



**HAL**  
open science

## Improving teachers' reasoning about sampling variability: A cross institutional effort

Bridgette L. Jacob, Hollylynne S. Lee, Dung Tran, Helen M. Doerr

### ► To cite this version:

Bridgette L. Jacob, Hollylynne S. Lee, Dung Tran, Helen M. Doerr. Improving teachers' reasoning about sampling variability: A cross institutional effort. CERME 9 - Ninth Congress of the European Society for Research in Mathematics Education, Charles University in Prague, Faculty of Education; ERME, Feb 2015, Prague, Czech Republic. pp.692-699. hal-01287076

**HAL Id: hal-01287076**

**<https://hal.science/hal-01287076v1>**

Submitted on 11 Mar 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Improving teachers' reasoning about sampling variability: A cross institutional effort

Bridgette L. Jacob<sup>1</sup>, Hollylynn S. Lee<sup>2</sup>, Dung Tran<sup>2</sup> and Helen M. Doerr<sup>3</sup>

1 Onondaga Community College, Syracuse, United States, [jacobb@sunyocc.edu](mailto:jacobb@sunyocc.edu)

2 North Carolina State University, Raleigh, United States

3 Syracuse University, Syracuse, United States

*The research reported here uses common items to assess statistical reasoning of teachers enrolled in a graduate-level education course to evaluate their reasoning about sampling variability. In particular, we discuss key aspects of a purposeful course design aimed at improving teachers' learning and teaching of statistics, and the resulting different ways of reasoning about sampling variability that teachers exhibited before and after the course.*

**Keywords:** Statistics, sampling, variability, reasoning, assessment.

Given the strong attention to statistics in the secondary curriculum in many countries (e.g., England Department of Education, 2014; CCSSI, 2010), many teacher professional development efforts and graduate courses include more opportunities for secondary teachers to develop their statistical reasoning and to learn pedagogical strategies for teaching statistics. Researchers have investigated the statistical knowledge needed for teaching using various frameworks (e.g., Burgess, 2011; Noll, 2011). Each of these frameworks has identified teachers' own statistical reasoning as a foundational aspect of their ability to teach statistics.

Our research is situated within the collaborative design and implementation of a graduate course across two institutions focused on teaching and learning statistics. Our design work is largely influenced by Pfannkuch and Ben-Zvi's (2011) recommendations for designing experiences to develop teachers' statistical reasoning. Though our courses aim to develop teachers' understanding for *teaching* statistics, in this paper we report on teachers' reasoning related to sampling variability, without regard to their under-

standing of how to teach others about sampling variability. To help assess the impact of the experiences we designed in our courses, we used both qualitative and quantitative data sources. The focus in this paper is to examine how teachers' reasoning about sampling variability changed.

Understanding sampling variability has been established as difficult, but key, in one's overall statistical reasoning (e.g., Shaughnessy, 2007). Two aspects of sampling variability are reported in this paper: representations of sampling variability and the effect of sample size on variability. Researchers have found that the understanding of variability from expected values among samples improves with experience and age (e.g., Watson & Kelly, 2004). The role of sample size in variation from expected values has also been the focus of much research. Concerning the effects of sample size, researchers have reported that students, and teachers, use equiprobable reasoning in determining the likelihood of two events without considering differences in the sample size of the two events. (e.g., Watson & Callingham, 2013). In a meta-analysis, Noll and Sharma (2014) discussed the Hospital Task, one of the four items reported here, which has been used to assess reasoning about the effect of sample size on parameter estimation. Since the original version presented by Kahneman and Tversky (1972), it has been revised and implemented with students from approximately 10 years of age through college with the predominant response being the equiprobable response. In 2013, Lee, Doerr, Arleback, and Pulis reported that after experiencing an earlier version of our graduate course, teachers still exhibited difficulties with sampling variability, including a tendency to apply equiprobable reasoning by ignoring the effect of sample size on variability. Our reflections on those findings led to the current work.

