GRADUATE TEACHING ASSISTANTS' STATISTICAL CONTENT KNOWLEDGE OF SAMPLING

JENNIFER NOLL
Portland State University
noll@pdx.edu

ABSTRACT

Research investigating graduate teaching assistants' (TAs') knowledge of fundamental statistics concepts is sparse at best; yet at many universities, TAs play a substantial role in the teaching of undergraduate statistics courses. This paper provides a framework for characterizing TAs' content knowledge in a sampling context and endeavors to raise new questions about TAs' content knowledge and its potential impact on the teaching of undergraduate statistics. The participants in this study were sixty-eight TAs from 18 universities across the United States. These TAs demonstrated considerable knowledge of theoretical probability distributions. However, they experienced tensions when attempting to quantify expected statistical variability in an empirical sampling situation and had difficulty explaining conceptual ideas of variability.

Keywords: *Statistics education research; Sampling distributions; Statistical knowledge for teaching; Teacher knowledge*

1. INTRODUCTION

Over the past two decades, mathematics education researchers, statisticians, and statistics education researchers have devoted greater attention to research in statistics education. In part, this attention stems from concerns over the statistical literacy of the general population, as well as a broad scope of professions requiring more sophisticated statistical skills (Ben-Zvi & Garfield, 2004; National Council on Education and the Disciplines, 2001). Although more recent attention has been allocated to the field of statistics education in general, there is a dearth of empirical research specifically investigating the statistical content knowledge of teachers (Groth, 2007; Shaughnessy, 2007). A review of the education research literature reveals a small number of studies (e.g., Canada, 2004; Heid, Perkinson, Peters, & Fratto, 2005; Liu & Thompson, 2005; Makar & Confrey, 2004) focused on K-12 teachers' statistical content knowledge, and a void in research investigating college and university teachers' statistical content knowledge.

At universities across the United States, enrollment in introductory statistics courses is increasing (Lutzer, Rodi, Kirkman, & Maxwell, 2007), and introductory college statistics is likely to be the first exposure many students have to statistics. Moore (2005) suggests that students form their attitudes and beliefs about the use of statistics from these beginning courses and these courses serve as a potential recruiting ground for future statisticians. Many of these courses are taught by teaching assistants, either teaching their own course or teaching recitation sections. Thus, these introductory courses serve a critical function, and TAs' role in the teaching team is integral (Lutzer et al.; Moore). Although TAs teaching undergraduate statistics courses is not inherently problematic, it is

Statistics Education Research Journal, 10(2), 48–74, http://www.stat.auckland.ac.nz/serj

not uncommon for TAs who majored in mathematics or the sciences as undergraduates to enter graduate school having never taken a statistics course. Green (2010) investigated statistics graduate TAs' experiences and perceptions teaching introductory statistics and noted TAs wanted more direction and support when they first start teaching introductory statistics. Yet many universities do not offer mentoring or professional development to TAs for teaching specific courses, although more are starting to do so (see for example, Froelich, Duckworth & Stephenson, 2005; Gelman, 2005; Harkness & Rosenberger, 2005). To be in a position to teach in a way that allows students to develop strong conceptual and procedural knowledge, teachers need to have a solid foundation themselves. Thus, it is imperative that TAs possess strong content knowledge before teaching introductory college statistics courses because these courses are fundamentally important for developing the attitudes and beliefs of future consumers of statistics. Consequently, improvements in statistics education are likely to remain limited without a careful examination of statistics TAs' content knowledge. In addition, already existing professional development and TA training programs need empirical evidence documenting content areas that TAs need to think more deeply about. This study makes a contribution to statistics education research by taking a first step towards examining TAs' content-specific subject matter knowledge in the context of sampling. This paper addresses the following research questions:

1. What strategies do TAs employ in solving problems in an empirical sampling context?
    a. To what aspects of distribution do TAs attend?
    b. What connections do TAs make between empirical distributions and theoretical distributions? How do they grapple with, and resolve, differences between theoretical models and empirical data?
2. What considerations about variability do TAs express in their thinking about sampling problems?

2. BACKGROUND

This section reviews two areas of the research literature relevant to the current study—teacher knowledge and research on the teaching and learning of sampling concepts.

2.1. TEACHER KNOWLEDGE

At the college level, statisticians and statistics educators (Cobb, 1993; Cobb & Moore, 1997; Zieffler, Garfield, delMas, & Reading, 2008) have called for reform in both the structure and content of introductory statistics courses. Statistics is fundamentally different from mathematics in that it must be taught with context in the forefront and must emphasize the omnipresence of variability—otherwise the subject loses all meaning (Cobb, 1998; Gould, 2004; Pfannkuch & Wild, 2004). But are statistics TAs prepared to teach their statistics courses in ways that meaningfully incorporate applications and stress the omnipresence of variability? This question points to a fundamental factor in teaching introductory college statistics courses—teacher knowledge and experience.

Mathematics educators have long been interested in the issue of teacher knowledge and experience, and over the past 20 years there has been a paradigmatic shift in the way researchers have conceptualized teacher knowledge (see Ball, Lubienski, & Mewborn, 2001; Shulman, 1986). Shulman described how past research on teacher knowledge either focused on teachers' specific content knowledge (e.g., knowledge of particular mathematical topics as measured by course work or grades) or their pedagogical knowledge (e.g., presentation of material, classroom management, etc.). Shulman's construct of pedagogical content knowledge (PCK) provided a link between content and pedagogy. He provided a compelling argument that the expert knowledge of a mathematician is not sufficient for teaching mathematics, and that qualities such as classroom management, which is completely void of subject matter, would be insufficient for a thorough understanding of the knowledge required to teach mathematics. In particular, Shulman defined PCK as content knowledge that "goes beyond knowledge of the subject matter per se to the dimension of subject matter knowledge *for teaching*" (p. 9).

