISI framework was used to categorize the text data into one or more ISI components. On the inductive side, codes that were short summaries of the text were attached to the text to describe the exact meaning. These codes were subsequently combined into codes with similar meanings or on closely related issues. Participants whose results were difficult to interpret were asked to comment on our interpretation of their data (Torrance, 2012). All results were discussed with an external researcher until consensus was reached about the results' validity. Atlas.ti and Excel were used for data analysis.

The results of the pretest and posttest were based on information from the 16 participants who completed both tests. The coded open-ended responses from the pretest and the posttest were used to summarize the main strategies employed. For each fictitious statement, the number of participants who evaluated the statement correctly was calculated.

The results of the intervention were based on information derived from all 21 participants. To trace what ISI-CK the participants displayed at particular moments during the activities, each activity was divided into several parts, such as group and class discussions. These parts, 18 in total, were analysed separately. In addition, the activities in the second half of the intervention, where the focus was on PCK, were analysed. For each part and each ISI component, a tabulated overview of the codes was used to summarize the main results. For the group discussions, the summaries were aggregations of the individual groups' results. Using the main results from all 18 parts, we described the development of ISI-CK for each component over the course of the intervention.

4 Results

For each ISI component, we first present the results of the tests (see Table 3). We then describe the participants' ISI-CK during the intervention and the possible role of the activities used

Table 3 Pre-service teachers' ISI-CK demonstrated on pretest and posttest ($n = 16$)

ISI component		Learning objective	Open-ended response or statement		Pretest[a]	Posttest
Data as evidence		1	2 Open-ended: use of data as evidence		15	16
		1	2.1[b] General opinion is not valid evidence		15	16
Generalization beyond the data		2	2 Open-ended: awareness task requires inference		1	1
		3	2.2 Generalization is possible		12	14
		4	2.4 Understands misconception in equiprobability bias		1	0
Probabilistic language	Sampling variability	5	1.5 It is unlikely that another large random sample gives entirely different results		8	12
	Sampling method	6–8	1 Open-ended: preferred sampling method	Random	2	3
				Distributed[c] sampling	10	13
				Other/none	4	0
		6	1.2 Random sampling is possible		9	11
		7	1.6 Distributed sampling not representative		1	3
		8	2.6 Convenience not representative		11	16
		9	1.1 Understanding of controlling external factors		2	6
	Sample size	10	1 Open-ended: remarks related to sample size	1000 is not a sample	0	1
				2000 to 4000 sufficient	0	2
				sample size depends on population size	0	3
		10	1.3 40 is insufficient		12	13
		10	1.4 10,000 is not necessary		11	14
	Uncertainty	11	2 Open-ended: awareness of uncertainty		1	1
		11	2.5 Complete certainty impossible		16	15
		12	1.5 Larger sample, more precise estimates		15	14

[a] Number of participants who gave the specified response to an open-ended question or correctly evaluated a statement

[b] Second task, first statement

[c] Distributed sampling: purposefully selecting individuals to obtain a distributed sample across critical population characteristics (see Watson & Moritz, 2000a)

during the intervention in the development of the participants' ISI-CK. The conclusion combines both data sources to show to what extent the participants' ISI-CK was in line with the learning objectives and summarizes the role of the activities.

4.1 Data as evidence

Pretest and posttest In both tests, (almost) all participants agreed that supposedly commonly held beliefs are not valid evidence for a conclusion and used descriptive statistics to compare two sample distributions, which signals that they used the data as evidence for their answers.

Intervention Overall, the participants valued data as evidence. During the model lesson, most groups based their conclusions on the data, and at various points, several participants stressed the importance of the quality of the research conducted by research institutes. However, during the model lesson, some participants combined the data with other sources of evidence in making inferences, probably at the expense of treating the data as evidence. They combined arguments relating to the data, such as sample size and variations in the sample distribution, with arguments

based on other sources, such as results found on the web and the participants' own knowledge, even though these sources pertained to different populations, such as adult books. These participants seemed to accept the outcome of the class investigation because it did not conflict with their own knowledge. As Astrid (all names are pseudonyms) stated:

> But that [the acceptance of the outcome of the class investigation] might be because we had the information in the back of our minds. Yes, I don't know, I immediately thought the word "the" because I once heard it on the news or so, that "the" is the most frequently used word. … So then I automatically think, yeah, that's what I heard and then we got it from our small test and then you think: OK, that's correct.

