VARIABILITY IN A SAMPLING CONTEXT: ENHANCING ELEMENTARY PRESERVICE TEACHERS' CONCEPTIONS

Daniel L. Canada
Eastern Washington University, United States of America
mathaction@gmail.com

While recent and ongoing research has begun to reveal ways that precollege students think about variation, more research has been needed to understand the conceptions of variation held by elementary preservice teachers and also how to shape the university courses where those preservice teachers learn. This paper, sharing an excerpt from an exploratory study aimed at preservice teachers, describes changes in class responses to a sampling task where variation is a key component. Overall, going from before to after a series of instructional interventions, responses reflected a more appropriate sensitivity to the presence of variation.

INTRODUCTION

The specific research question addressed in this paper is: How do elementary preservice teachers' responses concerning variation in a sampling context compare from before to after an instructional intervention? Previous research has begun to illuminate precollege student thinking about variation in several contexts, such as probability, data and graphs, and sampling situations. Meanwhile, research on how teachers reason statistically has also been emerging, with recent calls to examine how preservice teachers think about variation or variability in data. This paper focuses on the context of sampling, which is just one of many important situations for considering variability.

RELATED LITERATURE

As an example of research using a direct precursor to the task used in this paper, Rubin, Bruce, and Tenney (1991) interviewed a dozen high school seniors about a question in which the population was known and repeated samples could be drawn. In the Gummy Bears problems, students were told that packets of candy were filled with six Gummy Bears per packet. These candies were packaged after being drawn from a large vat containing two million green and one million red candies. Students were first asked about the number of green candies they thought would be in their own packet; then they estimated how many packets out of 100 would have that same number of green candies. All twelve subjects said that they would expect four out of the six candies in their own packet to be green. However, "when asked if every kid's packet would contain four green Gummy Bears, all of the students knew that there would be variation among samples" (Rubin, Bruce & Tenney, 1991, p. 5). The researchers were thus able to powerfully illustrate the twin ends of the continuum between sample representativeness and sample variability. They concluded by noting that students "lack experience thinking in terms of a *distribution of samples* generated from a particular population" (Rubin, Bruce & Tenney, 1991, p. 12, italics added).

Later, in what came to be known as the Candy Task (in America) or the Lolly Task (in Australia), researchers considered several different ways of framing a task involving a repeated sampling problem in which five samples, or pulls, of size ten were drawn (with replacement) from a known population of colored candies (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999; Torok & Watson, 2000; Reading & Shaughnessy, 2000; Shaughnessy & Ciancetta, 2002). For example, Reading and Shaughnessy (2000) asked elementary and high school students to describe how many red candies might be in a sample of ten candies drawn from a population that held 50% red. They then asked students to graph the results for 40 pulls. Finally, they altered the population itself from 50% red to 70% red. Results showed that students were better at describing reasons for their responses when talking about centers than when talking about variation.

In an exploratory study involving four students each from grades 4, 6, 8, and 10, Torok and Watson (2000) used the same expanded form of the Candy Task that Reading and Shaughnessy (2000) used. Students were asked if the results of their pulls were surprising and

 In C. Batanero, G. Burrill, C. Reading & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference.*

were given an opportunity to modify their earlier answers after doing the experiment. The strongest factors that differentiated students' responses were "the extent to which variation was acknowledged and … the recognition and use of the proportion concept to describe individual outcomes" (Torok & Watson, 2000, p. 153). These factors gave rise to four hierarchical levels that comprised a model for categorizing student reasoning. At the lowest level, subjects acknowledged variation but focused on individual outcomes. The four students at this level were easily swayed by experimental results. At the other end of the hierarchy, the two students in the highest level showed a high level of proportional thinking, balanced by a "very good and consistent appreciation of variation" (p. 160). These students were only moderately influenced by experimental results. The hierarchical levels proposed by Torok and Watson (2000) influenced the analysis of responses for yet other versions of the Candy Task given to elementary, middle, and high school students (Reading & Shaughnessy, 2004; Shaughnessy, Canada, & Ciancetta, 2004).

METHODOLOGY

The thirty subjects in the present study of elementary preservice teachers (24 women, 6 men) were enrolled in a ten-week preservice course at a university in the northwestern United States designed to give prospective teachers a hands-on, activity-based mathematics foundation in geometry and probability and statistics. During the first week of the course, prior to instruction in probability and statistics, subjects took an in-class survey (called a PreSurvey) designed to elicit their understanding on a range of questions about sampling, data and graphs, and probability. The sampling question from the PreSurvey that relates to the current paper is a version of the Candy Task mentioned earlier: Six handfuls of ten candies are to be drawn from a jar containing 60 red and 40 yellow candies (with replacement). Students were asked how many red candies might be in each of their six handfuls. They were also asked why they had chosen the numbers they did. Following the PreSurveys but prior to the class instruction on probability and statistics, individual interviews were conducted with ten subjects to allow further probing of their thinking. After instructional interventions took place in class, a similar PostSurvey question was asked concerning six handfuls of one hundred candies from a jar containing 600 red and 400 yellow candies (with replacement). Again, students were asked how many red candies might be in each of their six handfuls, and they were also asked why they had chosen the numbers they did. Finally, after the PostSurveys the same students who had been earlier interviewed were interviewed once again.