Ball and her colleagues (2001, 2008) built upon Shulman's (1986) work and have been driving research on teacher knowledge through the construct of mathematical knowledge for teaching (MKT) by addressing the question of "what do teachers need to know and be able to do in order to teach effectively" (Ball, Thames, & Phelps, 2008, p. 394). They argue that in order for teachers to be effective, their knowledge base must include subject matter knowledge, knowledge of pedagogy, knowledge of common student misconceptions, knowledge of student development, knowledge of common student solution strategies, and knowledge of curriculum and best practices for introducing material. They dissect MKT into six distinct components—*common content knowledge*, *specialized content knowledge*, *knowledge at the mathematical horizon, knowledge of content and students, knowledge of content and teaching* and *knowlege of the curriculum*. Figure 1 shows the domain map of Ball and her colleagues' framework of MKT.

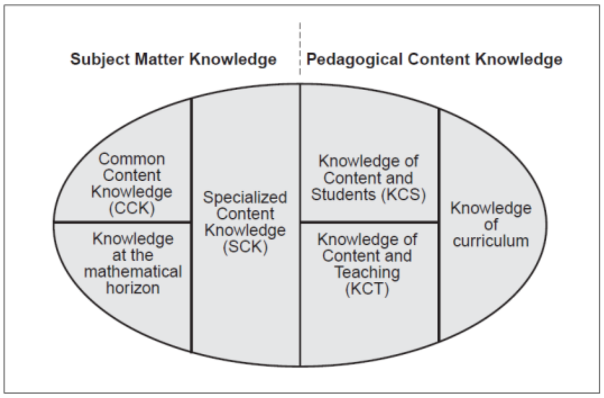

Subject Matter Knowledge
Pedagogical Content Knowledge
Common Content Knowledge (CCK)
Knowledge at the mathematical horizon
Specialized Content Knowledge (SCK)
Knowledge of Content and Students (KCS)
Knowledge of Content and Teaching (KCT)
Knowledge of curriculum

Figure 1: Domain map of mathematical knowledge for teaching (Hill, Ball, & Schilling, 2008)

Each of the six components of MKT is briefly defined below in order to situate the current study for the reader. The first three, *common* and *specialized content knowledge* and *knowledge at the mathematical horizon*, are types of knowledge that constitute what it means to know mathematics. *Common content knowledge* is defined as mathematical knowledge *not unique* to teaching. For example, the ability to subtract multi-digit numbers is necessary mathematical knowledge in many different settings. *Specialized content knowledge* is mathematical knowledge *unique* to teaching (Ball et al., 2008, p. 400). Specialized content knowledge includes "looking for patterns in student errors," determining whether a student's nonstandard approach to a problem works, understanding different representations or ways of seeing a problem, and justifying mathematical ideas

(p. 400). For example, understanding why the algorithm for subtracting multi-digit numbers works is mathematical knowledge teachers need in their work, whereas a person balancing his checkbook only needs to know how to do the computation and not necessarily understand why it works. Finally, *knowledge at the mathematical horizon* is subject matter knowledge that goes beyond the specific content taught to the next level. For example, an algebra teacher should know something about calculus because that represents where the students are headed and is knowledge at the horizon. The other three components of Ball's framework are considered pedagogical content knowledge components and define what it means to teach mathematics. *Knowledge of content and students* includes knowledge of common student misconceptions and common approaches students may take when learning a particular topic. *Knowledge of content and teaching* includes knowledge of how best to sequence instruction, examples that best highlight particular concepts or procedures, and what choices of representations to include during instruction (p. 401). *Knowledge of curriculum* includes global knowledge of the curriculum as well as how the curriculum plays out on a daily basis.

Ball and her colleagues have been studying elementary teachers in practice and finding ways to measure each of the different components of MKT. However, these distinctions are not really as mutually exclusive as their definitions suggest. For example, when a teacher is grading a student solution to a homework problem, it may be difficult to decipher whether she is using *specialized content knowledge*, mathematically examining a non-standard student approach, or using her *knowledge of content and students*, recognizing a common student misconception or stage of development, or a combination. Likewise, there are some mathematical concepts that may fall on the boundary between *common content knowledge* and *specialized content knowledge*. For example, one might argue that some professions need knowledge of why the algorithm for multi-digit subtraction works, a financial consultant for instance, and, thus, this knowledge is not unique to teaching. It is important to point out, though, that common and specialized content knowledge are both forms of subject matter knowledge.

Given that there exist significant distinctions between the disciplines of statistics and mathematics that impact the way the subject matter should be taught, research investigating *statistical knowledge for teaching* (SKT) is an important field of study distinct from research investigating MKT. Yet, because "statistics utilizes mathematics," and there is considerable overlap in the structure of statistics education and mathematics education, there is much that can be gleaned from research on MKT (Groth, 2007, p. 427). Groth suggests applying Ball and her colleagues' six components of MKT for use in statistics education research and he defines SKT as necessary knowledge for statistics teachers in order to be effective in their work. SKT then includes, but is not limited to:

- Knowledge of the concepts and procedures of statistics—statistical literacy and statistical thinking skills;
    - Analysis of statistical solutions to problems and determination of whether the results are reasonable and what types of thinking or analysis might lead to particular results;
    - Knowledge of informal and formal statistical inference ideas including connections between probability, sampling distributions, and statistical inference;
- Ability to formulate questions, collect data, analyze data, and interpret results;
- Ability to recognize and account for the key roles of context and variability in solving statistical problems, including formal procedures for calculating variability and conceptual ideas of variability;
- Ability to recognize the crucial role language plays in expressing statistical ideas;

- Ability to recognize common student misconceptions and understand how students come to learn specific statistics content;
- Ability to manage productive class discussions about statistics and answer student questions.

The list outlined above represents a synthesis of numerous recommendations from statistics education researchers (Franklin & Garfield, 2006; Garfield & Ben-Zvi, 2008; Groth, 2007; Rossman, Chance, & Medina, 2006) regarding the teaching and learning of statistics. For the purposes of this paper, I focus solely on TAs' *subject matter knowledge* of statistics in the context of sampling. Thus, this paper addresses aspects of TAs' SKT that pertain to the first three primary bullet points outlined above. The first three primary bullets are content focused, and thus include both *common* and *specialized* forms of content knowledge. Although the research questions that guide this study did not include teasing apart common and specialized forms of statistical content knowledge, the findings do point to some potential distinctions and raise some questions about these two forms of content knowledge and, thus, are included in the discussion section at the end of the paper. Whereas subject matter knowledge, both common and specialized, are only two of the six components of SKT, they are important for statistics education researchers to study in order to glean insights that will support the evolution of quality professional development for statistics teachers.

