ELEMENTARY PRE-SERVICE TEACHERS' CONCEPTIONS OF VARIATION IN A PROBABILITY CONTEXT

DANIEL CANADA
Eastern Washington University
dcanada@mail.ewu.edu

ABSTRACT

While other research has begun to contribute to our understanding of how pre-college students reason about variation, little has been published regarding pre-service teachers' statistical conceptions. This paper summarizes a framework useful in examining elementary pre-service teachers' conceptions of variation, and investigates the question of how a class of pre-service teachers' responses concerning variation in a probability context compare from before to after class interventions. The interventions comprised hands-on activities, computer simulations, and discussions that provided multiple opportunities to attend to variation. Results showed that there was overall class improvement regarding what subjects expected and why, in that more responses after the interventions included appropriate balancing of proportional thinking along with an appreciation of variation in expressing what was likely or probable.

Keywords: *Statistics Education Research; Teacher Education; Variation; Probability*

1. INTRODUCTION

The purpose of this paper is to report on research aimed at elementary pre-service teachers' conceptions of variation. Other research has already begun to illuminate pre-college student thinking about variation in several contexts, such as sampling (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999; Torok & Watson, 2000; Reading & Shaughnessy, 2004), data and graphs (e.g., Watson & Moritz, 1999; Meletiou & Lee, 2002; Reading, 2004), and probability situations (e.g., Truran, 1994; Shaughnessy, 1997; Shaughnessy & Ciancetta, 2002). Moreover, in keeping with the centrality of variation to the entire discipline of statistics (Cobb & Moore, 1997; Wild & Pfannkuch, 1999), an entire issue of the *Statistics Education Research Journal* focused on "research on reasoning about variation and variability" (Jolliffe & Gal, 2004).

As the picture begins to get painted about how pre-college students reason statistically, a continuing focus for research also needs to be on the teachers of these students: What sort of subject matter knowledge and pedagogical content knowledge do teachers have to prepare them for teaching and assessing in ways that enhance their own students' learning (Shulman, 1986)? Some researchers have incorporated inservice teachers into their studies, such as when Hammerman and Rubin (2004) looked at using statistical software tools in a professional development seminar and in the teachers' own classes. Garfield and Ben-Zvi (2005) begin to address the need to tie together the ways in which research informs practice as they present an epistemological model for teaching and assessing variability, but the research on how teachers reason about variation, or variability in data, remains thin.

Meanwhile, there is a paucity of research about how pre-service teachers think about variation. Makar and Confrey (2005) report on informal language used by secondary pre-service teachers to describe variation while reasoning about distributions, but little has been published regarding elementary pre-service teachers' conceptions of variation. Since university teacher preparation programs are concerned with both the subject matter knowledge as well as the pedagogical content knowledge of teacher candidates, it makes sense to attempt to identify the pre-service teachers' conceptions of variation. If a goal is for teachers to "provide students with authentic, inquiry-based tasks meant to develop children's reasoning about variation" (Makar & Canada, 2005), then a natural step in achieving this goal is to improve teacher training courses. By discerning components of pre-service teachers' reasoning, teacher educators can better design university experiences that promote an understanding of variation for pre-service teachers, as well as an understanding on how pre-college students come to learn this topic.

Therefore, doctoral research (Canada, 2004) was undertaken to explore which components of a conceptual framework help characterize elementary pre-service teachers' thinking about variation, how their conceptions of variation before an instructional intervention compared to those conceptions after the intervention, and what tasks were useful for examining their conceptions of variation in the contexts of sampling, data and graphs, and probability. Because so little is known about pre-service teacher knowledge in statistics education, the ultimate purpose of the research was to take first steps in discovering what pre-service teachers think regarding situations where variation was a key component. Since almost all of the subjects had no recollection of having ever taken prior courses that included any probability or statistics, their initial responses at the outset of the research may be considered intuitive. In that sense, the rationale behind the research was that by taking subjects who initially had little familiarity with the cognitive tasks being posed, and then studying how the thinking of those subjects changed from the more intuitive to the more substantive over the course of instructional interventions, key aspects of the subjects' thinking might be revealed.

In addition to summarizing the conceptual framework emerging from the doctoral study (within which the results of this paper are situated), this paper presents results for the following research question: How do elementary pre-service teachers' responses concerning variation in a probability context compare prior to and after class interventions? To address this question, first the overall methodology of the study will be presented, with particular emphasis on the structure of the class interventions and the nature of the survey and interview tasks. Then, the conceptual framework used to examine elementary pre-service teachers' reasoning about variation will be discussed, and connections will be made to past and recent models posited by other researchers. After presenting the research results and accompanying analysis, a summary and implications for teacher training programs follow.

2. METHODOLOGY

2.1. PARTICIPANTS AND DESIGN

The thirty subjects in the study of elementary pre-service teachers (24 women, 6 men) were enrolled in a ten-week pre-service course at a university in the northwestern United States. The course, Math for Elementary Teachers II (MET II), is designed to give prospective teachers a hands-on, activity-based mathematics foundation in geometry and probability and statistics. The only prerequisite course (MET I) focused mainly on whole and rational numbers and their operations. Thus, it was not expected that the subjects had

received any prior formal instruction in probability or statistics. When asked at the outset of MET II what classes involving probability or statistics they recalled taking (earlier in college or in high school), almost all subjects thought they had either not had any such classes, or they had taken such classes so long ago that they couldn't remember any specifics. For obtaining licensure to teach grades K-8, the two-course sequence of MET I and II represents the only content-specific math classes required at the university.

For the structure of the MET II course, Weeks 1 – 4 and 9 – 10 were on geometry and the four-week span from Week 5 – 8 was devoted to probability and statistics. During the first week of the course, 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 PreSurvey probability questions reported on in this paper considered a hypothetical person Mark tossing a fair coin 50 times (a *set* of 50 tosses), with a focus on how many of the 50 tosses landed heads-up. The questions are given in Table 1.

Table 1. PreSurvey Q7 (Sets of 50 Flips of a Fair Coin)

Subquestion	Nickname	Description
Q7ai	One Set - What	How many times out of 50 flips do you think the coin might land heads-up?
Q7aii	One Set - Why	Why do you think this?
Q7b	Compare Sets	After Mark's first set of 50 flips, he decides to do a second set of 50 flips. How do you think his results on the second set of 50 flips will compare with the results of his first set?
Q7ci	Six Sets - What	Mark actually has a lot of time on his hands, so the next day he does 6 sets of 50 flips. Write a list that would describe what you think might happen for the number of flips out of 50 the coin would land heads-up (in each of the 6 sets of 50 flips).
Q7cii	Six Sets - Why	Why did you choose those numbers?

The PreSurvey also contained an invitation for subjects to participate in individual interviews after regular class hours. The purpose of the individual interviews was to have an open-ended time where subjects could expand on their views and provide deeper explanations than were possible on the surveys. Eleven subjects volunteered to be interviewed, and all eleven were scheduled for one-hour interviews (called PreInterviews) before commencing instruction in probability and statistics so that their conceptions of variation could be further explored. The interviews were videotaped and included some of the same questions that were on the surveys so that subjects' verbal responses could be compared with what they had written earlier and extensions to their thinking could be probed. Thus, the interview script contained specific questions that were used with each subject, but the protocol also allowed flexibility to follow each individual subject's unique train of thought.

During Weeks 5 – 8, a series of activities was conducted in class specifically designed to offer opportunities to investigate and discuss variation. The activities (comprising the Class Interventions) were centered on the three realms of data and graphs, sampling, and probability situations. The instructor of the course, Sam, allowed me to co-lead many of the activities in class with him. Take-home surveys (called PostSurveys) were given after each class intervention. The PostSurvey probability questions reported on in this paper considered a hypothetical person Matt spinning a half-black and half-white spinner 50 times (a *set* of 50 spins), with a focus on how many of

the 50 spins landed on black. The questions (Table 2) were isomorphic to those asked in the PreSurvey:

Table 2. Probability PostSurvey Q1 (Sets of 50 Spins of a ½ - Black, ½ - White Spinner)

Subquestion	Nickname	Description
Q1ai	One Set - What	How many times out of 50 spins do you think the arrow might land on black?
Q1aii	One Set - Why	Why do you think this?
Q1b	Compare Sets	After Matt's first set of 50 spins, he decides to do a second set of 50 spins. How do you think his results on the second set of 50 spins will compare with the results of his first set?
Q1ci	Six Sets - What	Matt actually has a lot of time on his hands, so the next day he does 6 sets of 50 spins. Write a list that would describe what you think might happen for the number of spins out of 50 the spinner would land on black (in each of the 6 sets of 50 spins).
Q1cii	Six Sets - Why	Why did you choose those numbers?