## COURSE AND PARTICIPANTS

Following from the results of Lee and colleagues (2013), a team of four instructors from two institutions began conceptualizing ways to improve a graduate course focused on teaching and learning statistics. Related to sampling variability, suggested improvements included readings and discussions targeted to draw attention to students' tendency for equiprobable reasoning. These suggestions also included purposeful task design using technology tools for data exploration. The team met weekly via videoconference for an academic year (2013–14) to design a 15-week course, and to discuss issues and alter plans as the course was taught in Spring 2014. The course consisted of opportunities for teachers to engage in statistical investigations with real data and tasks designed to develop their understandings of distribution, samples and sampling distributions, and inferential statistics, especially using randomization approaches. The course used the dynamic software *Fathom* (Finzer, 2005) and *TinkerPlots* (Konold & Miller, 2011), as well as online applets and resources such as StatKey ([lock5stat.com/statkey](http://lock5stat.com/statkey)). The course included readings and discussions about (a) the nature of statistical reasoning, and (b) students' learning and reasoning related to the aforementioned topics. Software tools were used to support teachers' learning by allowing them to flexibly explore graphical representations, easily compare data sets, and make changes to data in displays to explore conjectures. The software provided the simulation tools necessary to create representations of a population, a sample, and an empirical sampling distribution. Given the research on students' struggle to understand sampling distributions (e.g., Saldanha & Thompson, 2014), we saw these representations as critical for developing teachers' knowledge of sampling variability.

Across institutions, the course served a variety of graduate students ( $n = 27$ , 8 in Course1 and 19 in Course2). Participants consisted of one pre-service teacher (5th year senior), six pre-service and 11 in-service teachers in masters programs, one full-time masters student in mathematics education, seven doctoral students in Mathematics or Mathematics Education (three currently teaching in post-secondary contexts), and one doctoral student with interests in statistics education. Twenty-one participants were female and six were male, with six participants for whom English was a second language. Most participants had com-

pleted the equivalent of an undergraduate major in mathematics, with all but two having had at least one course in statistics. Henceforth we refer to course participants as teachers.

## DATA SOURCES AND ANALYSIS

One source of data was participants' responses to a statistical concept inventory constructed to align with our course goals, content, and experiences. On the first day of class and during the final week of the course, all participants completed a 20-item multiple choice test with items in five categories: distributions (5 items), comparing distributions (3 items), probability (2 items), sampling variability (7 items), and formal inference (3 items). Eleven items were drawn from validated instruments (delMas, et al., 2007; Garfield, 2003), with seven selected from the ARTIST database ([apps3.cehd.umn.edu/artist](http://apps3.cehd.umn.edu/artist)), and two items adapted from research (e.g., Watson & Kelly, 2004; Zieffler et al., 2008). The 20-item test was agreed upon by instructors during the planning phase to ensure items had content validity to measure concepts to be addressed in the course. For two of the four sampling variability items we highlight in this paper, teachers were asked to justify their choices.

After the course, semi-structured interviews were conducted with selected teachers ( $n = 14$ ) across institutions to understand changes in their reasoning and perceptions of what might have influenced those changes. Interviews (45–90 minutes) were audio or video taped. Interviewees were purposely selected because of trends in their responses to items on the test. For example, some were selected because they had improved from an incorrect response to a correct response on several items, while maintaining incorrect responses on other items. During the interview, participants were shown an item, given time to reread, told which choice (A, B, C, etc.) they had selected on the pre and post-test, and then asked about their reasoning. Based on responses, the interviewer asked questions prompting them to elaborate about their reasoning.

Descriptive statistics and t-tests were used for the 20-item assessment to document the change in teachers' performance on the pre and post-test, both overall and on individual subscales. Teachers' responses to the two open-ended items and responses of the 14 teachers interviewed were open coded to identify emerging

themes to gain insight into the teachers' reasoning about the changes in their responses.

**RESULTS**

Analysis of the pre and post-test showed significant improvement in teachers' overall scores (out of 20), with a mean increase of 1.84 points (s.d.=1.98). Strong gains were found in the items related to sampling variability, with a mean increase in scores (out of 7) of 1.3055 (s.d.=1.19). In this paper, we report on changes in teachers' reasoning for two key categories: (1) the effect of sample size on the likelihood of outcomes, and (2) representations of sampling variability.