2.2. RESEARCH ON LEARNING SAMPLING CONCEPTS

Statistical inference is the central focus for college-level introductory probability and statistics courses. Statistical inference is the process by which conclusions about a particular population are drawn based upon evidence obtained from a *sample* of the population. Statistical inference is an important skill for those living in data-driven societies, and therefore a key topic in introductory statistics courses. Yet, research suggests there are substantial gaps in students' (Chance, delMas, & Garfield, 2004; Pfannkuch, 2005) and K-12 teachers' (Heid et al., 2005; Liu & Thompson, 2005) informal and formal understanding of statistical inference. Part of the difficulty lies in students' ability to make connections between probability models and statistical inference. Students need a strong foundational understanding of distribution, variability, samples, sampling distributions, and populations and parameters in order to construct salient connections between probability and statistical inference. Sampling distributions and their properties play a key role in the theory behind the statistical analysis of single or multiple samples drawn from a population. Thus, sampling distributions are important for understanding how estimates of population parameters are derived and represent an important building block to a coherent understanding of statistical inference.

Despite the fact that sampling distributions are foundational to introductory statistics curricula, the complexity involved in building a coherent understanding of sampling distributions is often underestimated. Saldanha and Thompson (2003) suggest two possible views of a sample—static and dynamic. They note secondary students' proclivities toward a static view of sample, wherein students see a sample simply as a subset of the population from which it was drawn. Saldanha and Thompson suggest that it is a far more complex task to conceive of a sample dynamically, whereby one imagines the sampling process as being repeated and considers the variability inherent to the sampling procedure throughout the process. They argue that the second, more dynamic, concept image (in the manner of Tall & Vinner, 1981) of samples is necessary to support a coherent understanding of sampling distributions and informal notions of statistical inference.

In addition to a concept image of a sample that entails a dynamic image of repeating the sampling process, other researchers (e.g., Bakker & Gravemeijer, 2004; Pfannkuch & Reading, 2006; Rubin, Bruce, & Tenney, 1991; Shaughnessy, 2007; Wild, 2006) suggest that *distributional reasoning* skills are necessary in order to rationally negotiate more complex sampling tasks, and that distribution as a conceptual structure constitutes a unifying theme in the introductory statistics curriculum. *Distributional reasoning* is a complex task which requires (1) coordinating two or more key features of a distribution (e.g., center, shape, outliers, and spread/variability—including density and skewness); (2) coordinating ideas about sampling and randomness; and, (3) making connections between empirical and theoretical distributions. Research suggests that students often have trouble coordinating multiple attributes of a distribution and tend to rely on a single, elementary attribute to reason through more complex sampling situations.

K-12 students' difficulties in reasoning about distributions of data in sampling contexts appear to reside in their inability to appropriately apply measures of variability (Konold & Pollatsek, 2002; Reading & Shaughnessy, 2004; Rubin et al., 1991). Wild (2006) argues that "statisticians look at variation through a lens which is 'distribution'" (p. 11), yet the insight with which statisticians account for variability and the ways in which they conceive of distribution are not easily acquired. Students do not appear to have prudent intuitions for how to quantify the expected spread in a sample and in a sampling distribution. A finding of major consequence in the research literature is that students' conceptions of samples and sampling distributions fall within a spectrum, with measures of center at one end and measures of variability at the other (Konold, 1989; Reading & Shaughnessy; Rubin et al.). As students traverse this continuum, they must grapple with the role that sample size and sample selection methods play in the variability of sampling processes. Unfortunately, much of the research indicates that many students reason at the extremes of this spectrum and not in the middle. On the one hand, students tend to express the belief that there is more variability in a sampling distribution than is probable and, as a result, they appear to be overly focused on obtaining an unusual sample or sample statistic. On the other hand, students tend to express the belief that samples and their statistics are identical to the parent population; in these instances, students appear to believe the sample provides all the information one needs about a population because they fail to think about issues of variability (Rubin et al.). It is not necessarily surprising that students experience difficulties identifying and resolving tension between measures of center and variability. Indeed, statistics is, in a sense, the study of variability. Variability is a complex idea and, depending on the context, statisticians must decide to minimize, maximize, estimate, model, analyze and/or tease apart variability in data (Gould, 2004; Wild). It seems plausible, then, that statistics TAs (that is, apprentice statisticians) may have a tenuous, novice, and developing relationship with variability.

Shaughnessy and colleagues (2004a, 2004b, 2005) have identified three features in the development of students' statistical reasoning in the context of sampling—*additive*, *proportional*, and *distributional*. To illustrate these characterizations, consider the following situation: A well mixed jar contains 100 candies, 60 red and 40 yellow; pull out a handful of ten candies, note the number of reds, put the candies back in the jar, mix them back up, and repeat this process 49 more times. Now predict the number of handfuls out of the 50 containing 0 red candies, 1 red candy, 2 red candies, … , 10 red candies. Shaughnessy et al. (2004a) used this task with middle and secondary school students (n = 272). They observed that *additive* reasoners attend to absolute frequencies (e.g., "there are more red in the jar so I expect more red than yellow in each handful"). *Proportional* reasoners primarily use the underlying ratios of the population as they reason in sampling situations (e.g., "each handful should contain about 6 red candies because the ratio of red to yellow candies in the jar is 60 to 40"). *Distributional* reasoners coordinate two or more attributes as they reason about sampling situations. For instance, a distributional reasoner might justify his/her predictions by discussing the ratio of red to yellow candies and the fact that one should expect the outcomes from handful to handful to differ slightly from that expected value. They also observed students in transitional stages. In these instances, students tended to focus on single attributes, such as mode, shape, and spread. For instance, if a student primarily attended to the shape of a distribution or the range of a distribution, then he or she was coded in the transitional category. Taken together the characterizations of additive, proportional, and distributional reasoning, along with transitional reasoning, these characterizations provide an initial conceptual framework, used in the present study, for how students come to form coherent statistical thinking skills when investigating distributions of data (see Figure 2).