Conclusion In the pretest and the posttest, participants valued data as evidence. This was confirmed during the intervention, but during the model lesson, there was also some evidence that participants based their inferences on a combination of sources of evidence at the expense of relying on the data.

4.2 Generalization beyond the data

Pretest and posttest In the pretest, 12 participants acknowledged that making generalizations is possible; this increased to 14 in the posttest. In both tests, only one participant was aware the second task required an inference. In the pretest, one participant noticed the misconception in the equiprobability bias, and in the posttest, no participants noticed.

Intervention Almost all participants agreed it is possible to make generalizations based on a sample. This was evidenced during the discussion of the homework assignment when the participants indicated that it was not necessary to sample the entire population.

From the start of the intervention onwards, the participants showed awareness that the activities required an inference because they discussed issues that presupposed this awareness. For example, when discussing the results of their investigations during the model lesson, all groups discussed the representativeness of the sample used. This awareness may have been raised by the homework assignment, as it explicitly distinguished between sample and population. For this assignment, 14 of the 19 participants who handed in the assignment paid attention to the inferential dimension, for instance, by referring to the quality of the sample used. The attention to inference might have been further triggered by the model lesson in which both the population (the pile of books) and the sample (the sampled books) were tangible and visible.

The discussion around the car choice task showed that many participants acknowledged that the chance of defects may differ between brands. However, apart from one participant, none applied the chance argument to one specific car. Alfred's conclusion is illustrative for many participants. He said that on the one hand, "in general one can also have just bad luck," but on the other hand "one still should look at [research]." Consequently, while most participants valued the results of research, they did not use these results to predict the outcome of an individual case.

Conclusion We found only a minimal change with respect to the possibility of making generalizations between the pretest and the posttest. The only two participants who during the posttest still denied this possibility were the two who denied this during the first session.

Throughout the intervention, the participants were aware the activities required an inference, their awareness probably sparked by the homework assignment. This awareness was not seen in the tests, as only one participant noticed the second task required an inference. Only one participant was able to use chance arguments to make predictions about an individual case and thus showed an understanding of the misconception in the equiprobability bias.

4.3 Sampling variability

Pretest and posttest The number of participants who understood sampling variability increased from eight in the pretest to 12 in the posttest.

Intervention At the beginning of the intervention, various participants struggled with the issue of sampling variability. As Menthe wrote in her homework assignment, "How can 1,082 people represent what all 17 million Dutch people have in mind?" The simulation seemed to have been crucial in changing participants' conceptions about this topic. By showing Menthe's quote to the other participants, the teacher educator brought their struggle to the fore, thus clarifying the issue in question and motivating the participants to learn from the simulation, as was evidenced from the (ironic) remark of one of the participants: "Finally, an answer to our questions."

At various points, participants indicated that the simulation was clear, and during the recap in the next lesson, six participants correctly explained that larger samples resemble each other more than smaller samples. As Astrid stated:

> At a certain moment, there are … not so large differences. For example, with a dice, … if you throw a hundred times, then you can still see that four is thrown many times, while three is not. But from a certain number, ehm, 1,000, 2,000, 3,000 ehm, there is little difference and, euh, at a certain moment you have reached the max, so then you have thrown 3,000 times and then everything is about the same and if you throw 6,000 times, that doesn't matter much.

While these participants correctly indicated that appropriately sized samples resemble each other, the understanding that inferring from one sample is therefore possible, remained implicit. During the remainder of the intervention, the issue of sampling variability was not discussed again, probably indicating that most participants agreed it is possible to make inferences from sufficiently large samples. This was evidenced during the model lesson when various participants expressed their uncertainty about their conclusions because of the small size of the individual groups' samples.

Conclusion The evidence both from the tests and from the intervention suggests that the simulation led to increased understanding of sampling variability.

4.4 Sampling methods

Pretest and posttest Although in the posttest two more participants agreed that random sampling is appropriate, as compared to the pretest (pretest: 9; posttest: 11), in both tests most participants preferred distributed sampling over random sampling. A large majority incorrectly agreed that distributed sampling is an appropriate sampling method (pretest: 15; posttest: 13).