After shifting topics in the course from probability and statistics back into geometry for the final two weeks of the course, the same eleven subjects who participated in PreInterviews also participated in PostInterviews. In the PostInterview, subjects were asked to elaborate on their PostSurvey responses concerning the spinner tasks, using a protocol similar to that in the PreInterview.

While interview data were gathered from eleven subjects, only six representative cases were chosen from among those eleven for this paper. The reason for this selection was that the grounded-theory approach (used in discerning the aspects of the conceptual framework that was a main contribution of the research) enabled a point of saturation to be reached, beyond which new data was not adding anything new to the framework. Thus, taken cumulatively, the responses from the six interview subjects profiled in this paper may be seen as representative of the class as a whole.

2.2. CLASS INTERVENTIONS

In this section, the class interventions, comprised of activities around which much of the class discussion was based, are described in more detail These interventions are presented in the order in which they occurred in the MET II course, leading off with the context of data and graphs, then sampling, and finally probability situations.

Class Intervention #1 (Data & Graphs) The two activities comprising the Class Intervention for the context of data and graphs were called "Four Questions" and "Body Measurements." The first activity offered a good opportunity to discuss both average and spread in data sets, and Sam started the class exploration of statistics in the fifth week by having the entire class gather data from one another in response to four questions:

Four Questions Activity Prompt
How many pets do you have?
How many years have you lived in this city (to nearest half-year)?
How many people are in your household?
How much change (in coins) do you have today?

After graphing the data in different ways, the class had a discussion about levels of detail provided by each type of graph and about "typical" values for an individual student and for the whole class. The tension between centers and spread of data was one theme to emerge from the discussion of the graphs. For the second activity, everyone's own armspan, height, handspan, head circumference, and pulse rate per minute were recorded. Also, all students in class measured a designated person's armspan, to see how multiple measurements of the same object would compare. Again, we had a class discussion about the data and graphs for the body measurements, this time focusing more on causes of variation.

Class Intervention #2 (Sampling) In the seventh week of class, the two activities "Known Mixture" and "Unknown Mixture" were conducted with Sam's students. Prior to the "Known Mixture," we started with a general discussion of what samples were, who uses samples, and the purpose of sampling. Then the following scenario for the Known Mixture Activity was given as a part of a handout:

Known Mixture Activity Prompt

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 of the band members in a hat and then they randomly draw out 10 names.

The class discussed initial expectations for this scenario, focusing especially on what would happen if the random draw of 10 names were to be repeated thirty times. After students talked about predictions for drawing thirty samples each of size ten, we simulated this activity using chips in a jar. Actual data were 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 then made a transition into the second activity in this intervention, the Unknown Mixture. Now we had larger jars that each contained 550 yellow and 450 green chips, and the use of opaque jars having only a narrow opening made it difficult to look inside at the contents. Students were only given the information that each jar had 1000 total chips, and that the mixture was identical across all jars (they were not told the true mixture). To make a conjecture about the true mixture of chips, the students were asked to decide in their groups what sample size they wanted to use and how many samples they wanted to draw. Then they were to carry out their plans, do the sampling, graph the results, and make their conjectures about the true mixture in the jar. After the sampling was carried out, we had a class discussion about the different choices made in sampling and the class results, and we tried to forge a class consensus about what the true mixture was.

Class Intervention #3 (Probability) There were two activities that made up this intervention, "Cereal Boxes" and "The River Crossing Game." These were chosen specifically because of the probability aspects involved in the activities and these were the main class activites involving random devices. Cereal Boxes relies on the use of spinners and River Crossing on the use of ordinary fair dice as random generators.

Cereal Boxes actually took place in the first class session of week 6, just before we gathered data for Body Measurements. As explained earlier, there was considerable overlap in the three contexts, and Cereal Boxes is a good example of this overlap. Cereal

Boxes is a sample-until scenario, assuming that any of five different stickers can be obtained within each box of cereal, and that the five stickers have equal chances of being obtained. The question concerns how many boxes need to be opened to obtain all five stickers, and the situation was simulated by using an equal-area five-region spinner. Cereal Boxes brings together probability, sampling, and data and graphs in a way that highlights variation.

The second activity for this intervention, the River Crossing Game, involved finding the sum of the scores on two dice. Both the Cereal Boxes activity and River Crossing Game are part of the *Math and the Mind's Eye* curriculum (Shaughnessy & Arcidiacono, 1993). Using two players, each player receives 12 chips to place on their side of a "river," along spaces marked 1 through 12. After configuring their chips in an initial arrangement along the spaces, players take turns tossing a pair of dice. If either player has any chips on the space showing the total sum for the dice, one chip can "cross the river" and be removed from the board. The winning player was the first one to remove all the chips on his or her side. Note that although a sum of 1 is not possible to obtain, this fact is left for the players to discover; the challenge remains for players to reason about the optimal placement of their 12 chips. As with Cereal Boxes, in the River Crossing Game we made predictions, gathered and graphed data, and discussed results.

The activities in all the interventions were designed to elicit discussion about variation. For instance, the intervention on data and graphs included different types of graphs and the amounts of variation they showed. Body Measurements got at the ideas behind multiple measurements of the same object, whereas the Known and Unknown Mixtures had students actually draw chips from a container to experience variation resulting in a sampling context. Cereal Boxes and the River Crossing Game had students use traditional random generators such as spinners and dice to get a sense of what was likely in a probability context. Software, including ProbSim® (Konold & Miller, 1994) and Fathom™ (Finzer, 2001), was used to aid construction of graphical representations and to extend the simulations that the class had already participated in manually.

3. CONCEPTUAL FRAMEWORK

The doctoral study (Canada, 2004) had as one of its goals the description of components of a conceptual framework to help characterize elementary pre-service teachers' thinking about variation. Although different frameworks have been described by other researchers (e.g., Jones, Mooney, Langrall, & Thornton, 2002; Watson, Kelly, Callingham, & Shaughnessy, 2002), pre-service teachers were not the subjects of such research. Thus, the approach used in this study was largely exploratory, relying heavily on grounded theory, to let the data provided by the subjects help fill in details of what was called an *Evolving Framework.* Data which helped develop the framework included the written PreSurvey and three PostSurveys, transcriptions from all of the PreInterviews and PostInterviews, and observations from the class interventions.

The framework provides a lens through which three different *aspects* of an elementary pre-service teacher's understanding of variation can be viewed. The three aspects address how subjects reason in terms of *expecting*¸ *displaying*, and *interpreting* variation; these aspects are then defined in terms of their constituent *dimensions* (Figure 1). The following subsections describe the dimensions of the framework in terms of major themes that emerged from the participants in the study, and connections to framework components posited by other researchers are also discussed.

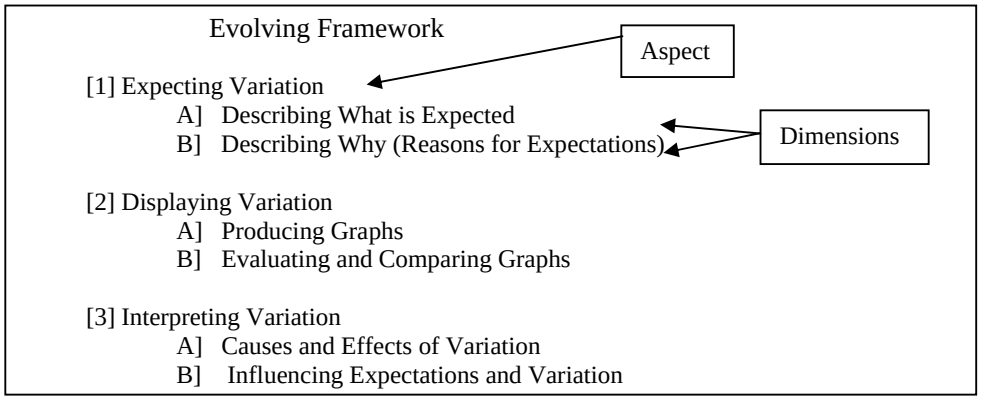

Figure 1. Framework for elementary pre-service teachers' conceptions of variation

3.1. EXPECTING VARIATION

When expecting variation, such as prior to sampling or conducting probability experiments, subjects expressed both *what* they expected and *why*. The expected value or average was a frequent theme concerning *what* subjects thought might occur in predicting results for experiments involving uncertain outcomes. A dominant type of response was how results should be close to, about, or near the expected value, and a more explicit type of response was how results might be higher or lower than the expected value. Another theme for *what* was expected concerned ranges or extreme values. More specific than just suggesting that results be above and below the expected value, some responses actually specified a numeric range.