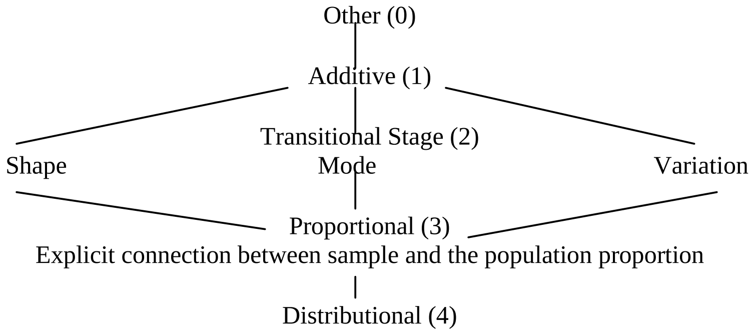


Figure 2. Conceptual framework of Shaughnessy and colleagues (2004a, 2004b)

The studies discussed in this section have been focused on K-12 students', tertiary students', or K-12 teachers' statistical reasoning and development in sampling contexts. Evidence from the research literature suggests that students need focused instructional support in order to develop a dynamic conceptualization of sampling distributions, coordinate multiple attributes of a distribution, and reasonably quantify their expectations for variability in sampling situations. There appears, however, to be no research investigating how TAs reason in sampling contexts, what distribution means to them, what difficulties they may encounter when working with and analyzing empirical sampling distributions, and how they describe variation within and between empirical sampling distributions. What is needed now is research that helps us better understand how TAs, with considerable mathematics and statistics backgrounds, reason about distribution and variability in sampling distributions, and the practical implications of creating professional development for TAs that facilitates their statistical thinking. The study reported in this article makes a small contribution to this end by expanding on and refining current frameworks in order to model TAs' reasoning. In addition, this study raises questions about the need for the statistics education research community to tease apart forms of common and specialized content knowledge.

3. METHODS

3.1. DATA COLLECTION METHODS AND STUDY PARTICIPANTS

TAs' reasoning was explored through a task-based web survey and through semi-structured interviews conducted with a subset of the survey population. The data corpus consisted of 68 survey respondents from 18 universities across the United States, and five interview participants taken from the larger survey population. Contacts were made at universities targeted through the Research in Undergraduate Mathematics Education (RUME) list-serve, through the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) website, and from the list of graduate programs in statistics found at the American Statistical Association (ASA) website. Both the survey and interview participants comprised a volunteer sample. Survey participants were selected based on three criteria: (1) willingness to participate in the study; (2) experience teaching (or assisting with) at least one introductory statistics course; and, (3) the prior completion of *at least one* graduate statistics course (in fact, over 50% of the survey participants had taken eight or more graduate statistics courses). Interview participants were selected from the larger group of survey participants based on three additional criteria: (1) TAs' willingness to participate in three hour-long interviews, (2) TAs' location, due to budgetary constraints, and (3) the researcher's selection of varied survey responses so as to represent as much of the gamut of responses as possible. The five interview participants were all from the same institution—a large research university in the Pacific Northwest. The TAs at this institution participate in a week-long orientation focused on general teaching strategies such as creating a syllabus, grading rubrics, consistency, and professionalism. The orientation does not cover specific mathematical or statistical topics because TAs may be teaching different courses. There is no mandatory mentoring program for TAs prior to teaching specific courses. Although some TAs are informally mentored by faculty prior to teaching a particular class, this was not the case with any TA in this study. The demographic information for the survey and interview participants is shown in Tables 1 and 2, respectively.

Table 1. Demographic information on the survey participants (n = 68)

ESL*	Gender		Undergraduate degree		Current field of study	
Yes 14(21%)	Male	36 (53%)	Statistics	12 (17.6%)	Statistics	54 (79.4%)
No 54 (79%)	Female	32 (47%)	Mathematics	39 (57.4%)	Mathematics	7 (10.3%)
			Math Educ.	1 (1.5%)	Math Educ.	3 (4.4%)
			Other	16 (23.5%)	Other	4 (5.9%)

*ESL denotes English as a second language

Table 2. Demographic information on the five interview participants

Pseudonyms	Amanda	Andy	Joe	Sandy	Sam
Program of study	Ph.D. Statistics	Ph.D. Math	Ph.D. Math Ed	Master's Statistics	Master's Statistics
Courses taken					
Undergraduate stat courses	1	4	0	4	2
Graduate stat courses	10+	9	1	10+	10+
Teaching experience					
Introductory Statistics I	Once	Once	Once	Multiple	Once
Introductory Statistics II	Multiple	Never	Never	Multiple	Never
Statistics for engineers	Once	Never	Never	Never	Never

The survey contained four tasks and took approximately 30 minutes to complete. Three 60-minute individual interviews were conducted with the five interview participants, for a total of 15 interviews. The interviews allowed the researcher to achieve the level of detail necessary for gaining a more robust understanding of these TAs' knowledge of sampling distributions. The first interview consisted of a series of follow-up questions to the survey tasks, and the second and third interviews contained four new sampling tasks, questions about definitions of different statistical terms, and questions about teaching statistics.

This paper reports on the survey responses to two of the sampling tasks that helped to provide a general framework that characterizes trends in the types of reasoning exhibited by these TAs. In addition, this paper includes excerpts from the first interview with Amanda, Sandy, and Andy to illustrate the conceptual framework. The interview excerpts provide deeper insight into these TAs' reasoning strategies. Amanda's interview, in particular, describes difficulties she experienced as she negotiated these two tasks. Although the information gleaned from the interviews is limited in scope, as it only applies to these specific TAs, the findings help corroborate some of the observations from the survey data, as well as pinpoint promising features of TAs' knowledge worthy of future research and larger scale studies.