The class interventions were a series of small-group and whole-class activities and simulations that engaged the contexts of data and graphs, sampling, and probability situations and were designed to provide opportunities to notice and wonder about variation. For example, prior to one of the sampling activities (called the "Known Mixture"), we started with a general discussion of what samples were, who uses samples, and what samples were good for. Then the following scenario for the Known Mixture Activity was given to the class:

The band at Johnson Middle School has 100 members, 70 females and 30 males. To plan this year's field trip, the band wants to put together a committee of 10 band members. To be fair, they decide to choose the committee members by putting the names of all the band members in a hat and then they randomly draw out 10 names. As the preservice teachers discussed their initial expectations for this scenario, they especially focused on what would happen if the random draw of 10 names were to be repeated thirty times. After talking about predictions for drawing thirty samples each of size 10, we simulated this activity using chips in a jar. Actual data was gathered and graphed. Then we had a discussion about how the graphs of the predicted data compared to one another, how the graphs of the actual data compared to one another, and also how the predicted graphs compared to the actual graphs. We also used the statistical software *Fathom* (Finzer, 2001) to extend our physical simulations by using the computer to generate data on larger numbers of samples.

Both parts of the sampling questions (*what* students expected and *why*) were taken into consideration for coding purposes, primarily to retain consistency with an analogous rubric derived for a similar question asked in a sampling context (Shaughnessy et al., 2004). The rubric places a higher value on responses that integrate proportional reasoning as well as variation. The

codes and class results for the questions are presented in Table 1. Inappropriate choices for listing *what* was expected (or blank answers) were automatically coded at Level 0, regardless of the reason given. Deciding what would constitute an appropriate choice for the results on six sets of flips or spins involves making a judgment, and the subcodes used for this subquestion question help identify inappropriate choices as (W)ide, (N)arrow, (H)igh or (L)ow.

RESULTS

Of the nine inappropriate PreSurvey responses, four were narrow and two were low (the remaining were left blank). Of the six inappropriate PostSurvey responses, three were high, two were wide, and only one was narrow. A few exemplars from each level are shown along with the actual choices made by the subject, starting with Level 0:

- Kate (Pre) {6, 6, 6, 6, 6, 6} If each student returns their candies to the jar, then the ratio of red & yellow would remain the same.
- Alice (Pre){4, 4, 4, 5, 5, 5} Hard to say. Never exact
- Rob (Post){60, 60, 60, 60, 60, 60} Theoretically you should always get 60 red and 40 yellow. This would be the most educated guess at the 6 outcomes.
- Susie (Post) {30, 50, 55, 60, 75, 85} I believe that the classmates would pull anywhere between 30-85 reds, somewhere between the lower quartile and the upper quartile.

Table 1. Results for PreSurvey and PostSurvey Questions

Code Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
L3	Explanation explicitly involves reasoning about centers as well as notions of variation	4 (13.3%)	9 (30.0%)
L2	Explanation uses either centers or notions of variation but not both.	8 (26.7%)	12 (40.0%)
L1	Explanation uses additive thinking or informal notions of chance (such as general likelihoods)	9 (30.0%)	3 (10.0%)
L0	Inappropriate choice (or Blank) PreSurvey (60R & 40Y) W(ide) = Range $\geq$ 8, N(arrow) = Range $\leq$ 1 H(igh) = All $\geq$ 1, L(ow) = All $\leq$ 6 PreSurvey (600R & 400Y) W(ide) = Range $\geq$ 50, N(arrow) = Range $\leq$ 3 H(igh) = All $\geq$ 60, L(ow) = All $\leq$ 60	9 (30.0%)	6 (20.0%)

By the rubric of Table 1, Alice's (Pre) choices are categorized as N(arrow) since her range is only 1, and moreover her choices are L(ow) since all the numbers are less than 6. Susie's (Post) choices are categorized as W(ide) since the range for her choices is more than 50.

Level 1 responses included appropriate choices for what was expected but the reasons why did not specifically reflect any aspect of distributional thinking:

- Sally (Pre) {5, 6, 6, 7, 7, 8} All are close to 6 or 6.
- Gary (Pre){5, 6, 6, 7, 7, 8} The odds are that each classmate would have more red because there are 20 more reds to begin with.
- Molly (Post) {58, 62, 63, 64, 68, 71}. I chose random numbers. But I think it would tend to be higher numbers, 50 or above, since there are more reds than yellows.
- Jerry (Post) {52, 58, 59, 60, 60, 62}. These numbers seems to co-relate to the number of red candies in the bucket.

