INFORMAL STATISTICAL INFERENCE AND PRE-SERVICE PRIMARY SCHOOL TEACHERS: THE DEVELOPMENT OF CONTENT KNOWLEDGE IN TEACHER COLLEGE EDUCATION

Arjen de Vetten[1], Judith Schoonenboom[2], Ronald Keijzer[3] and Bert van Oers[1]
[1]Vrije Universiteit Amsterdam, Section of Educational Sciences, and LEARN! Research Institute, The Netherlands. [2]Department of Education, University of Vienna, Austria. [3]University of Applied Sciences iPabo Amsterdam, Academy for Teacher Education, The Netherlands
a.j.de.vetten@fsw.leidenuniv.nl

Teachers who engage primary school students in informal statistical inference (ISI) must themselves have good content knowledge of ISI (ISI-CK). In this case study, the ISI-CK development was investigated in a class of 21 pre-service primary school teachers who participated in an intervention. Based on analyses of pre-test, post-test and intervention data, the results suggest that most participants acknowledged the possibility of making inferences. An assignment to search the media for inferential claims created awareness regarding inference. A simulation probably increased the participants' knowledge of sampling variability and random sampling, but many participants favored distributed sampling over random sampling. Further research on belief formation with regard to data as evidence, sampling methods and the expression of uncertainty is needed.

INTRODUCTION

In today's society, it is increasingly important to be able to reason inferentially. 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 inferences 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 for 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). If primary school students are to be introduced to ISI, their future teachers must have appropriate content knowledge (CK) of ISI (Groth & Meletiou-Mavrotheris, 2018). In this study, we conceptualized Makar and Rubin (2009) ISI framework 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 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.
    b. Sampling method: The inference includes a discussion of the sampling method and the implications of the sample representativeness.
    c. Sample size: The inference includes a discussion of the sample size and the implications of the sample representativeness.
    d. Uncertainty: The inference is expressed with uncertainty and includes a discussion of what the sample characteristics 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, Schoonenboom, Keijzer, & Van Oers, in press-a, in press-b; Watson, 2001). Furthermore, they lack awareness that ISI tasks

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

require an inference over and above a descriptive analysis of the data, while mixed results were found regarding the value of data as evidence and the possibility of making inferences (De Vetten et al., in press-a, in press-b). Because there is little scientific understanding of how (pre-service) teachers can be supported in their ISI-CK development (Groth & Meletiou-Mavrotheris, 2018), this paper, which is an abridged version of a paper to be published elsewhere, reports on the development of pre-service teachers' ISI-CK during a short intervention. 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?'

Table 1. ISI-CK learning objectives of the intervention

ISI component		*Learning objectives – The pre-service teachers*
Data as evidence		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		are aware when a task requires an inference.
		know that it is possible to use a sample to make general claims about the population.
		know that generally not each possible outcome of a random process has equal probability of occurring (equiprobability bias).
Probabilistic language	Sampling variability	understand that when a sample is relatively large (e.g. 1,000) and randomly selected, the probability is small that another similar sample will give an entirely different result.
	Sampling method	know that random sampling is an appropriate method to obtain a sample.
		prefer random sampling over distributed sampling (i.e. purposefully selecting individuals to obtain a distributed sample across critical population characteristics).
		know that convenience sampling, such as sampling one's own class, is an inappropriate sampling method to obtain a representative sample.
		understand why an appropriate sampling method yields a sample in which aggregate characteristics are close approximates of the population.
	Sample size	know what sufficient sample sizes are in different contexts and understand why this is the case (e.g. understand why a sample size of 1,000 is a sufficient sample size for the entire Dutch population of 17 million people).
	Uncertainty	acknowledge uncertainty of inferences and impossibility of absolutely certain inferences.
		know larger samples are more likely to yield precise estimates of the population parameters.

METHOD

Participants and intervention

One class of 21 second-year pre-service teachers from a Dutch teacher college for primary education participated. All participants had encountered some basic descriptive statistics during their first year of study, while a minority did during their secondary education. The average age of the participants was 20.95 years (SD: 2.19); six were male. The first author was the teacher educator.

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. 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. While the intervention also had the participants teach an ISI lesson in their placement schools, the preparation, delivery and evaluation of these lessons is beyond the scope of this paper. Based on our ISI framework, 12 learning objectives were formulated (see Table 1). Table 2 describes the activities that were intended to support the development of the participants' ISI-CK.