Of the five TAs interviewed, Amanda, Sandy, and Andy were ideal candidates for informing the framework because (1) they were exceptional at thinking aloud and could clearly articulate their thinking; (2) they had each taken a significant number of graduate statistics courses; (3) they had each taught at least one college-level introductory Statistics course; and, (4) each had a strong interest in teaching. Amanda is highlighted here because she appeared to struggle more with these two tasks and her ability to express her thinking aloud highlights her struggles. Joe and Sam were removed from consideration because: (1) both had difficulty thinking aloud, making it difficult to analyze their reasoning strategies and draw conclusions about their thinking; (2) Sam displayed significant difficulties with the tasks, which may account for his resistance in sharing his thinking aloud; (3) Sam was not a native English speaker (although neither was Sandy) and his English communication skills were poor; and, (3) Joe was different from the majority of TAs taking the survey and from the interviewees in that he had only taken one graduate level statistics course and his graduate studies were not focused on statistics—thus, he was less representative of the larger survey population.

3.2. TASK DESIGN

The two sampling tasks are referred to as the Prediction Task and the Real/Fake Task (see Figures 3 and 4, respectively). These tasks were originally used with middle and high school students (see Shaughnessy et al., 2004a, 2004b, 2005). Similar tasks have appeared previously in the literature (see Reading & Shaughnessy, 2004; Rubin et al., 1991). These tasks were chosen because they emphasize core statistical concepts such as reasoning about empirical data, variability, and distributions of data, and they align with recommendations made in the Guidelines for Assessment and Instruction in Statistics Education (Franklin et al., 2007). In addition, the tasks align with the framework suggested by Zieffler et al. (2008) for tasks that support the development of informal inferential reasoning. In particular, these tasks require informal statistical inference skills, inductive reasoning, and the ability to reach reasonable conclusions under uncertainty. Rossman et al. (2006) argue that "[s]tatisticians often come to different but reasonable conclusions when analyzing the same data" and that the " quality of conclusions lies in the analysts' ability to support and defend their arguments" (p. 329). It is equally important for teachers of statistics to be able to reason and communicate their statistical thinking in a sophisticated manner in order to assist their students' development of robust statistical thinking skills. Thus, the tasks used in this study provided the opportunity to assess how TAs reason and communicate about fundamental statistics concepts. In addition, because these tasks present opportunities for reasoning about statistical ideas both formally, through probability models, and informally, they provide opportunities to investigate whether or not TAs have multiple strategies and ways of reasoning about these tasks, as well as an opportunity to discuss which types of statistical knowledge are unique to teaching and which are not. The tasks are limited in the sense that the theoretical probability is known, and thus they do not provide an opportunity to study TAs' strategies when only empirical data are present.

3.3. THE PREDICTION TASK

The Prediction Task (see Figure 3) can be a straightforward application of a binomial or hypergeometric probability model, and thus, it was anticipated the majority of TAs would apply one of these two models. The nature of the task suggests that it measures common content knowledge. The task was designed to address the following questions:

- Would TAs reason more formally (apply a particular probability model) or informally (generally discuss the distribution's center, shape, and/or spread) when justifying their predictions? (Corresponds to research questions 1 and 2)
- To which aspects of the distribution would TAs attend? (Corresponds to research questions 1 and 2)
- Would TAs' reasoning strategy be consistent with their prediction? (Corresponds with research question 1b)

A jar contains 1000 candies, 750 are red and 250 are yellow. The candies are mixed well. Suppose that you pull a random sample of 10 candies from the jar, record the number of reds, put the candies back in the jar and mix them up. Suppose you do this 50 times. How many times out of 50 do you think you would get a handful of 10 candies with:

Number of Red Candies in Handfuls of 10	Prediction
0 red	
1 red	
2 red	
3 red	
4 red	
5 red	
6 red	
7 red	
8 red	
9 red	
10 red	
Total	50

Figure 3. Prediction Task

Prior to analyzing TA predictions, the author and another statistics educator created a rubric, quantifying what would constitute a *reasonable* prediction based on criteria previously established by Shaughnessy and colleagues (2004a, 2004b, 2005) (Table 3).

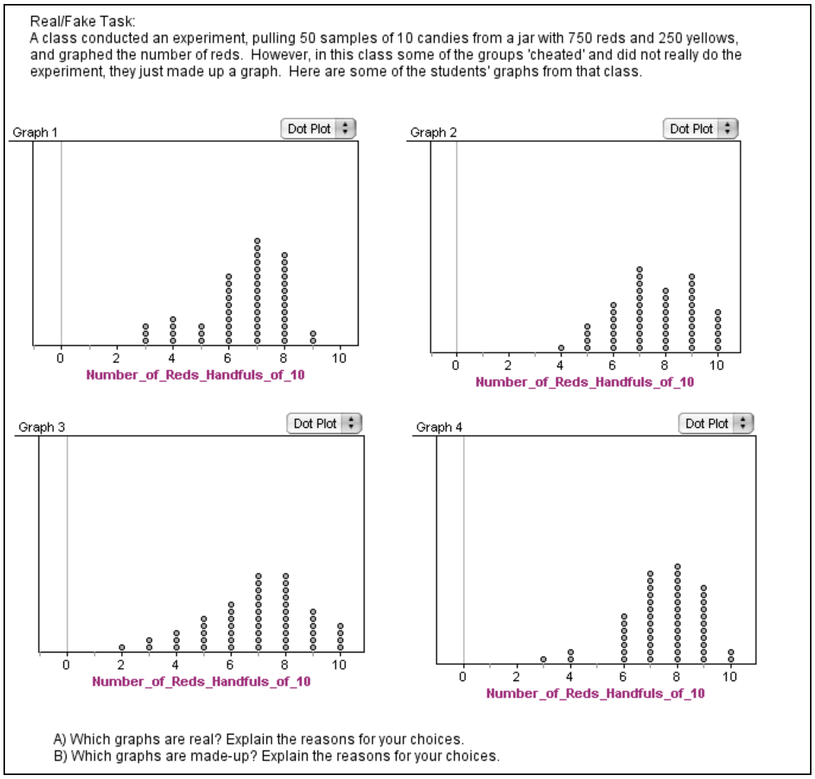
Table 3. Criteria for assessing reasonable predictions for the empirical sampling distribution