Two items, Brown Candies and Two Hospitals, pertained to the effects of sample size on the likelihood of outcomes from a sample. At the beginning of the course, about half of the teachers correctly answered each of these items. Two additional items, Sample Means and Sample Proportions, asked teachers to reason about expected variability in a distribution of sample statistics when sampling from a given population distribution. For both of these items, a larger proportion of teachers were able to correctly respond at the beginning of the course, 78% and 70% respectively.

**Effect of sample size on the likelihood of outcomes**

Table 1 illustrates the distribution of responses for the Brown Candies item. There was a major shift to 83.2% responding correctly on the post-test. Most notable was the decrease in the number of teachers choosing the equiprobable response (E).

By the end of the course, the teachers gained a clearer understanding that smaller sample sizes have greater variability than larger sample sizes, thus resulting in small samples being *more likely* to have larger deviations from the expected percentage of 50% brown candies. The following is representative of teachers' responses when asked to explain their reasoning about the Brown Candies item.

The large [bag] is more like 50%, the small [bag] is more unlikely because of smaller sample size. For example you flip a coin, you have more chance to have 8 and 2 whereas you flip 200, you are more likely to get 50%. (Teacher 27)

The single teacher to choose 'B', the larger bag having more variability, on the post-test, chose the equiprobable response 'E' on the pre-test. However, during the interview, she realized she had chosen in error and stated her reasoning:

Maybe at first when I answered it, at first I think I didn't have any idea about sample size. But after we learned something about sample size in class and how it will affect, you know, like the variability... [reading the problem] So Sam is the one having a larger, a large family sized bag? So that implies that large family sized bag will have a large sample size? And that implies that we should have less variability? I think I was wrong. So it should be 'C'. (Teacher 1)

This teacher appeared to have difficulty with the complex terminology in the item rather than a misunderstanding of the underlying concept. Three teachers

<b>Brown Candies:</b> A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag. Which bag is more likely to have more than 70% brown candies?		
	Pre-Test	Post-Test
Sam, because there are more candies, so his bag can have more brown candies.	0.0% (0)	0.0% (0)
Sam, because there is more variability in the proportion of browns among large samples.	3.7% (1)	3.7% (1)
<b>Kerry, because there is more variability in the proportion of browns among smaller samples.</b>	<b>51.9% (14)</b>	<b>83.2% (23)</b>
Kerry, because most small bags will have more than 50% brown candies.	3.7% (1)	0.0% (0)
Both have the same chance because they are both random samples.	40.7% (11)	11.1% (3)

**Table 1:** Results of the Brown Candies Item

chose 'E', the equiprobable response, on both the pre and post-test. One teacher explains:

I was thinking about this idea of flipping a coin... flipping it head, the next time you flip, it is 1 over 2. So ... making an analogy to this it says that a certain manufacturer claims that they produce 50% brown candies... So it doesn't really matter whether the bag has ten candies or a thousand candies... there is 50% chance I mean 50% of them... would be brown. So long as the bag contains candies from this manufacturer. (Teacher 7)

This teacher is equating samples in the large and small bags of candies to tossing a single coin rather than a series of coin flips in the long or short run.

For the related Two Hospitals item, teachers needed to reverse their thinking by choosing the hospital that was *less likely* to record a high percentage of female births. Teachers were also asked to write about their reasoning for this item on the test. Table 2 shows results for the Two Hospitals item in which gains were made in teachers' correct responses. Again, we saw a decrease in the number of teachers choosing the equiprobable response (C) and an increase in the correct choice (A).

Of the 22 teachers answering this correctly on the post-test, the responses below are representative of their thoughts about sample size and variability.

So Hospital B is more likely to have 80% or more. And then I also thought about the numbers, if you're doing 50 births a day, 40 girls out of 50 seems like a lot compared to 8 out of 10. (Teacher 3)

The larger hospital will have less variability from the expected value of 50% boys and 50% girls. (Teacher 17)

These teachers illustrated an understanding of the relationship between sample size and variability. In the open-ended responses, four teachers also referred to the actual number of births, stating 8 out of 10 female births was a likely outcome for the smaller hospital but 40 out of 50 female births was unlikely for the larger hospital. During interviews, six teachers noted the conceptual similarity between the Brown Candies and Two Hospitals items and that they had to "reverse" their thinking for the latter.