While in the pretest five participants thought a convenience sample could be representative, none did in the posttest. Finally, an increased number (from two to six) agreed that when an appropriate sampling method is used, the sample is representative for all population characteristics.

Intervention Throughout the intervention, most participants showed a preference for distributed sampling. For instance, during the group discussion of the homework assignment, all participants suggested this sampling method; random sampling was not considered. Some groups even made long lists of factors that would need to be included in appropriate quota. In addition, during the model lesson, four of the seven groups suggested using distributed sampling, in particular sampling from the three difficulty levels (A, B and C) of the books. The other three groups could not agree on whether to use random sampling or distributed sampling, even though during the following class discussion the participants eventually agreed on using random sampling.

Evidence from the model lesson hints at two possible reasons why most participants preferred distributed sampling, although they acknowledged that random sampling could yield a representative sample. First, one group chose distributed sampling because it allowed them to control the representativeness of the sample:

> Astrid: OK, so you don't want to do it randomly?
> Sander: No. I don't think that's handy.
> …
> Astrid: I don't know what, what- In my head it sounds much more reliable if you take from each difficulty level.
> Sander: [Random sampling] seems a bit too easy to me.

Sander found random sampling not "handy" and "too easy" in this context, and Astrid said that "in her head" distributed sampling seemed more reliable. They might have thought they had more control when using distributed sampling and thus more certainty about the sample's representativeness.

Second, most participants might not have realized that when using distributed sampling, the proportions of relevant population characteristics must be known. During the model lesson, although the proportion of the three difficulty levels in the pile of books was unknown, three groups proposed sampling the same number of books from each level. In the class discussion, Nico used the example of a non-uniform population distribution to explain why distributed sampling was not correct: "If your library consists of 1,000 books of level C and 100 of level B and 100 of level A, … then you shouldn't sample proportionally [i.e., uniformly]…." Building on this example, Nico and the teacher educator tried to explain why distributed sampling is inappropriate. Various participants agreed that selecting four A, four B and four C books would not be representative in Nico's example. The teacher educator then concluded that random sampling solves the problem of not knowing the population proportions. Although his proposal of using random sampling was accepted, none of the participants explicitly indicated that they understood Nico's line of reasoning.

Conclusion The simulation seemed to have helped a number of participants to acknowledge that random sampling is an appropriate sampling method because during the model lesson the participants agreed to use random sampling and because in the posttest two more participants agreed that random sampling is an appropriate sampling method than in the pretest. Still, a majority continued to prefer distributed sampling over random sampling. There is some evidence that some participants thought distributed sampling helped them to control the

representativeness of the sample. Moreover, most participants might not have realized that to obtain a representative sample when using distributed sampling, the proportions of relevant population characteristics must be known.

4.5 Sample size

Pretest and posttest Little development was found in the knowledge about sample size, and, overall, none of the participants reasoned entirely in line with the learning objective. Some progress was seen in the knowledge about the minimum sample size required, as the number of participants who thought that a sample needed to be at least 10,000 decreased from five to two. In both tests, in the first task involving sample selection little attention was paid to sample size.

Intervention At the beginning of the intervention, many participants did not appear to have much knowledge about appropriate sample sizes, as five participants explicitly indicated. Before the simulation, there seemed to be a consensus that a sample of 1000 was too small to be representative of a population of 17 million. Even the teacher educator's demonstration of sample size as part of the simulation did not convince the participants that a sample of 1000 was sufficient. What might have been influential for at least some participants was the web-based sample size calculator that Nico found on the internet. Nico repeatedly put forward the idea that an optimal sample size is somewhere around 4000. Some participants seemed to have taken up this number, as evidenced by the remarks of the participants during the recap of the simulation (see the quote in the Sampling Variability subsection).

Not accepting a sample size of 1000 for a population of 17 million might also be related to the idea that the required sample size is proportional to the population size. During the group discussion around the homework assignment, this idea was discussed in four of the five groups. Percentages between 10 and 30% of the population were mentioned, although some participants stated that the sample size is proportional to the population size only up to a certain point.