Criterion 1: *Whole Number Predictions*	The predictions should be whole numbers, as decimals do not make sense in the context of the problem.
Criterion 2: *Appropriate Center*	The center should be located between 7 and 8 red candies because the population proportion is 75% red. Somewhere between 40% (20 handfuls) and 68% (34 handfuls) of the outcomes should be placed at 7 or 8 red candies. If there are fewer than 40% or more than 68%, the sampling distribution has an unreasonably low or high number of outcomes at 7 or 8 red candies, respectively.
Criterion 3: *Appropriate Spread*	The distribution should be concentrated in the 6 to 9 red candies range with a few handfuls containing 3, 4, 5 and/or 10 red candies. An interval length of 3 or fewer units constitutes an *unreasonably narrow* range for the sampling distribution. For example, it is highly unlikely that all 50 outcomes would be stacked at 6, 7, and 8 red candies. An interval length of 9 or more units constitutes an *unreasonably wide* range for the sampling distribution. For example, it is highly unlikely that the 50 outcomes would be spread from 0 or 1 red candy through 10 red candies.
Criterion 4: *Appropriate Shape*	The distributions should be approximately mound-shaped around the population center. The distribution should not look too uniform or have a large 'spike' at 7 and 8 red candies. Drops in frequency of more than 9 units from the 7 and 8 to the 6 and/or 9 constitutes an unusually large change in frequency and a change of 2 or fewer from the 7 and 8 to the 6 and 9 constitutes an unusually small change in frequency. Likewise, too much density at the low end of the distribution would be unlikely. For example, placing 3 or more outcomes at 2 or 3 red candies would be unreasonable, or placing 0 outcomes at 10 red candies and only 1 or 2 outcomes at 9 red candies would be unreasonable.

3.4. THE REAL/FAKE TASK

The Real/Fake Task (see Figure 4) is an extension of the Prediction Task. The experimental situation is the same, but the Real/Fake Task evaluates TAs' abilities to detect fraud by determining a reasonable expectation for the shape, center, and spread of the sampling distributions. Graphs 1 and 3 were manufactured ('fake') and Graphs 2 and 4 were generated via computer simulation ('real') (see Shaughnessy and colleagues, 2004a, 2004b). Graph 1 was designed with an appropriate range, but shifted to the left. Thus, Graph 1 has an unusually high number of outcomes at four and below and too few at nine and ten. Graph 3 was designed to have a 'smooth' distribution in terms of frequencies, as well as a range that is unusually wide.

Based on previously established criteria (see Shaughnessy et al., 2004a, 2004b), Table 4 illustrates one method for assessing the empirical sampling distributions in the Real/Fake Task. The criteria in the table focus on variability and the tails of the distribution. For example, the table shows that the probabilities for a simulation producing graphs with criteria similar to Graphs 1 and 3 are less likely than for Graph 2. If we consider the population proportion it makes sense that we should expect more outcomes with 9 and 10 red candies than outcomes with 2, 3, and 4 red candies.



Real/Fake Task:
A class conducted an experiment, pulling 50 samples of 10 candies from a jar with 750 reds and 250 yellows, and graphed the number of reds. However, in this class some of the groups 'cheated' and did not really do the experiment, they just made up a graph. Here are some of the students' graphs from that class.

A) Which graphs are real? Explain the reasons for your choices.
B) Which graphs are made-up? Explain the reasons for your choices.

Graph 1
Dot Plot
0
2
4
6
8
10
Number_of_Reds_Handfuls_of_10
Graph 2
Dot Plot
0
2
4
6
8
10
Number_of_Reds_Handfuls_of_10
Graph 3
Dot Plot
0
2
4
6
8
10
Number_of_Reds_Handfuls_of_10
Graph 4
Dot Plot
0
2
4
6
8
10
Number_of_Reds_Handfuls_of_10

Figure 4. Real/Fake Task

Table 4. Criteria for characterizing unlikely classes of graphs: P(red)= 0.75

Types of distributional outcomes	Quantifying types of outcomes in the tails	Probability
Many outcomes at the low end of the distribution	6 or more of the 50 samples of size 10 result in 4 or fewer red candies (e.g., Graphs 1 and 3)	0.0004
Many outcomes at the high end of the distribution	17 or more of the 50 samples of size 10 result in 9 or more red candies (e.g., Graph 2)	0.0782
Few outcomes at the high end of the distribution	2 or fewer of the 50 samples of size 10 result in 9 or more red candies (e.g., Graph 1)	0.00013

Although we were interested in whether or not TAs would correctly identify the computer simulated graphs versus the fraudulent ‘knockoff’ graphs, the task served a greater purpose: investigating the ways in which TAs utilize statistical thinking when reasoning about empirical sampling distributions. In some sense this task is more pedagogical than the Prediction Task in that it is a task designed for teaching about variability in distributions of data and can be used with beginning statistics students. The task is certainly statistical in nature, but may better assess the kinds of statistical knowledge a teacher of statistics may need rather than other professionals who use statistics. In particular, this task was designed to answer the following questions:

- What types of sampling distributions, if any, would TAs conceive of as unusual for this situation? How narrow or wide could the distribution be before a TA considered it unusual? (Corresponds with research questions 1 and 2)

- How much variability would TAs expect in the percentage of occurrences at the center or how much variability do TAs expect in a left-skew shape? (Corresponds with research question 2)
- Would TAs expect the shape of the graph to be smooth or would they expect ‘bumps,’ ‘dips’ and ‘ups and downs’? (Corresponds with research question 2)
- What are TAs’ expectations for the statistical variation of the data set? (Corresponds with research question 2)

Table 5 provides a brief description of the framework for characterizing TAs’ reasoning on the Prediction and Real/Fake Tasks. Much of the framework shown in Table 5 was built upon the work of prior statistics education researchers (e.g., Reading &