Of the five teachers choosing incorrectly on the post-test, three of them indicated the smaller hospital was less likely to record a high percentage of female births (B). During interviews with two of these teachers, they indicated they had misread the problem on the post-test. "Yeah 'A', the big one. The reason is because there is more variability... [laughs]. The answer is wrong but my explanation is correct. There is more variability in the small sample." (Teacher 6) Both teachers had responded instead to which hospital would be *more likely* to record 80% female births.

The remaining two teachers responding incorrectly on the post-test chose the equiprobable response. One of them chose the equiprobable response on the pre and post-test, as he had done for the Brown Candies problem. His open-ended response on the post-test revealed his reasoning, "Each birth is independent of the other births and there is a 0.5 probability that each birth (independent) of others would result in a boy or a girl." (Teacher 7) Both teachers demonstrated the difficulty of dispelling the notion of equiprobability of events with small and large samples.

### Representations of sampling variability

The Sample Means and Sample Proportions items assess teachers' understanding about sampling distribution by asking teachers to predict variability from expected outcomes, using different representations.

<b>Two Hospitals:</b> Suppose about half of all newborns are girls and half are boys. Hospital A, a large city hospital, records an average of 50 births a day. Hospital B, a small, rural hospital, records an average of 10 births a day. On a particular day, which hospital is less likely to record 80% or more female births?		
	Pre-Test	Post-Test
Hospital A (with 50 births a day)	51.9% (14)	81.5% (22)
Hospital B (with 10 births a day)	14.8% (4)	11.1% (3)
The two hospitals are equally likely to record such an event	18.5% (5)	7.4% (2)
Not able to determine based on given information.	14.8% (4)	0% (0)

**Table 2:** Results of the Two Hospitals Item

The Sample Means item provides a graphical display of the population distribution with the mean and standard deviation, and asks teachers to choose the most likely dotplot of five sample means of size 10. The Sample Proportions item gives the population proportion numerically and asks which set of five proportions from random samples of size 20 is most likely.

Table 3 shows the distribution of teachers' responses on the pre and post-test for the Sample Means item. Initially 78% of the teachers were able to identify reasonable variation from expected. This number increased by 11% after the course. Further examining the open responses and interviews gave insight into their reasoning.

Seven of the 14 teachers interviewed were able to eliminate too little variation (response a), and too much variation (response b), indicating a strong sense for reasonable expectations in variation. For example, Teacher 15 stated, "Here with only five samples, I think 'a' would be too perfect. I would throw out 'b' because the chance you have 10 values with the mean of 8.5 would be very slim, if even ever [pointing to dot above 8.5 on dotplot]". Four of the teachers were able to distinguish between the population, samples and sample means, and the sampling distribution. "I try to

grasp my head around five students with 10 values, so that this dot represents the mean of 10 values, not just one value I picked out" (Teacher 15). In addition, they could incorporate the sample size into the expected variability of the sampling distribution. As an example, Teacher 11 replied:

Sample 10 seems to me not a large sample size set of data so I am certainly expecting... if I take 1000 values, I am convinced that the sample mean should be put together. With 10 values I am suspicious. That's possible but less probable. And by exactly the opposite reasoning, 10 seems big enough that I see so much variation that I see. No way I can quantify... that is too much variation of sample 10. That dot [pointing to right dot in 'b'] I should have many values over here [pointing to right tail of the population distribution] that's simply outrageous; I believe if it is 1 sample [pointing to 'b'].

When reasoning about the population, samples, and the distribution of sample means, the teachers either went back to the graph of the population to estimate the means, or used the numerical statistics given to evaluate the possibility of the sample means.