The Level 2 responses included a specific indication of reasoning using an average, proportion, or a measure of spread:

- Emma (Pre) {3, 4, 5, 6, 7, 8} I'm sure you always pull at least 3 reds and at least a few yellows, so I just went from 3–up to 8. No formula, just guessing.
- Julie (Pre) {3, 5, 6, 6, 8, 8} I tried to give different results that together average 60%.
- Becky (Post) {56, 60, 60, 60, 61, 63} I chose these numbers because they have a mean, median, and mode of 60. All three are 60.
- Scott (Post) {40, 55, 58, 62, 65, 80} Because they reflect a mean of 60 or 6:4 which is the actual ratio of red to yellow in the container.

What distinguished the Level 3 responses was an indication of reasoning using *both* centers and spread:

- Daisy (Pre) {3, 4, 5, 6, 6, 7} The more times candies are grabbed, the more chance of the number of reds deviating from 60%.
- Roger (Pre) {4, 5, 6, 6, 7, 8} Reality does not obey the estimates of probability, so while 6 red candies remains the average outcome, variation is likely.
- Greg (Post) {56, 59, 60, 60, 61, 63} The numbers will vary. I gave a range of a minimum of 56 and a maximum of 63. 60 is the most frequently recurring number.
- Maria (Post) {48, 52, 55, 57, 63, 68} If I expect the average to be about 60, then I would guess that the amount chosen would vary above and below 60 & pretty close to 60.

Daisy's response, for example, clearly uses proportional reasoning to explicitly identify the 60% red in the underlying population. Yet she also explicitly acknowledges the expectation not only of a central anchor (such as 6 out of 10 candies being red) but also of variability in repeated samples. Her inclusion of both a measure of central tendency as well as an acknowledgement of variation is what makes this an example of a Level 3 response.

ANALYSIS

In considering the Level 0 responses, Kate and Rob's narrow expectations are obviously over-influenced by the expected value, but it seems surprising that more subjects did *not* put all 6s or 60s for their choices in the PreSurvey or PostSurvey, given results discussed by other researchers (e.g., Shaughnessy et al., 1999). In fact, the relative number of preservice teachers who gave narrow responses was less than that reported by Shaughnessy et. al. (2004), whose research involving 93 high school students and a sampling task showed almost 26% of responses being narrow. In contrast to those subjects who were over-influenced by the expected value, Alice's low response raises questions about whether she is able to identify the expected value at all. Although Susie's explanation does acknowledge variation, her choices result in an unlikely wide range with a particularly implausible lower bound. This compares with responses reported by Reading and Shaughnessy (2004) in their discussion about how students may overestimate the kind of variation one could reasonably expect.

In Level 1, while Sally and Gary have identical choices, they give different reasons. Sally reveals what "close to 6" means for her (within two). A theme common to the other three responses (including Gary's) is the reliance on additive as opposed to proportional strategies. That is, the focus is on the actual numbers in the jars as opposed to the ratios of red to yellow candies. Note the explicit nature of Gary's response, where he looked at the difference "20 more reds". Molly and Jerry's responses also had an additive flavor, where the numbers themselves seemed to influence their choices. This idea of using additive instead of proportional reasoning was also detected in middle and high school students by Shaughnessy et al. (2004). Reading and Shaughnessy (2004) describe how their subjects–six elementary and six secondary students–discussed frequencies of colors as a cause of variation.

Looking at the Level 2 responses, Emma appeals to her sense of a reasonable spread, and in fact she distributes her six choices so there are no repeated values–every possibility from three to eight (inclusive) is listed. Her language suggests she'd be surprised if two reds occurred, since she believes results would "always" include at least three reds. The other three

responses have an interesting commonality in that the choices were deliberately selected to reflect an average that was equal to the expected value of 6 or 60 reds. From class discussion and subsequent interview probes, it seems a widespread belief that while results even from a small sample such as six handfuls might vary, the average should still be the expected value. Hierarchies reported by other researchers (e.g., Torok & Watson, 2000; Reading & Shaughnessy, 2004) include the notions of deviations from a central anchor, but here attention has been drawn to the idea that the average of results from repeated samples should reflect the proportion of the underlying population. Scott's choices were even more interesting in that they did not actually include the expected value of 60.