Table 5. Statistical reasoning framework applied to TA justifications

Category		Description
Idiosyncratic (I)		No reasoning supplied, reasoning was unclear, or not pertinent.
Additive (A)		Primary reasons give attention to frequencies, “more red,” primarily additive or frequency only type reasoning.
Single Attribute Reasoning	Center (C)	Reasoning states or strongly implies the use of centers (e.g., modes, %, probability) or mention of population proportion.
	Shape (S)	Overall attention appears to be based on the shape of the distribution. Shape language includes skewness, normally distributed, too perfect, too formulaic, evenly distributed, smooth, not smooth, bumpy, goes up and down, gaps, piled up. Shape language also relates to the distribution’s density—relative frequencies, modes. If a survey respondent uses the term variability, but appears to be referring to the graph’s frequencies this would be coded as shape.
	Spread (V)	Variation responses reference ideas of statistical variation such as range, standard deviation, interquartile range, and spread. A separate code of Tails was used when TAs focused on one or both of the ends of the distribution.
	Tails (T)	A separate code of T if the reasoning refers to *tails*, normally the number of 3s, 4s, and/or 10s. If TAs paid attention to both ends of the distribution, their response was just coded T**.** However, often they expected more or fewer outliers at either the high or low end of the distribution, and in those cases the secondary codes were added: MH—expect *more high* (10s) LH—expect *fewer high* (10s) ML—expect *more low* (3s, 4s) LL—expect *fewer low* (3s, 4s)
Informal Distributional (ID)		*Explicit* distributional reasoning involves the coordination of at least two of the attributes of a distribution (S, C, V, T) on the same graph. For example, if a TA says “mostly 7 and 8s, but will decrease from there” that would be evidence of ID because the TA mentioned the center and is acknowledging some spread around that center.
Formal Distributional (FD)		Mentioning the binomial or hypergeometric probability distribution and/or written use of these formulas. Also, mention of combinations of different outcomes would be evidence of FD. If a survey participant mentions computing the probability of a certain event then that would also be coded as FD and the type of event should be noted. For example if a survey participant computes the probability (or says they have computed the probability) of getting four handfuls containing four red candies that would receive a code of FD-tails.

Shaughnessy, 2004; Shaughnessy et al., 2004a, 2004b, 2005). However, the original framework (recall Figure 2) was further developed and refined during the data analysis of this study. The author and another statistics education researcher reviewed the survey data, using the previous framework (recall Figure 2), looking for new categories of thinking as well as discrepant events. Once the first author developed a detailed framework, the second coder independently reviewed the survey responses in order to test the coding scheme. Inter-rater reliability scores are given with each task in the sections that follow. The result of that analysis is the framework presented in Table 5.

Most survey and interview participants identified the underlying probability distribution when the population parameters were known, and subsequently a formal probability distribution category naturally developed and the additive category essentially disappeared. Thus, the original framework of Shaughnessy and colleagues (2004a, 2004b) (Figure 2) was further refined for use with TAs by distinguishing both a *formal probability distribution* (*FD*, when explicit reference was made to probability models) and *informal probability distribution* (*ID*, the distributional reasoning classification as described in the literature review) category. FD was applied when TAs made explicit reference to the underlying probability distribution. The next section illustrates the framework via specific survey and interview responses.

4. RESULTS

What is particularly compelling is that certain trends in reasoning among K-12 students and teachers also appeared among TAs, and thus, the initial framework was both relevant and useful for categorizing TAs' reasoning about introductory statistics concepts. What follows is an analysis of TAs' reasoning for each task applying the author's framework from Table 5 (as well as the other scoring criteria established in Tables 3 and 4). In addition, the TAs' predominant forms of reasoning are discussed.

4.1. PREDICTION TASK ANALYSIS

During coding of the Prediction Task, TAs received a reasoning code (Table 5) and a score of 0-4 for their predictions (Table 3), with one point provided for each criterion that was met. For two coders, the inter-rater reliability was 95.6%. Table 6 provides examples of TA predictions. For instance, survey respondent #68 identified the Prediction Task as a hypergeometric situation (last column of the table), and was coded as FD; yet, his predictions are inconsistent in that they sum to more than 50 trials, and the range he provides extends from two handfuls containing 1 red candy to 17 handfuls containing 10 red candies. His prediction score is one out of four as he only satisfied Criterion 1. Survey participants #23, #65, and #68 had unreasonably wide predictions because the number of outcomes they predicted at four and fewer red candies in a handful is highly unlikely. In fact, when pulling 50 samples of size 10 the probability that one would draw a handful where none or one of the candies is red is extremely low (0.0000416 and 0.00129 respectively).

As expected, most (approximately 68%) of the TAs provided an FD justification and their predictions appeared to be based on their calculations of the binomial or hypergeometric probability distribution function (see Table 7). In general this suggests that these TAs had strong common content knowledge on this task. Yet, several of these FD and ID reasoners provided overly wide ranges in their predictions.

Table 6. Examples of survey and interview participants' responses to the Prediction Task

Survey participant	Prediction score(Table 3)	Number of Red Candies Predicted in 50 samples of size 10											Justification and reasoning code (Table 5)
		0	1	2	3	4	5	6	7	8	9	10	
#54	4	0	0	0	0	1	3	8	11	15	9	3	"This is a binomial distribution prob with *n* = 10 each trial" – FD
#68	1 (Fails Criteria 2, 3, 4)	0	2	4	5	7	9	10	12	14	15	17	"Hypergeometric distribution" – FD
#52 Sandy	4	0	0	0	0	1	3	8	13	14	9	3	"Find prob to get x red at one draw; multiply by 50 to simulate the number of times out of 50 we get x red." – FD
#06 Andy	3 (Fails Criterion 1)	0	0	0	0.1	0.8	2.9	7.3	12	14	9.4	2.8	"Hypergeomtric. Compute expected values: 50*Prob(x=#of red candies)." – FD
#08 Amanda	4	0	0	1	1	1	3	7	12	13	9	3	"It's highly unlikely that we will get 0 or 1 red, so out of the 50 tries I do not expect any draws to turn out that way. However, for each draw I expect 7 or 8 reds, so I think most of the draws will have that results." – ID
#65	1 (Fails Criteria 2, 3, 4)	1	1	2	3	4	5	6	8	9	6	5	"Have to be honest, I assumed that for the most part you would pick 7 or 8 (75%) of the red ones, and just kind of distributed it evenly about 7 and 8." – ID
#23	2 (Fails Criteria 3, 4)	0	1	2	3	4	5	6	15	15	5	4	"7 or 8 red candies, since the probability that we get a red is 0.75." – C

*Table 7. Survey results for types of reasoning in the Prediction Task (n = 68)**

FD	ID	Center	Other (includes the additive and idiosyncratic responses)
46(67.6%)	10(14.7%)	5(7%)	7(10.3%)

*Inter-rater reliability 95.6%

Approximately 17.6% (12 of the 68 participants) of the survey responses failed to meet Criterion 3, providing predictions that were unusually wide. Of the 12 predictions that were coded as unusually wide, 8 were made by participants coded as FD or ID, suggesting that these TAs' predictions were inconsistent with their reasoning.