Data collection and analysis

The pre-test and post-test consisted of two tasks, adjusted from De Vetten et al. (in press-b). Both tasks started with an open-ended question. Next, participants were asked to evaluate the correctness of fictitious statements of primary school students concerning the same task. Task 1 investigated the selection of a representative sample. In Task 2, participants were asked to compare two sample distributions and to generalize from these samples. During the intervention, all written work was collected and class and small groups interactions were recorded. The transcripts from the intervention data and the open-ended responses from the tests were coded. The ISI framework was used to categorize the text data into one or more ISI components. Codes that were short summaries of the text were attached to the text to describe the exact meaning.

The results from the pre-test, the post-test and the intervention data were first analyzed separately and then combined. The results of the pre-test and post-test were based on information from the 16 participants who completed both the pre-test and the post-test. The coded open-ended responses from the tests 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 18 parts, such as group and class discussions. These parts were analyzed separately. Using the main results from all 18 parts, we described the development of ISI-CK for each component over the course of the intervention. All results were discussed with an external researcher.

Table 2. Activities used during the intervention

Homework assignment 'Samples in the media': (Session 1: 60 minutes)
This activity aimed at creating awareness of the use of inferential reasoning in the media. Before the first session, the participants searched 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 research. During the first session, errors identified in the news items and appropriate sampling methods, sample sizes and the certainty of inferences were discussed.

Simulation: (Session 1: 20 minutes; session 2: 10 minutes)
The teacher educator used a computer simulation to explain random sampling and the effects of sample size on sampling variability. During the next session, the learning points from the simulation were discussed.

Model lesson: (Session 2: 70 minutes)
The teacher educator taught a model lesson that involved a statistical investigation the participants could use in their placement schools. The lesson centered on a large pile of Dutch children's novels, and the research 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 to leave ample time was left for discussing ISI. Through discussions a hypothesis was formed about the top five most likely words and consensus was reached about the sampling method to be used. Next, the groups conducted an investigation using the agreed sampling method. By pooling the sample data and comparing group results, a discussion about sampling variability was elicited.

Car choice activity: (Session 3: 20 minutes)
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 (2000).

RESULTS

For each component, we describe the development of the participants' ISI-CK, combining evidence from the tests and the intervention, and explicate the possible role of the activities in this development. Table 3 summarizes the results for the tests. The subcomponents Sampling variability and Sampling methods are discussed in more detail.

Data as evidence: In the tests and during the intervention, participants valued data as evidence, but during the model lesson there was also some evidence that participants based their conclusions on a combination of sources of evidence at the expense of relying on the data.

Generalization beyond the data: We found only a minimal change with respect to the possibility of making generalizations between the pre-test and the post-test. 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 participants paid attention to inference. The attention to inference might have been further fostered by the model lesson in which both the population (the pile of books) and the sample (the sampled books) were tangible and visible. 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.

Table 3. Participants' ISI-CK during pretest and posttest

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

Sampling variability: The evidence suggests the simulation led to increased understanding of sampling variability. 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 presenting Menthe's quote, the teacher educator brought their struggle to the fore, thus clarifying the issue in question and motivating the participants to learn from the simulation. 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. If you throw 6,000 times, that doesn't matter much.

During the remainder of the intervention, 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.

Sampling methods: 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 post-test more participants agreed that random sampling is an appropriate sampling method than in the pre-test. Still, 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. 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.

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. Probably they thought to have more control when using distributed sampling and 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.

Sample size: 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. Little evidence was found that the participants accepted a sample size of 1,000. 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 found on the Web by one of the participants. 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 2,000 to 3,000 is a safer number than 1,000.

Uncertainty: 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 may have lacked the tools for how to express the certainty of their inferences.

CONCLUSION

This study investigates how teacher college education can contribute to the development of pre-service teachers' ISI-CK. 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. There are some notable issues. First, while previous research found that pre-service teachers tend to describe data only (De Vetten et al., in press-a, in press-b), the awareness for inference seen during this intervention might imply that ISI tasks are more effective if participants have first been made sensitive to inference and if tangible populations and simple descriptive analyses are used. Second, although the demonstration of a simulation did not involve participants in conducting simulations themselves, an explanation why the demonstration was effective in fostering the participants' understanding of sampling variability could be that it was shown at the moment the participants were aware of their ignorance regarding sampling variability. Third, although a growing number of participants accepted random sampling, almost all participants stuck to their preference for distributed sampling. One reason for this finding might be that they felt a loss of control when using random sampling. Another reason might be that they lacked an understanding of the workings of random sampling and distributed sampling. An explicit comparison of random and distributed sampling could, therefore, be added in future versions of the intervention. 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 extent to which the participants are able to introduce ISI to the primary school students are questions we hope to answer in future studies.

REFERENCES

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

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

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

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

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.

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

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

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.