In describing *why* they held their expectations, almost all subjects' reasoning at some point involved the language of possibilities and likelihood. For example, many subjects explained how extreme results were possible but unlikely. Reasons involving personal or shared experience constituted another theme. Some subjects mentioned informal out-of-class experiences, such as games they had played, and other subjects recalled their experiences with the in class activities. The theme of proportional reasoning can be a useful anchor to help center expectations appropriately, and this theme was a part of many subjects' reasons for *why* they expected what they did. An over-reliance on proportional reasoning can lead to a restricted expectation of variation, but an under-reliance on proportional reasoning can also lead to poor expectations. Some subjects were less influenced by proportional reasoning than by additive reasoning, as in the case of sampling experiments where the quantities in a sample or a population were more persuasive than the proportions.

3.2. DISPLAYING VARIATION

Subjects showed their skills and reasoning along the two dimensions of *producing graphs* as well as *evaluating and comparing graphs*. In considering how subjects *produced* graphs for tasks such as predicting the results of 50 samples of size 10 taken from a jar of 60 red and 40 yellow candies, the technical details of their graphs were a reflection of the subjects' own graph sense. Subjects often drew smooth bell curves in situations when a bar chart or dotplot would have been a better choice. Some graphs had detailed axes with an appropriate scale, while others had unlabeled axes or inappropriate scales. How the characteristics of the distribution get conveyed is another theme. Subjects

generally gave centers that were reasonably placed, but they often provided ranges that were too wide. Spreads were occasionally too tight or too scattered, and shapes were often unnaturally symmetric.

When *evaluating and comparing graphs*, such as comparing average annual traffic rate deaths between two regions of America, the four themes exhibited by subjects' responses corresponded to four components of distributional reasoning: Average, range, shape, and spread. A focus on average, was reflected in most but not all the subjects' responses. Many subjects were able to move beyond a focus on average to include references to other features of the distribution, but some made it clear that the average was their primary consideration in answering any question having to do with graphs. The theme focusing on range or extremes was often reflected in questions having to do with which graph had more variation. Subjects had some standard ways of talking about shape, using language like "symmetric" or "skewed," "normal" or "uniform." There were also some non-standard ways of referring to the shape of a distribution, including the use of hand gestures to try to communicate the picture in the subject's mind. Responses that focused on the theme of spread depended on the type of display that subjects were considering. For example, when using dotplots, some subjects referred to the way data were "clustered" or "scattered" along the horizontal axis to indicate how they saw the way data were grouped or spread out in the graph. In using boxplots, class discussions had focused on how the interquartile range was one measure of spread, and many subjects referred to that measure in their responses.

3.3. INTERPRETING VARIATION

Two dimensions that arose in the data for this aspect were *causes and effects of variation*, and *influencing expectations and variation*. Under *causes and effects of variation*, one theme reflected in the data was naturally occurring causes, such as the geographical and meteorological causes subjects listed for differences in rainfall patterns between two cities. In the sampling and probability situations, many students seemed to point to randomness as a naturally occurring cause. The other theme of physically induced causes included those causes which were deliberate or intentional as opposed to naturally occurring. For example, lining up the spinner in the same spot for each spin and trying to apply the same amount of force each time was seen a physically induced cause for reduced variation. The *effects of variation* were seen in terms of two distinct but related themes: the effect of variation on students' perceptions and the effect of variation on their decisions. For example, some students perceived a difference between theoretical predictions and real-life outcomes, and many students perceived that "anything can happen" in situations involving variation. Also, some students expressed a lack of confidence in making decisions, reflected in their "I don't know" responses. In making inferences, it seems that the two themes for *effects of variation* were often linked. For example, a student who thinks that "anything can happen" may be thinking that there is no way to decide what might happen, and thus the student may respond with "I don't know."

The two themes for *influencing expectations and variation* were quantities in sampling (i.e., the numbers of candies in the population or in the sample) and also the numbers of samples taken. The first theme applied primarily to the context of drawing samples where there was a discrete population, such as samples of candies from a jar containing 60 red and 40 yellow candies. Several subjects focused on the sheer numbers of candies in the jar, and in some cases it seemed that the probabilities of getting different outcomes were linked to these quantities. Particularly for subjects who are not strong

proportional reasoners, there may be a tendency to see the quantity and not the ratio as the influential factor in the behavior of the sample outcomes. The second theme, involving the numbers of samples taken, was reflected in many different ways. Almost all of my subjects pointed out that more samples would widen the overall range, while very few subjects suggested that more samples would also tighten the subrange capturing most of the results. Other ideas included how additional samples offered more chances to attain the expected value, and how additional samples provided a better picture of the underlying distribution.

3.4. RELATION TO OTHER MODELS

The summary of the Evolving Framework provided in this section captures some of the main ways in which subjects expressed their intuitive and emerging conceptions of variation throughout the doctoral study (Canada, 2004). Grounded in survey, interview, and classroom observation data, the framework provides structure for characterizing elementary pre-service teacher thinking about variation in the contexts of sampling, data and graphs, and probability situations. The framework is "evolving" because there are no doubt more ways in which elementary pre-service teachers' understandings of variation can be modeled, and the framework is expected to grow as more comparisons to other models of thinking are made. Already the aspects of the evolving framework reflect facets of other models. For example, Wild and Pfannkuch (1999) incorporated acknowledging, measuring, modeling, and explaining variation within their components of a model for statistical thinking. Acknowledging variation is involved when explaining what is expected regarding variation, and also relates to producing, evaluating, and comparing graphs when dealing with displays of variation. Explaining variation relates both to explaining why people expect what they do and also to causes of variation. To the model of Wild and Pfannkuch, Reading and Shaughnessy (2004) added the two components of describing and representing variation. Reading and Shaughnessy's description hierarchy reflected what was expected in terms of extreme and central values, and also how expectations deviated from an anchor. Also, a causation hierarchy included extraneous (physical) causes of variation as well as the reason why results might vary, such as additive or proportional reasoning. More recently, Garfield and Ben-Zvi (2005) proposed the following seven dimensions of a theoretical framework representing key facets of understanding variation, or variability in data:

(1) Developing intuitive ideas of variability
(2) Describing and representing variability
(3) Using variability to make comparisons
(4) Recognizing variability in special types of distributions
(5) Identifying patterns of variability in fitting models
(6) Using variability to predict random samples or outcomes
(7) Considering variability as part of statistical thinking

The framework proposed by Garfield and Ben-Zvi (2005) provides a comprehensive structure for looking at how people reason about variation and incorporates multiple aspects of other researchers' models of conceptualizing variation. For the evolving framework looking at elementary pre-service teachers' conceptions, certainly their intuitive ideas were explored in terms of what variation they expected and why. How elementary pre-service teachers dealt with displays of variation addressed the ways in which they described and represented variation, and also how they used variation to

compare distributions. Subjects also had primitive ways of using variation in making predictions for what they expected when sampling or considering probability outcomes.

While the evolving framework had as its three aspects *expecting, displaying,* and *interpreting* variation, the focus of this paper is on survey and interview results for a small subset of questions, namely the PreSurvey questions on coin flipping and the analogous PostSurvey and PostInterview questions concerning half-black and half-white spinners. These questions (presented earlier in Tables 1 and 2) did not encompass displays of variation, so the aspects of the framework that situate the results for this paper are how elementary pre-service teachers *expected* and *interpreted* variation.

4. RESULTS AND ANALYSIS

Results are first presented showing how (as a class) subjects' written survey responses concerning variation in a probability context compare before and after the class interventions. The comparison is facilitated by a scoring rubric derived from previous research and relating to the conceptual framework. Assessing paired-data results according to the rubric bolstered the claim of overall class improvement from the PreSurvey to the PostSurvey. Then, in the second and third subsections, results from both survey and interview questions are used to focus more sharply on changes in subjects' thinking according to two aspects of reasoning about variation in a probability context that emerged. These two aspects concern the subjects' expectations and interpretations of variation. Thus, the first subsection provides more of a quantitative backdrop for looking at overall class shifts in thinking, while the second and third subsections give more of a deeper, qualitative picture of elementary pre-service teachers' thinking about variation in accordance with the conceptual framework for the study.

4.1. CLASS SURVEY PERFORMANCE

To compare class performance on questions from the two surveys, coding schemes were adapted from rubrics used in a similar set of questions involving sampling candies out of a jar (Shaughnessy, Ciancetta, & Canada, 2004). After assessing the PreSurvey responses using the coding schemes, I had two colleagues (who were familiar with the original sampling rubrics) independently assess the responses. There was an initial inter-rater agreement of 91% on the PreSurvey, and all disagreements were subsequently resolved.

What follows are the class results for the PreSurvey and PostSurvey probability tasks described earlier, organized according to the task subquestions. Although class enrollment was 30, there were three absences on the day of the PreSurvey, which was completed during class time. After all the class interventions had taken place, the PostSurvey for Probability was the final written research instrument given in class, and was completed by 29 of the 30 enrolled students. The percentages are given of students who were coded at each of the levels for the different subquestions, and example responses from both surveys are also provided. Then changes in overall class performance on the different subquestions are discussed.