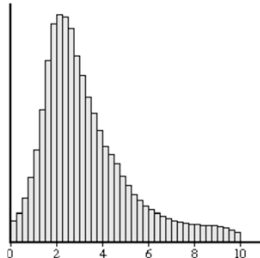
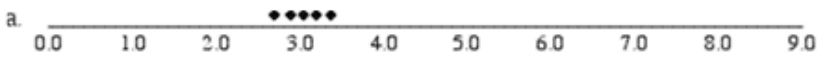
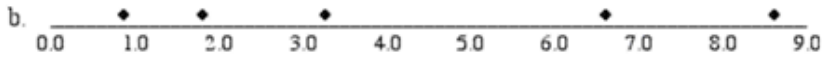
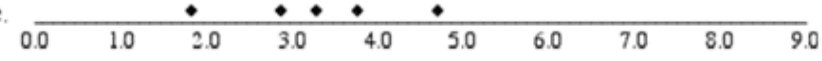
<p><b>Sample Means:</b> The distribution for a population of measurements is presented at the right. The mean is 3.2 and the standard deviation is 2. Suppose that five students each take a sample of ten values from the population and each student calculates the sample mean for his or her ten data values. The students draw a dotplot of their five sample means on the classroom board so that they can compare them. Which of the following dotplots do you think is the most plausible for the one they drew on the board?</p>		
<p>a. </p>	<p><b>Pre-Test</b> 14.8% (4)</p>	<p><b>Post-Test</b> 3.7% (1)</p>
<p>b. </p>	<p>7.4% (2)</p>	<p>7.4% (2)</p>
<p>c. </p>	<p>77.8% (21)</p>	<p>88.9% (24)</p>

Table 3: Results of the Sample Means Item

I think the new SD should be between 1/2 and 1. I don't know if 10 is large enough, but with large number of sample, the sampling distribution is approximate a normal distribution so I think if the mean is here [point to about 3.2 in 'c'] so other will be 1/2 SD from the mean and other is about 1 SD within the mean. (Teacher 18)

It is about 76%, then 2SD of 95%... I am looking for most of the things within that. This is like 8.5, it is way out of range. This [pointing to 'a'] does not have enough of deviation within 1 SD, they cramp together. 'C' seems more within the 2SD. (Teacher 30)

These responses reveal teachers' reasoning about sample means using the sampling distribution and the Central Limit Theorem.

Three teachers who got this item wrong on the post-test did not show a robust understanding of population, sample, and sampling distribution. For example, one teacher seemed to believe the distribution of sample means should resemble the population distribution "I think 'b' is making the most sense because I can see the skewness to the right." The sampling distribution remained complicated for her even though the course focused extensively on that construct.

Table 4 shows the distribution of teachers' responses on the Sample Proportions item. While teachers (70.4%) began our course with a good intuition about variation from expected, almost all correctly responded to this item after taking the course.

Teachers' interview comments indicated they were able to eliminate wrong options based on their sense of variability from expected. For example:

I eliminate 'C' because of the 5% and 95%. If I know 35% of the candies are yellow, I know it is not impossible but, I just don't see someone picks 20 candies all but one being yellow and I know there are enough candies in there just 35% of 1000, it could happen. I just, 20 candies is not enough to see the perfect 35% every time. One kid might see it; not every kid might see it. (Teacher 13)

For the teachers who chose an incorrect response, they seemed to be using either equiprobable reasoning or the thinking that anything can happen.

### Reasoning across Items

At the beginning of the course, only 10 of 27 teachers (37%) answered both the Brown Candies and Two Hospitals items correctly. Of the 13 teachers choosing incorrectly on the Brown Candies item, the predominant response chosen by 11 of them was 'E', the equiprobable response (Both have same chance because both are random samples). Of these 11, only four teachers chose the equiprobable response for the Hospital item as well. Another common misconception that larger sample sizes have greater variability was demonstrated by one teacher for the Brown Candies item and four teachers for the Hospital Problem.

After the course, 20 out of 27 teachers (74%) correctly responded to both items; a marked increase from the pre-test. Of the remaining seven teachers, three chose an equiprobable response to one of the items, repeating their error from the pre-test. There was only one teacher who exhibited an equiprobable misconception for both items on the pre-test and the post-test.

For the Sample Mean and Sample Proportion items, before the course 16 out of 27 (59%) teachers correctly answered both items. By item, 21 judged the variation

<b>Sample Proportions:</b> Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 35% yellow candies. Five students each take a random sample of 20 candies, one at a time, and record the percentage of yellow candies in their sample. Which sequence below is the most plausible for the percent of yellow candies obtained in these five samples?		
	Pre-Test	Post-Test
30%, 35%, 15%, 40%, 50%.	70.4% (19)	92.6% (25)
35%, 35%, 35%, 35%, 35%.	14.8% (4)	3.7% (1)
5%, 60%, 10%, 50%, 95%.	3.7% (1)	0% (0)
Any of the above.	11.1% (3)	3.7% (1)

Table 4: Results of the Sample Proportions Item

in a graphical format correctly whereas 19 selected a correct response for the question concerning variation in a numerical format. By the end of the course, almost all teachers correctly answered both items (24/27, 88.8%).