Conclusion Little evidence was found that the participants accepted a sample size of 1000. Over the course of the intervention, fewer participants thought that a sample needs to be at least 10,000, probably due to the simulation and a sample size calculator. Around a quarter of the participants still thought that a sample of 40 is sufficiently large or that a sample needs to be at least 10,000. Probably about half of the participants accepted a sample size of 2000 to 3000. These participants might have combined the information from the simulation and from the sample size calculator and concluded that a sample size of around 2000 to 3000 is a safer number than 1000. Overall, little development was found in their knowledge about sample size, and none of the participants' knowledge was entirely in line with this learning objective.

4.6 Uncertainty in inferences

Pretest and posttest (Almost) all participants acknowledged the impossibility of making absolutely certain inferences (pretest: 16; posttest: 15) and agreed that a larger sample leads to greater certainty in relation to the inference. In both tests, only one participant incorporated uncertainty in their open-ended response in the second task. The other participants only described the data, which makes reference to uncertainty superfluous.

Intervention The idea that a larger sample yields more certainty was widely shared by the participants. For example, during the model lesson, six out of seven groups suggested increasing the sample size to have a more certain inference. In addition, during the first session, all participants agreed with this idea. Some participants refined it by arguing that the additional benefit of a larger sample decreases when the sample size increases. Most pre-service teachers did not consider the effect of the sample variance on the certainty, as only two pre-service teachers observed that the large difference between the number 1 and 2 increased the certainty of the inference.

It appeared to be problematic how to express the certainty of the inference. For example, Romy stated that 98.5% is not very certain—a percentage one would typically regard as very certain. Other participants were extremely certain, calling out percentages such as 100 and 99.7%, or else they made wild guesses, such as 62.3%. Various participants admitted explicitly that they found it difficult to correctly articulate the certainty of an inference. They may have lacked the tools to express their certainty—tools that were not provided by the teacher educator. Nico found such a tool in the sample size calculator, which calculated the required sample size for given levels of certainty. Although he indicated he did not know how this calculator worked, he still put it forward regularly, as a way to express the certainty of the inferences.

Conclusion While the ideas that any inference is inherently uncertain and that a larger sample yields more certainty were adhered to by all participants, the participants lacked the tools for how to express the certainty of their inferences.

5 Discussion

The present study investigates how teacher college education can contribute to the development of pre-service teachers' content knowledge of informal statistical inference. It is encouraging to see that three quarters of the participants seemed to understand the core ISI elements, such as the value of data as evidence, sampling variability and the possibility of making uncertain inferences based on a sample. The first activity, the assignment to search the media for inferential claims, might have created awareness regarding inference and the need to distinguish between a sample and a population, setting the stage for a discussion of other ISI issues. The demonstration of a simulation involving random sampling and varied sample size may have led to increased understanding of sampling variability and acceptance of random sampling, but no development was seen in the participants' knowledge of sufficient sample sizes. The statistical investigation conducted by the participants during a model lesson appeared to have further strengthened their awareness of ISI, but also revealed that many participants continued to favour distributed sampling over random sampling. Only one participant understood the misconception in the equiprobability bias. Finally, participants might have lacked the tools to express the certainty of their inferences.

While previous research found that pre-service teachers and other types of learners tend to describe data only (De Vetten et al., 2018a, 2018b; Pratt, Johnston-Wilder, Ainley, & Mason, 2008), starting the intervention with having the participants search for media articles that inferred from a sample may have made them attentive to the issue of inference and to the distinction between sample and population. This awareness may have been further fostered by

the model lesson's use of a tangible population and simple descriptive analyses. These results might imply that ISI tasks, such as suggested in the literature (Zieffler et al., 2008), are more effective if participants have first been made sensitive to inference and if tangible populations and simple descriptive analyses are used. Although for first introductions into ISI, it could be beneficial to limit the attention for descriptive statistics, our finding that most pre-service teachers did not consider the effect of the sample variance on the certainty of the inference, might imply that previous involvement with exploratory data analysis may help to acknowledge the importance of variation, and thus support the development of inferential reasoning (Makar et al., 2011).

Our evidence shows that the simulation seems to be effective in fostering an understanding of sampling variability. This effectiveness could be called surprising because the simulation did not involve participants in conducting simulations themselves. Lane and Peres (2006) argue that such demonstrations may be ineffective, as learners remain passive. An explanation for the demonstration's effectiveness could be that it was shown at the right moment, as the preceding discussion had made participants aware of their ignorance regarding sampling variability. A deeper understanding of sampling variability might be attainable when participants conduct simulations themselves, in particular if the simulation software allows them to repeatedly make inferences in a short time span, thereby allowing them to rapidly gain experience in making inferences, such as in Arnold, Pfannkuch, Wild, Regan, and Budgett (2011).