One Set The codes and class results for the first part of this subquestion are presented in Table 3:

Table 3. Results for One Set – what (PreSurvey Q7ai & PostSurvey Q1ai)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
2	Either gives a range around 25 such as 22-28, or else writes (for example) "Around 25."	1 (3.7%)	10 (34.5%)
1	Gives only 25 as answer.	21 (77.8%)	14 (48.3%)
0	Gives one number other than 25, such as 23.	5 (18.5%)	5 (17.2%)

As has been noted with similar sampling tasks involving predicted outcomes for one trial, most students put down the expected value (Shaughnessy et al., 2004), which in this case is 25. However, the number of students who volunteer some form of variability in their Level 2 response increased from PreSurvey to the PostSurvey. In fact, the average coding levels for class performance on this subquestion for both surveys go from 0.85 on the PreSurvey to 1.17 on the PostSurvey.

Even stronger evidence of class improvement is offered by the reasoning component to the subquestion, where subjects described why they held their particular expectation, and Table 4 has the codes and class results for the second part of this subquestion.

Table 4. Results for One Set - why (PreSurvey Q7aii & PostSurvey Q1aii)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
3	Uses proportional reasoning with some explicit statement about what else might happen.	2 (7.4%)	8 (27.6%)
2	Uses proportional reasoning (for example: ratio, average, or percent).	17 (63.0%)	17 (58.6%)
1	Uses additive reasoning, or gives a reasonable response which makes sense but lacks specificity.	4 (14.8%)	4 (13.8%)
0	No reason, a vague reason which makes no sense, or an irrelevant reason.	4 (14.8%)	0 (0.0%)

Here are some examples of responses from each coding level:

[Level 0]

Sarah (Q7aii) Maybe a little more than half 'cause it started on heads: I have no idea really.

[Level 1]

Rosie (Q7aii) Because you have the same chances.

Ross (Q1aii) There is no reason to expect that either white or black would result any more than the other, but an exact result isn't possible to predict.

[Level 2]

Emma (Q7aii) He has a 50% chance of landing on heads.

Brita (Q1aii) Because there is a 1 out of 2 (or 50%) chance that he will get black. So theoretically half of the spins will be black.

[Level 3]

James (Q7aii) The coin has a 1:2 chance of landing on heads. The more often you flip, the chances of the 1:2 ratio will be closer to that – 1 in 2.

George (Q1aii) There is a 50% chance of landing on white or black, in the long run it balances out closer and closer to 50%, but the short run it varies wider. 25 would be 50%, so 18 is very probable. It could be 28 or 20 or 16, but the more times, the closer it will be toward 25.

Additive reasoning was initially conceived as a level in connection with sampling tasks, where prior research had shown subjects to focus on the sheer numbers used in the sample or population (as opposed to a consideration of the proportion). In transferring the idea of additive reasoning to probability tasks, I had anticipated a focus on the amount of shaded area on a spinner, for example, or the number of sides for the coin. Some students did comment on the amount of shaded area in dealing with other spinners such as a 1:3 white-to-black spinner, but they did not use the same language for the 1:1 spinner used in this question.

Overall, there was a lack of specificity in the Level 1 responses, such as when Rosie wrote about the "same chances" without identifying what those chances were. Level 2 responses were characterized by the use of percentages, odds, or ratios, but little other information was usually given. In James' Level 3 PreSurvey response, he uses the ratio defined by the fair coin, but his response also suggests that the cumulative average of many flips approaches that ratio. Thus he shows thinking that aligns with the Law of Large Numbers, as does George's Level 3 PostSurvey response. Although George's written comment about "...the more times, the closer it will be toward 25..." doesn't make it clear what he is thinking about, in subsequent interviews it became apparent that he was envisioning the proportion of heads gravitating toward the theoretical 50% with increasing numbers of flips. One noticeable feature of Table 4 is how more students gave a Level 3 type of response in the PostSurvey than in the PreSurvey. Also, the average coding levels for class performance on this subquestion for both surveys go from 1.63 on the PreSurvey to 2.14 on the PostSurvey, again showing a sizable increase.

Compare Sets For this subquestion, a key idea to learn from responses was whether or not subjects believed that results on a second set of flips or spins would or should match the results from the first set. In sampling tasks used on the PreSurvey and PostSurvey and by other researchers (Reading & Shaughnessy, 2004; Shaughnessy et al., 1999; Shaughnessy et al, 2004), a similar question asked was "If several samples were taken, do you think you'd get the same results each time?" Subjects who were unduly influenced by the expected value often did answer affirmatively, with the idea being that if the expected value was reasonable for a single sample, then that same value was reasonable for several samples. In one study of 188 high school students, 25% agreed that results should be the same every time (Shaughnessy et al., 2004).

On the probability subquestions using flips and spins, the wording was changed (from what had been used on the sampling question described in the previous paragraph) so that subjects were invited to describe how results on a second set of 50 flips or spins might compare to the first set. The reason for the change in wording was to allow more flexibility in how subjects responded, since "...do you think you'd get the same results each time?" seemed to invite a straightforward yes-or-no kind of response. One thing to take into account with the less straightforward responses that the wording of the *Compare Sets* subquestions invited was that subjects often used "similar" to mean "similar but not

the same." Thus, the terms "different" and "similar" both occurred to signify "not the same." Classifying responses at different levels involved looking at other information provided showing *what* subjects expected and *why*. The codes and class results for this subquestion are presented in Table 5:

Table 5. Results for Compare Sets (PreSurvey Q7b & PostSurvey Q1b)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
3	[Different or Similar] with Explicit mention of a range or spread.	3 (11.1%)	15 (51.7%)
2	[Different or Similar] with Some additional information, such as use of ratio, average, percent, or giving specific alternatives for results.	11 (40.7%)	10 (34.5%)
1	[Different or Similar] with No additional information provided.	7 (25.9%)	4 (13.8%)
0	Mentions how results will be the same.	6 (22.2%)	0 (0.0%)

Over 20% of the PreSurvey responses indicated results would be the same, with responses such as:

[Level 0]

Ross (Q7b) In the absence of any change of approach, the results are most likely to be the same.

Maya (Q7b) Same. Probability will remain the same.

Ross's response points more to the physical aspects of doing the coin flips, implying that if the coin is flipped in the same manner, then obtaining the same results is "most likely." Maya shows she is influenced primarily by the constancy of the theoretical probability. No responses in the PostSurvey expressed a sense of expectancy that identical results would occur.

Level 1 responses only included an expression of difference or similarity, such as:

[Level 1]

Sally (Q7b) They will be similar but not the same.

Julie (Q7b) Similar, though probably a little different.

Jackie (Q1b) Could be slightly different, but basically the same.

Susie (Q1b) Probably very similar to the first set of results, keeping in mind that it is 'chance.'

Sally's response lends credence to the assumption that "similar" connotes "not the same," and judging by that assumption, most of the class in the PreSurvey (and all students during the PostSurvey) held the idea that results on the second set would likely not be identical to the first set.

Level 2 responses added additional information about what might happen or why:

[Level 2]

Carrie (Q7b) Probably different. But still has a 50/50 chance.

Emma (Q7b) He might get a few extra tails-up so his results should vary.

Cassie (Q1b) The comparison should be somewhat the same. It has the same odds again, 50%.
Robbie (Q1b) I think they will be very similar to the first set of 50 spins because the probability of getting black remains ½.

Whereas Emma clearly indicates an expectation of variation in results, what set apart the Level 3 responses was an explicit statement of *what* variation might result, or *how* results might vary:

[Level 3]
Maria (Q7b) It will be nearly the same, or the same. The variation may be only 2-3 one way or the other.
George (Q7b) Could be 30, 25, 20, 27.. .If he was super super super super lucky he'd get 50.
Molly (Q1b) Maybe a little different but still somewhere around 20-30.
Sofia (Q1b) Similar. Maybe a little wider range, 18-32.

Over half of the PostSurvey responses were at Level 3, suggesting that the class interventions helped attune students to thinking in terms of a range of expectations. In terms of class averages, again there was an increase in means, from 1.41 on the PreSurvey to 2.38 on the PostSurvey.