## DISCUSSION AND CONCLUSIONS

Overall, the teachers improved their understanding about sampling variability, in particular the relationship between sample size and variability, and variability from expected. This could be attributed to the extensive focus on statistical investigation and many experiences with simulations in which attention was drawn to expectations from a population distribution, collecting samples and sample measures, and discussing the distribution of sample measures. A few teachers still had difficulty on these items, corroborating prior findings that sampling distribution and sampling variability is complicated to understand (e.g., Saldanha & Thompson, 2014).

The results of this study also confirm that equiprobable reasoning can be misapplied in reasoning about samples of different sizes, and that this reasoning may become more stable for some teachers (e.g., Watson & Callingham, 2013). We observed that for two of the teachers, even with intensive experiences with variability, they still held a deterministic understanding of probability. This might be rooted in their early exposure to theoretical probability that they need to revisit and re-evaluate in order to build up a robust understanding.

We also observed that teachers could develop a sound understanding about sampling variability and reason correctly about an item, yet still choose a wrong answer. This resulted from a misreading of a problem or a misunderstanding of a particular word. Also, for the Sample Means item, teachers could give a correct answer, reasoning with a sense of variability from expected without necessarily understanding the relationship between a population, samples, and the sampling distribution. Thus, we are concerned that this item may be more useful for measuring understanding of variation from expected values rather than sampling distributions.

This study adds to the sparse literature related to teachers' reasoning about statistics, focusing on their understanding of sampling variability. It illustrates

how a carefully designed graduate-level course in teaching and learning statistics improves teachers' understanding of important statistical concepts. In particular, the focus on statistical investigation and reasoning experiences, and the emphasis on a simulation approach for inference, seems to improve teachers' knowledge about sampling variability.

## REFERENCES

- Burgess, T. A. (2011). Teachers' knowledge of and for statistical investigations. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education: A joint ICM/IASE study* (pp. 259–270). New York: Springer.
- Common Core State Standards Initiative (CCSSI). (2010). *Common core state standards for mathematics*. <http://www.corestandards.org>.
- delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28–58.
- England Department of Education. (2014). *The national curriculum in England: Key stages 3 and 4 framework document*. <https://www.gov.uk/government/publications>
- Finzer, W. (2005). *Fathom Dynamic Data Software*. (Version 2.1) [Computer Software]. Emeryville, CA: Key Curriculum Press.
- Garfield, J., delMas, R., Chance, B., & Ooms, A. (2006). *Assessment resource tools for improving statistical thinking*. [Website]. Retrieved from: <https://apps3.cehd.umn.edu/artist/index.html>
- Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, 2(1), 22–38.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Konold, C., & Miller, C. (2011). *TinkerPlots*. (Version 2.0) [Computer Software].
- Lee, H. S., Doerr, H. M., Arleback, J. B., & Pulis, T. (2013). Collaborative design work of teacher educators: A case from statistics. In M. V. Martinez & A. C. Superfine (Eds.), *Proceedings of the 35<sup>th</sup> annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 357–364). Chicago, IL.
- Noll, J., & Sharma, S. (2014). Qualitative meta-analysis on the hospital task: Implications for research. *Journal of Statistics Education*, 22(2), 1–26.



- Noll, J. A. (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, 10(2), 48–74.
- 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: A Joint ICMI/IASE Study* (pp. 323–333). New York: Springer.
- Saldanha, L. A., & Thompson, P. W. (2014). Conceptual issues in understanding the inner logic of statistical inference: Insights from two teaching experiments. *Journal of Mathematical Behavior*, 35, 1–30.
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (vol. 2, pp. 957–1009). Charlotte, NC: Information Age Publishing.
- Watson, J., & Callingham, R. (2013). Likelihood and sample size: The understanding of students and their teachers. *Journal of Mathematical Behavior*, 32, 660–672.
- Watson, J. M., & Kelly, B. A. (2004). Statistical variation in a chance setting: A two-year study. *Educational Studies in Mathematics*, 57(1), 121–144.
- 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.