Over the course of the intervention, more participants accepted random sampling as an appropriate sampling method. Still, when asked to select a sample themselves, almost all participants stuck to their preference for distributed sampling, emphasizing the need for a representative sample, such as the grade 3, 6 and 9 children in Watson and Moritz (2000a). One reason might be that they felt a loss of control when using random sampling. This is in line with the findings of Schwartz, Goldman, Vye, and Barron (1998), who report that fifth and sixth grade students tended to accept random sampling in chance contexts but preferred distributed sampling in opinion research contexts. Another reason might be that they lacked an understanding of the workings of random sampling and distributed sampling (Chi, 2013). An explanation of why distributed sampling does not work when the population quota are unknown arose only spontaneously during the model lesson, while the intervention did not contain an explicit comparison of random and distributed sampling. The latter could be added in future versions of the intervention.

The period during the model lesson when the participants combined the evidence from the investigation with their prior knowledge is of interest in relation to how pre-service teachers acquire knowledge. The combination of sources of evidence resembles Bayesian reasoning where new information is used to update a priori probabilities based on prior knowledge. Because in our study the data confirmed the participants' prior knowledge, they may not have felt the need to change their knowledge. In future investigations, situations could be created where the evidence from the data conflicts with pre-service teachers' prior knowledge. Such conflicts can be drivers of inquiry (Makar et al., 2011) and may reveal to what extent pre-service teachers adjust their knowledge to accommodate new evidence (Tversky & Kahneman, 1974).

The participants' understanding of the equiprobability bias appeared to be very limited, even when compared to the 11th graders studied by Watson and Moritz (2000b), of whom about 75% displayed the equiprobability bias. This might be due to the task design that required the participants to predict a single value rather than look at a set of values (Garfield, 1998). Therefore, in teaching the equiprobability bias, making a prediction about a population might be more effective than predicting a single value.

The expression of uncertainty levels seemed to be problematic for the participants, and this could have been partly due to the task design of the model lesson. In the context of ISI, although formal confidence levels might not be available, appropriate statistical tools and probabilistic intuitions are still required to support ISI (Makar et al., 2011; Rossman, 2008). The activity could be adjusted in such a way that multiple samples of two different sample sizes can be compared and the proportion of samples with the same most frequently used word can be used as an approximation of the certainty of the inference. Another way to support the quantification of uncertainty could be to take a modelling approach (Biehler, Frischemeier, & Podworny, 2017) and to use hands-on activities (Zapata-Cardona, 2015) or computer simulations to model sampling distributions (Braham & Ben-Zvi, 2015; Kazak & Pratt, 2017), which could lead to precursors of confidence intervals (Arnold et al., 2011).

Several issues warrant a cautious interpretation of the results. First, this was a small-scale study in the Dutch context where students enter teacher college immediately after secondary education. Therefore, the results are not readily generalizable to other contexts. However, similar processes may occur in countries where students enter teacher college with similar backgrounds and with similar statistics curricula in primary and secondary education. Second, sometimes the tests did not elicit the knowledge that the participants appeared to have, based on evidence from the intervention. For instance, the tests yielded no precise information about what sample size the participants deemed sufficient. Future research could incorporate items that elicit more precise responses regarding sample sizes.

In conclusion, our study is an example of how within a limited time frame teacher college education can facilitate the development of pre-service teachers' ISI-CK. This might be of interest for researchers and teacher educators in contexts where only limited time is available for ISI; therefore, the study complements previous intervention research with pre-service primary school teachers (Groth, 2017; Leavy, 2010). The development of the pre-service teachers' ISI-CK was not an end in itself but primarily a means to support the pre-service teachers in introducing primary school students to ISI. The role of the ISI-CK acquired and the extent to which the participants are able to introduce ISI to the primary school students are questions we hope to answer in future studies.

Compliance with ethical standards

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

References

Arnold, P., Pfannkuch, M., Wild, C. J., Regan, M., & Budgett, S. (2011). Enhancing students' inferential reasoning: From hands-on to "movies". *Journal of Statistics Education, 19*(2), 1–32.