Six Sets Both parts of this subquestion (the *what* and the *why*) were taken into consideration for coding purposes, primarily to retain consistency with the analogous rubric derived for the similar questions in a sampling context (Shaughnessy et al., 2004). Only inappropriate choices for listing *what* was expected (or blank answers) were coded at Level 0. Deciding what would constitute an appropriate choice for the results on six sets of flips or spins involves making a judgment call, and the subcodes used for this subquestion question help identify inappropriate choices as (W)ide, (N)arrow, (H)igh or (L)ow. Of key interest was how many subjects had a narrow response consisting of just a list of six identical values, namely the expected value of 25. In research involving 93 high schoolers and a sampling task, almost 26% of responses were narrow, which was conjectured to be because of "an influence of probability instruction, or just lack or exposure to statistics tasks involving variability" (Shaughnessy et al., 2004, p. 6). The codes and class results for this subquestion are presented in Table 6.

Table 6. Results for Six Sets (PreSurvey Q7c & PostSurvey Q1c)

Coding Level	Description of Category	Number of Students (Pre)	Number of Students (Post)
3	Appropriate choice & Explanation explicitly involves proportional reasoning as well as variation.	2 (7.4%)	9 (31.0%)
2	Appropriate choice & Explanation reflects proportional reasoning or notions of spread.	10 (37.0%)	15 (51.7%)
1	Appropriate choice & Explanation left blank or lacks any specific reasons relating to details of the distribution.	4 (14.8%)	3 (10.3%)
0	Inappropriate choice (Regardless of Explanation). W(ide) = Range > 19, N(arrow) = Range < 2, H(igh) = Choices > 24, L(ow) = Choices < 26	11 (40.7%)	2 (6.9%)

Of the eleven inappropriate PreSurvey responses, one was narrow, one was high, one was low, and four were wide (the remaining were left blank). It was clear from subsequent discussions in class that students initially felt uncomfortable venturing a guess for six results, often demonstrating that it was difficult to guess correctly. Such an attitude toward expectation has much in common with the *Outcome Approach* to random events, whereby subjects look at the goal of probability as correctly determining ahead of time what will be the next outcome (Konold, 1989). Of the two inappropriate PostSurvey responses, both were wide.

A few of the Level 0 examples are:

[Level 0]

Alice (Q7c) {25, 25, 25, 25, 25, 25} I don't see how the chances of getting heads will change if he does more sets of 50 flips.

Brita (Q7c) {7, 21, 23, 25, 29, 31} I chose numbers close to 25 because I think with a 50% probability, the results would come out pretty close to 25. I put the oddball 7 in for fun, because there is always that element of chance.

Susie (Q1c) {5, 15, 30, 40, 45, 50} It is chance.

Alice's narrow response is obviously over-influenced by the expected value, but it seems surprising that more subjects did *not* put all 25s for their choices in the PreSurvey, given results discussed by other researchers (e.g., Shaughnessy et al., 1999). Brita's choice of 7 is extremely unlikely and makes her overall range too wide, although her upper bound of 31 is plausible). Susie's choices are too extreme at both the upper and lower ends. Level 1 responses had appropriate choices for *what* was expected but the reasons *why* did not specifically reflect distributional thinking:

[Level 1]

Carrie (Q7c) {22, 23, 24, 25, 26, 27} It's usually not the same.

Maria (Q1c) {20, 23, 25, 25, 26, 30} I think he will hit 25/50 one time. The rest of the times, he will be close, but not exactly on. Also I think he will be controlling the way he hits the spinner more on the second day, which accounts for no 23 or 28.

Maria points to *causes* of variation in noting the physical manipulation of the spinner, and other subjects also seemed to indicate that spinners are not viewed as true random devices because the user can ostensibly control outcomes by altering the way the pointer is spun. The Level 2 responses included an indication of reasoning using an average, proportion, or a measure of spread:

[Level 2]

Sofia (Q7c) {20, 20, 24, 25, 26, 27} Because they average to about 25.

Sally (Q7c) {22, 23, 24, 25, 26, 27} They are all close to 25, ½ of 50.

Leila (Q1c) {23, 24, 24, 25, 25, 26} The numbers are pretty close to half or 50%.

Rocky (Q1c) {20, 22, 23, 27, 28, 30} These numbers represent a distribution across range of likely results.

Note how Sally's Level 2 response includes the same choices as in Carrie's Level 1 response shown earlier. However, Sally gives more specificity than Carrie in describing

her reasoning, which is proportional in Sally's case. Rocky doesn't include the expected value in his choices, but feels he has given a likely range and his sophisticated language borders on a Level 3 response. What distinguished the Level 3 responses was an indication of reasoning using *both* centers and spread:

[Level 3]

Maya (Q7c) {23, 24, 25, 25, 25, 26} Because there should be variation around the mean. The average should be 25.

Ross (Q7c) {22, 23, 24, 26, 27, 28} While 25 flips are likely to be heads, in reality some variation is likely, so my numbers represent a range that averages 25.

Sally (Q1c) {21, 24, 25, 26, 28, 29} All numbers are 25 or close to 25 (1/2 of # of spins). Not all are 25 in order to account for variation.

Daisy (Q1c) {18, 21, 24, 26, 28, 31} Because they are close to the 50% chance to get 25 hits of black allowing for variation due to random spinning hits. But none of the #'s are too high or too low (far from the 25) which would be hard to hit based on the 50% odds.

There were more Level 3 responses in the PostSurvey than in the PreSurvey, and the relative sophistication is apparent as subjects reconcile the tension of having results close to an average value while also acknowledging the presence of variation. Also, there were more subjects in the PostSurvey than in the PreSurvey whose choices did not include the expected value of 25 (such as Susie, Rocky, and Daisy), 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 subsequent trials. As in the other subquestions, average class performance on *Six Sets* also increased, from a mean of 1.11 in the PreSurvey to 2.07 in the PostSurvey.

Further Analysis Although the PreSurvey and PostSurveys were completed individually, the entire structure of the class interventions was geared towards small-group and whole-class discussions. With the constant exchange of ideas, opinions, and explanations that went on throughout the course, it made sense to look at classwide changes from the PreSurvey to the PostSurvey. On each subquestion, average class performance increased, with more students being rated in the highest coding level for each subquestion's scoring rubric.

Although there were 27 subjects who took the PreSurvey and 29 who took the PostSurvey, there were 26 subjects who took both. Thus, *t*-tests for differences in mean scores were applied to the paired data, with μ_1 as the mean for the PreSurvey and μ_2 the mean for the PostSurvey.

Using a one-sided test with H_a: $(\mu_2 - \mu_1) > 0$, and a significance level of $\alpha = 0.05$, statistically significant gains were found in all subquestions. Table 7 contains the p-values associated with each subquestion.

4.2. EXPECTATIONS OF VARIATION

Having seen evidence of classwide changes on the surveys, the following subsections focus on situating these changes more fully within the aspects of the conceptual framework. To help describe key aspects of understanding variation which emerged from the subjects, and how their expectations and interpretations changed throughout the course, examples of thinking provided by six interview subjects (Ross,

Table 7. Paired-data results for difference of means

Subquestions	Nickname	$\mu_2 - \mu_1$ Mean (Std. Dev)	*t*-test (25 df) p-values
Q7ai & Q1ai	One Set – What	0.27 (0.67)	0.0250
Q7aii & Q1aii	One Set – Why	0.46 (1.03)	0.0150
Q7b & Q1b	Compare Sets	0.96 (1.25)	0.0003
Q7c & Q1c	Six Sets – What & Why	1.00 (1.30)	0.0003

George, James, Daisy, Emma, Sandy) will be used along with examples from the class surveys and class discussions.

Describing What is Expected An important change from before to after the class interventions was how initial expectations that were overly influenced by the expected value became tempered by an increased appreciation for how variation occurs in multiple trials. Many students initially were inclined during the PreSurvey and PreInterview to think that the theoretical expected value was what *should* happen on any given trial (whether a sample drawn from a population or a set of flips or spins). These students also thought that even if results varied, the average of results *should* be the expected value. Such a perspective reflects the Law of Small Numbers described by Kahneman and Tversky (1972). By the end of the course, many students were speaking more about expectations in terms of a range rather than in terms of a single given value. For example, on the PreSurvey all six interview subjects expected 25 for the *One Set* of 50 flips, pointing out how 25 was the expected value. But for the *One Set* of 50 spins, five of the cases described their expectations on the PostInterview in terms of ranges. Ross expected "somewhere between 21 and 29...I would say – it's probably within that range," and James thought the result would be "approximately" 25, adding that "it will be, you know, plus or minus, maybe, 20% of that number – Somewhere in there." Daisy talked about the result being "within 2 or 3" of the expected value, and Emma and Sandy both said it would be "between 20 and 30 spins" for the result of black. Moreover, on *Compare Sets* in the PreSurvey, Ross felt the results on the second set "most likely were the same" as on the first set, whereas he had a different idea on PostSurvey, writing that the second set would likely show "not the exact same results." As he explained during the PostInterview, "I think that it's likely to fall in a same range, similar range." Sandy's comment was that "I think the range would still be somewhere very similar to that (first) one."