There were more Level 3 responses in the PostSurvey than in the PreSurvey, and the relative increase in sophistication is apparent as subjects reconcile the tension of having results be close to an average value while also acknowledging the presence of variation. For example, note how Maria clearly has a sense of the expected value as "60 Red" but also is attentive to how that value may not show up at all in six repeated samples. Moreover, she is careful to claim the average of her six choices as "about 60" in contrast to other subjects who specifically wanted variability in the six choices while still maintaining an average of exactly 60 (i.e., by having symmetry around 60 with choices such as 50, 55, 60, 60, 65, 70). This idea reflects the upper level in Reading and Shaughnessy's hierarchy in describing deviations from a central anchor, in which responses indicate "consideration had been given to both a center and what is happening about that center" (2004, p. 216).

An interesting feature in the responses was that there were more subjects in the PostSurvey than in the PreSurvey whose choices did not include the expected value (such as Molly, Scott, and Maria), suggesting that the class experiences helped counter the natural tendency to pin expectations solely to a theoretical average without an appreciation of the variation in repeated trials. For example, when we considered the large population of 600 red and 400 yellow, students could see for themselves that a handful of 100 candies usually held something other than 60 red, and so some students seemed to deliberately avoid listing 60 in their expectations.

Another interesting feature in responses was the tendency to avoid repeating choices when making predictions for multiple trials in the PostSurvey. For example, when responding to the PreSurvey, most students gave some repeated values for their choices, such as Sally and Gary's {5, 6, 6, 7, 7, 8} or Julie's {3, 5, 6, 6, 8, 8}. There is nothing wrong per se with having repeated values in six conjectured results, especially for handfuls from the smaller jar used in the PreSurvey, but most of the PostSurvey choices contained no repeated values for handfuls from the larger jar. In fact, only in the PostSurvey and during subsequent interview probing did subjects comment about their choices being "similar but not identical" and how "there are no repeats" in their list. Molly, Scott, and Maria, like many others, not only made all six choices different, but they seemed to deliberately avoid including the expected value among their choices in the PostSurvey.

DISCUSSION

One reasonable hypothesis for why the class as a whole seemed to shift to a greater awareness of variation in results stemming from a sampling experiment is that their collective engagement in the activities, simulations, and subsequent class discussions made them more expectant of variability in data. More than half of the students referred directly to class activities or simulations in explaining their PostSurvey thinking, and comments like these are representative:

- Dixie: In our class experiments, when I repeated an experiment you'd often have some new variations pop into the picture, but the central probability remains the same.
- Rosie: Because we had the same activity in class, the same concept: The more chances or tries you have more different answers you can get.
- Frida: I based it on the activities we have done in class with computer program as well as hands-on activities.
- Sheila: I know this because we saw it on the computer program in class.

The hypothesis that the class interventions had an effect on improving student responses has credence from other research. In similar work with precollege students, Shaughnessy et al. (1999) found that there was "considerable improvement in the students' responses after they actually did the experiment" (p. 15) in simulating the Candy Task. Also, Reading and Shaughnessy (2000) suggest that a computer simulation would be useful to display to students, which is what happened with this class of preservice teachers.

A sampling environment is only one of many contexts for looking at variability, but having subjects reason about both centers and spread gives a strong foundation for considering distributions that relate to real data as well. For example, in research using a task concerning the national average height of 18-year-old American males, results indicated that the undergraduate subjects used information on sample size more accurately when dealing with centers of distributions rather than the tails. However, even when the subjects had received instruction on sampling distribution, "many of them still did not understand how sample size influenced the variability of the sample mean" (Well, Pollatsek, & Boyce, 1990, p. 310). The point brought out by Well et al. and others is that a consideration of both centers and spread is critical to distributional reasoning.

By way of conclusion, it should be remembered that although this paper only reports on one facet of the preservice teachers' understanding–the conceptions of variability in a sampling context–the relevance for training teachers extends to other contexts such as probability experiments and gathering data as a part of statistical inquiry: Have students discuss their predictions ahead of time, then complete the experiment, and finally discuss their actual findings in comparison to their predictions. In doing so, more students are apt to pick up on variability inherent in the situation, hopefully strengthening their overall attention to variation.

REFERENCES

Finzer, W. (2001). Fathom! (Version 1.16) [Computer Software]. Emeryville, CA: Key Curriculum Press.

Reading, C., & Shaughnessy, J. (2000). Students' perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Annual Meeting of the International Group for the Psychology of Mathematics Education*. Hiroshima: Japan.

Reading, C., & Shaughnessy, J.M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201- 226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics*, Vol 1, (pp. 314-319). Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society, Cape Town, South Africa*. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy, J. M., Canada, D., & Ciancetta, M. (2004). Types of student reasoning on sampling tasks. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (vol. 4, pp. 177-184). Bergen, Norway: Bergen University College.

Shaughnessy, J., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgement of statistical variation. In C. Maher (Chair), *There's more to life than centers.* Presession Research Symposium, 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematical Education Research Journal, 12*(2), 147-169.

Well, A., Pollatsek, A., & Boyce, S. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47,* 289-312.