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

Biehler, R., Frischemeier, D., & Podworny, S. (2017). Reasoning about models and modelling in the context of informal statistical inference [special issue]. *Statistics Education Research Journal, 16*(2), 8–334.

Braham, H. M., & Ben-Zvi, D. (2015). Students' articulations of uncertainty in informally exploring sampling distributions. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 57–94). Minneapolis, MN: Catalyst Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

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. (2006). Elementary pre-service teachers' conceptions of variation in a probability context. *Statistics Education Research Journal, 5*(1), 36–64.

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer.

Chi, M. T. H. (2013). Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 49–70). New York, NY: Routledge.

De Vetten, A., Schoonenboom, J., Keijzer, R., & Van Oers, B. (2018a). Topics and trends in current statistics education research: International perspectives. In G. Burrill & D. Ben-Zvi (Eds.), *The growing samples heuristic. Exploring pre-service teachers' understanding about informal statistical inference when generalizing from samples of increasing size*. New York, NY: Springer. (in press).

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

Fennema, E., & Franke, L. M. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147–164). New York, NY: Macmillan.

Garfield, J. (1998, april). *Challenges in assessing statistical reasoning*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

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

Groth, R. E. (2013). Characterizing key developmental understandings and pedagogically powerful ideas within a statistical knowledge for teaching framework. *Mathematical Thinking and Learning, 15*(2), 121–145.

Groth, R. E. (2017). Developing statistical knowledge for teaching during design-based research. *Statistics Education Research Journal, 16*(2), 376–396.

Groth, R. E., & Meletiou-Mavrotheris, M. (2018). Research on statistics teachers' cognitive and affective characteristics. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 327–355). Cham, Switzerland: Springer.

Kazak, S., & Pratt, D. (2017). Pre-service mathematics teachers' use of probability models in making informal inferences about a chance game. *Statistics Education Research Journal, 16*(2), 287–304.

Lane, D. M., & Peres, S. C. (2006). Interactive simulations in the teaching of statistics: Promise and pitfalls. In A. Rossman & B. Chance (Eds.), *Proceedings of the seventh international conference on teaching statistics*. Voorburg, The Netherlands: IASE.

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

Lecoutre, M.-P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557–568.

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. (2018). Learning about statistical inference. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 261–294). Cham, Switzerland: Springer.

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.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1), 1–20.

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.

Pfannkuch, M., & Ben-Zvi, D. (2011). Developing teachers' statistical thinking. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 323–333). Dordrecht, The Netherlands: Springer.

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal, 7*(2), 107–129.

Pratt, D., & Kazak, S. (2018). Research on uncertainty. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 193–227). Cham, Switzerland: Springer.

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

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

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

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.

Schwartz, D. L., Goldman, S. R., Vye, N. J., & Barron, B. J. (1998). Reflections on statistics: Learning, teaching, and assessment in grades K-12. In S. P. Lajoie (Ed.), *Aligning everyday and mathematical reasoning: The case of sampling assumptions* (pp. 233–273). Mahwah, NJ: Lawrence Erlbaum.

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

Torrance, H. (2012). Triangulation, respondent validation, and democratic participation in mixed methods research. *Journal of Mixed Methods Research, 6*(2), 111–123.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Van Blokland, P., & Van de Giessen, C. (2016). VUSTAT [computer software]. Amsterdam, the Netherlands: VUSOFT.

Van Zanten, M. A. (2010). De kennisbasis voor pabo's: Ontwikkelingen en overwegingen [Knowledge base for primary education teacher colleges: Developments and considerations]. *Panama-post, 29*(1), 3–16.

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., & English, L. D. (2016). Repeated random sampling in year 5. *Journal of Statistics Education, 24*(1), 27–37.

Watson, J. M., & Moritz, J. B. (2000a). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*(1), 44–70.

Watson, J. M., & Moritz, J. B. (2000b). Development of understanding of sampling for statistical literacy. *The Journal of Mathematical Behavior, 19*(1), 109–136.

Wells, G. (1999). Dialogic inquiry: Towards a socio-cultural practice and theory of education. Cambridge, UK: Cambridge University Press.

Zapata-Cardona, L. (2015). Exploring teachers' ideas of uncertainty. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 163–181). Minneapolis, MN: Catalyst Press.

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.