Another trend in responses was to avoid repeating choices when making predictions for multiple trials in the PostSurvey and PostInterview. For example, in giving choices for *Six Sets* on the PreSurvey, most students gave some repeated values for their choices, such as James' (20, 22, 25, 25, 26, 27) or Daisy's and Emma's (23, 24, 25, 25, 26, 27). There's nothing wrong per se with having repeated values in six conjectured results, but it is interesting how all of the interview subjects (and most of the other class members) had PostSurvey choices for *Six Sets* that contained no repeated values. Ross's choices on the PostSurvey *Six Sets* were (22, 23, 24, 26, 27, 28), and in the PostInterview he amended those to have a slightly wider range (21, 22, 23, 27, 28, 29), pointing out how his choices still were "similar, but not identical" and how "there's no repeats." Sandy said of her PostSurvey choices (20, 23, 24, 26, 28, 29) that "they could repeat, but I just did a

range – from 20 to 30, just to choose... different numbers, but still somewhere in that range." Ross and Sandy, like many others, also seemed to deliberately avoid including the expected value among their choices in the PostSurvey.

Describing Why (Reasons for Expectation) Improvements in reasoning for *why* students held their expectations included an emphasis on proportional reasoning combined with an understanding of what was probable or likely in the face of variation. Reading and Shaughnessy (2004) placed likelihood based proportional reasoning at the top of their causation hierarchy, which was related to why students gave their responses. In responses from the PostSurveys and PostInterviews, repeated results were unexpected because they were seen as unlikely, and extreme values were often described as unlikely but possible. Subjects also used probabilistic language in a general way, for instance talking of how the chances for events were seen as high or low, or about what might or could happen. For example, note the subjective use of language as Daisy reasons in the PostInterview about getting or not getting the expected value of 25 in *One Set* of 50 spins:

> Daisy: 50% would be 25, and I think it'd be rare that we'd get exactly 25 on our first spin. Well, not rare, but unusual. I mean, it's possible, but I think probably your first set of 50, it would be unusual that we'd get exactly 25 blacks. There's no guarantee that... you're going to get exactly 25 out of 50.

For multiple sets, she thought that "to get every single set of spins to be 25 would just be unlikely," and George suggested that results "could be higher, it could be lower...to not get a 25, it's possible that's not happening" Similar reasoning was used when discussing extreme results resulting from *Six Sets*:

> Emma: Sometimes you CAN get as low as a 16, and sometimes you can get as high as 34... It just seemed out of six (sets) that it's unlikely to get 45.
>
> George: Well, 25 would be half, and 20 is possible. It's possible to get a high number. You know, is it possible to get 36? Could it happen? Sure! Sure it could happen. It's very unlikely.
>
> James: I thought it was kind of unlikely that out of (six sets), to have a 10 and a 45...they just seemed too far out. Very unlikely in six (sets), but – Possible.
>
> Ross: (Commenting on a conjectured range) You've got a range here of 19 to 32, so it's hovering around that 25, and there is some variation but it doesn't strike me as extreme, and so… It seems possible, reasonable.

The examples provided by the interviewed subjects reflect the trend shown by most of the rest of the class to talk more in terms of what was likely or unlikely than in terms of what would or would not happen. Having had students become more sensitive to presence of variation, they were less strident in their predictions than they were at the beginning of the course, making them more cautious in their predictions. Many simply pointed to the presence of variation in their explanations, such as when Daisy, James, and Sandy commented about results for *Six Sets*, saying:

Daisy: 'Cause there's gonna be variation, because the spinner CAN land anywhere, but probably on average it'll be close to 25. You have variation from your 50%, a little variation from the 25, but not too much!

James: Well, he's just going to have some variation, even though … We know that the probability is 50%.

Sandy: Again, you wouldn't expect to get the same exact thing, I expect more variation.

Giving the presence of variation as a reason for an appropriate distribution of results, coupled with the use of proportional reasoning, and couching explanations in terms of possibilities and likelihoods were all indicators of improved reasoning about expectations.

One hypothesis about why so many students in the PostSurvey gave expectations in terms of ranges and choices without repeat values, and often even stayed away from including the expected value, is that the class experiences and discussions began to persuade them of the rather extreme unpredictability inherent in small numbers of repeated trials. For instance, in the River Crossing Game, some students talked about how they knew a sum of seven was the most likely outcome for the sum of two dice. Eventually the whole class knew the theoretical probability of a sum of seven for a pair of dice, namely $^1/_6$. However, it was clear that even if we threw the pair of dice six times, we might not see any sum of seven. Similarly, it became rather unremarkable in class to toss a fair coin ten times and *not* get exactly five heads, or spin the 5-Spinner in Cereal Boxes ten times and *not* get exactly two 1s as theory suggested. Through discussion and experimentation, students became more comfortable with ranges than with point estimates, and less comfortable with just sticking with the expected value in making predictions. Certainly many students discussed the influence of the class activities in explaining *why*, such as when Emma justified her prediction for *One Set* by recalling that "from doing the activity in class, I know it won't be exactly 50% but somewhere close." Other examples show the scope of the class comments about class activities:

Dixie: In our class experiments, I found 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. I think that as we practiced in class, the more chances or tries you have the more different answers you can get.

Taha: Because due to the data shown in class, the majority of the data will be in the middle but there will be more variety with more data.

Sergio: I choose this answer judging my prediction on the exercise done in Monday's class because as demonstrated in class, every (trial) is different.

A particular impression was made by the computer simulations (using ProbSim and Fathom) that we did as a class, whereby we had a class discussion even as we continually ran the simulation with more and more trials. In later comments on the likelihood of getting extremes results, subjects clearly recalled the use of the computer:

Emma: After seeing the simulations in class on the computer, it seemed almost impossible.

Sheila: I know this because we saw it on the computer program in class.

Dixie: When we did over 5000 tests via the software program, we STILL didn't get the lower #.
Daisy: When we did the test on the computer it took 5000 (trials).
Loni: I remembered in class the computer simulation took 5000 (trials).
Sandy: I was thinking about the simulation in class and how many trials we had to enter in the computer.
Frida: I based it on the activities we have done in class with computer program as well as hands-on activities.

More than a few sample responses have been shared to emphasize the impressions that the class experience made on the students.

4.3. INTERPRETATIONS OF VARIATION

In addition to reasoning about expectations, subjects also revealed changes in how they thought about their interpretations of variation according to causes, effects, and influences on expectation and variation.

Causes For causes of variation, while there was heavier attention paid by some students in the PreSurvey and PreInterview to the physical nature of performing the flips or spins in the probability context (or drawing the samples in the sampling context), there was relatively less concern with these human causes of variation in the PostSurvey and PostInterview. Prior to the class interventions, many subjects expressed concerns about how samples were drawn, coins were flipped, and spinners were spun. Particularly in the case of spinners, the class as a whole seemed initially skeptical about whether or not spinners could actually be a true random device. Some initial responses from the PreSurvey were about the use of two half-black and half-white spinners, and whether or not the chance of both spinners landing on black was 50%:

Molly: Only if the spinner starts spinning in between both is it a 50-50. I think.
Rosie: A lot I think depends on how you spin.
Sarah: I think it depends somewhat on where the spinner is started from and the spinner is not on the same point in both pictures.
James: Depends on the force used to spin, the resistance of the spinner, the direction of the spin.

The representative responses shared above help to illustrate the concern with how the user operates the spinner, hinting that the user can cause more or less variation depending on the technique used. James helped further explain his concern about spinners in the PreInterview:

James: I want to look at the engineering of the spinner, where do you start the spin, you know, I mean…. Do you start it in white, you know, the velocity, or the force… None of that really matters, I guess…
I: I'm asking…
James: I mean, it CAN matter of course, yeah. Well, of course, it WOULD matter, you know, I mean, you play like a game that has a spinner, and, if you're a kid, you know if you hit it just the right way, and you start it at just the right the spot, there's a chance of it being in one spot are greater than in another spot.

I: So this is very well-oiled spinner…Very, very fair spinner
James: Ok, so this is a GOOD spinner. Yeah. Ok. A fair spinner. And the spinner is flat? A flat plane? It's a fairly spun game?

Rather than being contentious, James was expressing notions about fairness that were shared by others in class. Once the class interventions got underway, it became apparent that a major point of discussion was how children (and themselves) might strive to impede variation by, for example, flipping a coin in a certain way, or hitting the needle of a spinner. Even in sampling candies from a jar, subjects wondered about the plausibility of reaching into the jar in a special way so as to minimize variation. After the interventions, in the PostSurveys and PostInterviews, very little was expressed by the subjects about their concerns over causes of variation. One reason for the lack of commentary may be because the class had seemed to resolve the issue of deliberate causes. That is, they clearly knew a great deal about how children might tamper or try to tamper with random devices, and even in their own activities the subjects sometimes struggled with one another over how to fairly use the devices

The class seemed to come to a consensus that the point of doing a probability experiment really hinged on the assumption of randomness, and that their job was to help and not hinder the natural variation of outcomes. That is, they were not to try and spin a certain way to get a certain result, they were just supposed to spin and let the pointer land where it may. Thus, there may have just developed an acceptance of the myriad forms of physical, deliberate causes of variation. Having expressed their concerns in the PreSurvey and PreInterview, and having discussed these concerns in the class activities, they may have reconciled the issue of physical causes, leaving them more sensitive to the natural random variation inherent in the probability activities.

Effects As a part of the framework discussed earlier, the effects of variation were seen in terms of how students perceived probability situations and how they decided on their predictions. The focus of the *effects* component of the framework is therefore aimed at the effects on how students think, and a noticeable change reflected in class responses was a shift away from an "Anything can happen" and an "I don't know" mentality. In terms of trajectory of thinking, a precursor to how "Anything can happen" seemed to be the idea of how reality was different from theory. For example, in the PreSurveys and PreInterviews some students expressed the "Reality versus Theory" mindset in explaining their reasoning:

Daisy: Because probably outcomes aren't for sure outcomes.
Ross: Reality does not obey the estimates of probability.
Sergio: You are dealing with chance, like gambling. In theory there is probably an answer…But if you do it for real, 100 times, the numbers change but the ratios do not.

In the PreInterview, Ross was able to expand on his thinking, and he described a "probability-dictated reality, as distinct from described likelihoods." When asked for further explanation, he said: "I thought, okay, reality is going to impinge on the strict likelihood by a given thing." Thus, an effect of variation for Ross and others is that reality does not always match with what probability says should happen.

From discussions in class, it became apparent the "Reality versus Theory" mindset was held by many. However, a potentially unhelpful result of the "Reality versus Theory" mindset seemed to be that if theoretical predictions couldn't be counted on in reality, then

"Anything could happen" For example, in considering the prediction of probability outcomes in the PreInterview, some subjects were deliberating about what outcomes to choose:

> Sarah: Just choose randomly – Anything is possible.
> James: Well, they're all likely.
> George: You could just get any number.
> Sandy: Logically that's what my brain is telling me, is it can be absolutely anything.

A major problem with the "Anything can happen" mindset is that subjects who held this view tended to think of *all* outcomes not only as possible, but also as somewhat equiprobable. As Sandy said later on in the PreInterview, "I feel like it really can be anything. And so making a guess is just like… Just saying anything." Sandy's comment gives no regard to the relative likelihoods of different outcomes, and implies a complete lack of guidance in making predictions.

Along with the "Anything can happen" view, a strong undercurrent of the class discussions prior to the interventions swelled toward the idea that it wasn't possible to even make a prediction, which was likened to guessing – the "I don't know" mindset. The following excerpts illustrate what subjects wrote when asked to make predictions on the PreSurvey:

> Alice: You can make a prediction, but not a concrete answer.
> Leila: Always getting (25 heads) is hard to predict.
> Carrie: Hard to say. The odds are never exact.
> Frida: Couldn't hazard a guess, or could but it would be random.
> Rosie: This one I don't know. I have to do it physically.

The key feature that emerged from PreSurvey and PreInterview responses as well as from the class discussions prior to the interventions was that many students were extremely reluctant to make predictions, often using language to the effect that they "couldn't guess." The PreInterviews helped show that what was meant by the "I don't know" mindset was not really that students couldn't guess or predict, but that they couldn't know ahead of time whether or not their predictions would be correct:

> Emma: You just never know what you're going to get.
> James: It's impossible to know. Because you can't predict the future. I mean, I don't know what I'm going to get.
> Sandy: I dunno, I can't guess. I have trouble making guesses because… I can never know.

Sandy in particular encapsulated the view of many in the class, saying in the PreInterview that "you can never really guess. Because there's always a chance that any of those numbers could be anything." The trajectory of thinking held by many students prior to the class interventions was that (1) Probability theory may suggest a given result, but in reality results will vary; (2) Since results can vary, anything is possible, even to the point of being equally likely; (3) Since anything can happen, one can't know ahead of time what will occur, so it isn't possible to know ahead of time what will occur.

Thus, it is the effect of variation upon perceptions ("Anything can happen") that also interferes with the effect of variation upon decisions ("I don't know"). In other words, it

is the variation inherent in the probability situations that results in uncertainty, leading in turn to the difficulty students have with making a prediction. The hypothesis is that variation (and the resultant uncertainty) means one doesn't know for sure what will occur, and if one doesn't know what will occur, then results could be anything, thus confounding expectations. What is really striking is that virtually none of the "Anything can happen" and "I don't know" views were expressed after the class interventions. The discussions from the class, along with PostSurvey and PostInterview responses, suggest that most subjects thought that although one may not know for sure about a given outcome, one can still make reasonable statements of expectation. Also, subjects had less difficulty in making choices and decisions in the PostInterviews, and choices were more reasonable than in the PreInterviews.

Influencing Expectations and Variation Finally, the conceptual framework dimension of influencing expectations and dimensions came through more strongly and credibly after the class interventions, chiefly in the way that subjects referred to the number of sets of spins used in the PostSurvey and PostInterview. The number of sets was related to the average, extremes, and the overall distribution of results from multiple trials (sets of spins), with richer notions being expressed after the class interventions. Prior to the class interventions, for example, Sandy mentioned expecting "an average of 25, if you did many of these sets," and Ross agreed that "if you're going to see a range of results, the average of that range will be 25, but not every result will be 25." The ideas put forth by Sandy and Ross were shared by others in the class who thought that even if results varied, the average for multiple sets would or should be the expected value (or close to that value). After the class interventions, there were more comments that reflected how performing more sets of spins would draw the cumulative average closer to the expected value:

> Sandy: The more that you would do these sets of 50 spins, the more it would probably come back towards that 25.
> Sergio: The more times, the closer it will be toward 25.
> Loni: The more times he spins, the closer he will actually get to the 50/50 chance.
> Sheila: The more he spins the closer the results will match the probability (1/2).
> Maya: It will be even closer to 25 because of the Law of Large Numbers.
> James: So, the theoretical should come close to the experimental… Over the long run, if we do enough trials, chance are, they'll come pretty close if we do a fair number of sets.

The richness of the type of thinking really comes through in James' comment above, as it exemplifies the ideas from the class discussions how experimental probability relates to theoretical. Most importantly, instead of thinking that the average needed to always be the expected value, after the class interventions there were more comments such as George's: "Your mean and median will probably get closer and closer – the more and more you do – you know, the closer and closer you would get to 25." In particular, George pointed out in the PostInterview, with fewer numbers of sets "you're going to have a lot more variation in where the median and the mean are going to go." George's remarks, shared by others in the class, clearly show an appreciation for how even averages can vary.

As for influencing the extremes with performing more sets, there was an appreciation both before and after the interventions of how more sets would expand the range, but the

comments were rather thinly expressed at the outset of the course. Typical notions were that “the range will increase with increasing attempts” (Cammy), and “the more sets you do, the more often you’d expect to get that low chance” (Sandy). After the interventions, again the same kind of idea was expressed about an increasing range happening with more and more sets, but the language was more sophisticated:

Ross: As the number of trials goes up, so expands the range of possible outcomes towards the extremes.
Sandy: You would expect with more sets you do, the more sort of outliers you would get, or the ‘unexpecteds’ you would get.
Emma: The more sets, the more opportunity you have for outliers.
George: The more you do, the better chance of getting those extreme numbers.
Rocky: And so, the more sets you do, the more opportunity that exceptional event has of occurring, the more chance there is of getting an outlier, or an extreme value.

Of course, the language of outliers and extremes was a part of the course discussions, and the class had seen how doing more sets had in fact made it more likely to get an outlier, so it made sense to see these ideas expressed after the interventions.

Regarding distributional thinking, mostly primitive notions came through prior to the interventions, with some students mentioning how results might vary with more sets. However, the few responses were not very specific:

Daisy: The more you do something, the more chances you have that it’s going to vary from the percentage.
Dixie: If you have more friends doing it, then I think there’s more chance of more variation.
Jackie: The more sets done, the more likely you will get less likely results.
Julie: The more people that do the experiment, the more varied the results.

After the interventions, responses were more articulate. For example, Daisy expressed that “the more times you do it, you’ll have variations on each end, which might get wider, but you’ll have more in the center, around the 25.” Julie mentioned that with more sets, “the results would get tighter, the grouping would accumulate around 25.” More importantly, subjects gave reasonable comments aimed at the shape of the underlying distribution:

Daisy: The more pulls you do, the more evenly shaped your graph is going to be. Where fewer pulls, you’re going to have a little more unevenness in your curve.
Ross: The more trials run, the more normal the distribution, but the chance of outliers also increases. I expect to see a certain bell curve, given more trials.

Note how Daisy mentions both the influence of more and of fewer sets, with fewer sets attributed to an “unevenness” in the graph of results. Others in class agreed with this idea, and Sandy expressed that with fewer sets, “you expect there to sort of be this more random look to it, so it’s going to look a little bit more scattered.” With more sets, Sandy thought “it would become more conformed to this perfect bell-curve, and that it would pull out a little bit more” meaning the tails of the graph would extend further. Even

though Ross and Sandy appropriately discuss bell curves in the context of the probability problems, they and others in class also tended to use the language of a "bell curve" or a "symmetric" distribution even on other sampling and probability questions where the underlying distribution was not normal.

5. SUMMARY AND IMPLICATIONS

In this final section, first a summary of the main insights of this study is given, and then implications for future teacher training are discussed

5.1. MAIN INSIGHTS

One of the main insights from this research is how subjects became more attentive to variation throughout the course. Written class responses showed improvements from the PreSurvey to PostSurvey, when evaluated according to rubrics that placed higher value on responses recognizing variation in probability situations. Each of the subquestions of *One Set*, *Compare Sets*, and *Six Sets* showed overall class improvements regarding what subjects expected and why (reasons for expectations). In *One Set*, more subjects gave range expectations and gave evidence of reasoning using both variation and proportions. For *Compare Sets*, more subjects incorporated a range or some kind of spread into their explanation of why results would not necessarily be the same for the results of the second set as on the first set. With *Six Sets*, students gave more appropriate choices that were backed up with reasoning explicitly using proportions and variation. Despite the improvements shown by subjects towards the end of the course, it should be acknowledged that there were still areas in which a substantial percentage of students showed a less than optimal performance. For example, as Table 3 showed earlier, on *One Set* almost half of the students still gave 25 as an answer on the PostSurvey, rather than some other response that might better acknowledge the effect of variation.

Another main insight from this research is the usefulness of the conceptual framework in characterizing the thinking of elementary pre-service teachers. While the coding rubrics were useful for gaining a quantitative picture of overall class changes, the evolving framework was a useful lens for looking more closely at key aspects of reasoning about variation which changed over the course. The interview responses combined with more detailed survey responses helped paint a more detailed picture of the richer understanding that emerged from subjects in terms of their expectations and interpretations of variation. Overall, subjects drew from their collective learning to better reason in terms of what they expected and *why*. Their predictions in the PostSurveys had better attention to range considerations and less emphasis on repeated values for results, particularly involving the expected value. Their reasoning included appropriate balancing of proportional thinking along with an appreciation of variation in expressing what was likely or probable. Class experience clearly had an influence on the reasoning of many students after the interventions, particularly the use of computers. Their interpretations included a reconciliation of physical causes of variation, leading them to focus more on natural causes of variation, namely the randomness inherent in the probability situations. Instead of interpreting variation as simply leading to an "Anything can happen" mindset, accompanied by an "I don't know" regard for making predictions, more students were able to express reasonable predictions. Also, students showed some reasonable interpretations of the effect that performing more trials might have on the cumulative average, presence of outliers, and shape of distribution of results.

5.2. FUTURE IMPLICATIONS

Implications for teaching elementary pre-service teachers include the suggestion that having hands-on activities, bolstered by small-group and whole-class discussion focused specifically on variation, can be a powerful way to move them toward a better appreciation of how variation plays a role in statistical thinking. The class interventions involved all three main aspects of understanding variation (expecting, displaying, and interpreting) in the contexts of sampling, data and graphs, and probability situations, all of which are important for elementary school children to address. If school teachers are to shape their lessons so as to encourage statistical thinking in their own students, then university teacher training programs need to provide an environment where pre-service teachers can learn in a similar way that they themselves will aim to teach (National Council of Teachers of Mathematics, 1991). In the environment where this research took place, the teaching philosophy of the course encouraged a great deal of discourse among students, which served to naturally provide springboards from which class discussions of variation could emerge. Also, a key design component of the surveys, interviews, and class interventions was that subjects were expected to make conjectures and discuss their reasoning before actually doing any activities. By laying out ahead of time what everyone in class thinks, groundwork can be established for making comparisons after actual data have been collected by doing the probability experiments. The computer simulations, brought out only after students have physically run simulations themselves, also seem to hold much promise for getting subjects to understand long term trends. Elementary pre-service teachers, like the children they will one day teach, need to investigate variation in probability settings by conjecturing, reasoning together, doing experiments, and discussing findings. The task for teacher educators is to continue to develop ways to structure their college classes to support elementary pre-service teachers' reasoning about variation.

However, in designing instruction for elementary pre-service teachers, it is important to keep learning more about the initial conceptions of variation that they hold, and how those conceptions change with different instructional interventions. That is, there is an iterative sense in the way instruction for pre-service teachers is designed and then refined based upon what has been learned about how they think about variation. The research described in this paper has been largely exploratory because little has been known about the conceptions of variation held by elementary pre-service teachers. An implication for research is that more needs to be learned about how elementary pre-service teachers' conceptions of variation compare with those of elementary students. For example, what are some similarities and differences in the responses of elementary pre-service teachers and school children? How can elementary pre-service teachers increase their own knowledge of variation while also learning how children reason about variation?

6. CONCLUSION

As research in the field of statistics education advances, one goal is that teacher education can improve not only the subject matter knowledge of elementary pre-service teachers, but also the pedagogical content knowledge of teaching about variation. Steps toward improved pedagogical content knowledge can certainly be informed by recent research about how pre-college students learn. Meanwhile, steps toward improved subject matter knowledge can be informed by a consideration of the conceptions of variation held by pre-service teachers as they enter university programs. Collective discourse in the class, bolstered by activities and simulations targeted at eliciting conceptions of variation

and developing these concepts, hold promise as a way of building elementary pre-service teachers' knowledge while also reflecting the kinds of practice they themselves will want to demonstrate in their own classrooms.

REFERENCES

Canada, D. (2004). *Preservice teachers' understanding of variation.* Unpublished doctoral dissertation, Portland State University, Portland, Oregon (USA). [Online: www.stat.auckland ac.nz/~iase/publications/dissertations/dissertations.php]

Cobb, G., & Moore, D. (1997). Mathematics, statistics, and teaching. *American Mathematics Monthly, 104*(9), 801-824.

Finzer, W. (2001). *Fathom™ Dynamic Statistics* [Computer software, v. 1.1]. Emeryville, CA: KCP Technologies.

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92-99. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_Garfield_BenZvi.pdf]

Hammerman, J., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41. [Online: www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Hammerman_Rubin.pdf]

Jolliffe, F., & Gal, I. (Eds.). (2004). *Statistics Education Research Journal, 3*(2). [Online: www.stat.auckland.ac.nz/~iase/serj/SERJ3(2).pdf]

Jones, G., Mooney, E., Langrall, C., & Thornton, C. (2002). Students' individual and collective statistical thinking. 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.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-451.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59-98.

Konold, C., & Miller, C. (1994). *ProbSim: A Probability Simulation Program.* Santa Barbara, CA: Intellimation Library for the Macintosh.

Makar, K., & Canada, D. (2005). Pre-service teachers' conceptions of variation. In the *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education*. Melbourne, Australia.

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27-54. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_Makar_Confrey.pdf]

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. 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.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal, 3*(2), 84-105. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Reading.pdf]

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201-227). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Shaughnessy, J. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Bidduch & K. Carr (Eds.), *Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 6-22). Rotorua, New Zealand: MERGA.

Shaughnessy, M., & Arcidiacono, M. (1993). *Visual encounters with chance (Unit VIII, Math and the mind's eye)*. Salem, OR: The Math Learning Center.

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., Ciancetta, M. & Canada, D. (2004). Types of student reasoning on sampling tasks. In the *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*. Bergen, Norway.

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

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

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

Truran, J. (1994). Children's intuitive understanding of variance. In J. Garfield (Ed.), *Research Papers from the 4th International Conference on Teaching Statistics (ICOTS 4)*. Minneapolis, MN: International Study Group for Research on Learning Probability and Statistics.

Watson, J., Kelly, B., Callingham, R., & Shaughnessy, J. (2002). The measurement of school students' understanding of statistical variation. *The International Journal of Mathematical Education in Science and Technology, 34*, 1-29.

Watson, J., & Moritz, J. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145-168.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 233-265.

DANIEL L. CANADA
Eastern Washington University
203 Kingston Hall
Cheney, WA 99004
USA