Survey respondents #06, #08, and #52 represent Andy, Amanda, and Sandy, respectively. During the first interview, participants were asked to elaborate on their thinking, describe how they interpreted the Prediction Task, and to judge the reasonableness of other predictions. Andy and Sandy were coded as FD reasoners and during the interview continued to judge predictions based on computing probabilities. Amanda was coded as an ID reasoner on the survey, but during the interview she indicated that she did some calculations with the binomial distribution function in order to help guide her predictions. However, Amanda appeared to struggle when she reasoned informally about the left tail of the distribution. During the interview, she mentioned that she imagined the sampling situation by picturing a big jar filled with red and yellow candies all mixed up, and it was just as difficult for her to imagine pulling out a handful containing all 10 red candies as it was to imagine a handful containing no red candies. When queried further about this comment Amanda retracted it; yet, the following discussion of how she visualized the sampling situation ensued:

Interviewer: Do you think it is more likely that you would get a handful with 3 red candies or with 10 red candies?

Amanda: I have a harder time imagining 10 red candies.

Interviewer: Okay. So you're thinking it's more likely I'm going to get 3 red candies?

Amanda: Just in a visualizing sense, yeah.

Interviewer: Okay, so when you say that, in a visualizing sense, that might not have anything to do with how the actual theoretical probabilities work out?

Amanda: Right, exactly. … and actually, when you phrase it in terms of do you feel like it would be *more likely* then instantly my gut reaction is to say, well I can't say that because I would have to sit down and calculate probabilities. But just in my mind's eye, visualizing way.

In this excerpt, Amanda made a distinction between her statistical knowledge with regard to the sampling situation versus her concept image (in the manner of Tall & Vinner, 1981) and intuition of the situation. This excerpt provides some evidence that Amanda was experiencing difficulty balancing her expectations for sample to sample variability versus sample representativeness—a finding consistent with studies involving K-12 students (e.g., Reading & Shaughnessy, 2004; Rubin et al., 1991). Amanda did not appear to coordinate the population proportion with her expectations for variability in this context. Several of the TAs from the survey also provided overly wide ranges for their predictions despite using FD or ID arguments, which may be an indication that Amanda is not the only TA who may struggle with finding balance between sample representativeness and sampling variability. Amanda's struggle to balance her expectations of representativeness and variability, as well as several overly wide ranges in the predictions of TAs taking the survey, lend some small support for the author's hypothesis that TAs (novice teachers of statistics) may have similar developing conceptions as their students.

4.2. REAL/FAKE TASK ANALYSIS

The Real/Fake Task provided an additional opportunity to investigate TAs' expectations for variability in empirical sampling situations. Responses on the Real/Fake Task were given two codes, a score of 0-4 for the number of correct identifications, and a reasoning code (recall Table 5). Table 8 provides examples of responses coded as FD, ID, S, and/or T reasoning along with the number of correct identifications. Some TAs focused their reasoning on *different attributes* from graph to graph (e.g., shape on one graph and tails on another graph). These TAs were given multiple reasoning codes depending on the

Table 8. Examples of survey responses to the Real/Fake Task

Participant	Reasoning code and TA responses	# Correct
#48	Code: FD & ID (Tails LL and Shape) Response: "Graph 1—There are *too many low samples*. The probability of getting 3 reds in a sample of 10 is 0.0031. Just getting one of these samples is very unlikely, so it is even more unlikely that they would get three of them. Graph 2—This is a very reasonable plot given the expected values. Graph 3—Again there are *too many low samples*. The probability of getting 2 reds out of 10 is 0.0004. It is highly unlikely that the students would get one sample with 2 reds and two samples with 3 reds. The overall distribution is also a little *too perfect*. Graph 4—Although this graph is a little too heavily concentrated in the 7-9 range, it does appear to be a reasonable graph. Even though I have speculated that two of these graphs are made-up, they are all theoretically possible."	4
#24	Code: ID (Tails MH and LL; Shape; Center) Response: "1: *Too few 10s and possibly too many 3s*. 2: Seems to have the right shape center and spread. 3: *Too smooth* for just 50 simulations (like my own prediction earlier—made up). 4: Seems to have the right shape, center, and spread, and the gap at X=5 is something only a very clever student would include."	4
#45	Code: S & T (Shape & Tails; LH and ML) Response: "1 seems roughly in line with expected values. The tails of the normal distribution are about where they should be. 2 weights the high values too heavily—*there are a large number of 9s and 10s, and only one 4 with nothing below*. 3 seems like a *smooth distribution*, but almost *too smooth*. 4 has large disparities between the number of 6s and 5s, and seemingly too many 9s."	0
#08 Amanda	Code: S & T (Shape and Tails; LH) Response: "2 seems to have *too many observations for 9 and 10*, and it makes me feel uncomfortable with the graph. 3 just looks *too perfect*."	2
#52 Sandy	Code: S (Shape) Response: "Graph 3 seems to be *too close to the expected* results. It doesn't seem to account for *empirical variation* like the other three graphs do."	3
#06 Andy	Code: S (Shape) Response: "The last two graphs seem *too perfect*. Real data samples have lots of *variance* to them. The first one seems the most correct. Graph 2 is biased somehow. Graph 3 and 4 are *idealized*."	2

attributes to which they attended. Some TAs *coordinated* two or more attributes of the distribution (e.g., shape and tails) on the same graph. TAs who specifically *coordinated* two or more attributes of the distribution on the *same* graph were coded as ID (or FD if they applied a probability model). For example, survey participant #48 was coded FD and ID because she coordinated the attributes of Tails and Shape on Graph 3—indicating that there were too many low samples *and* that the shape was too perfect. In addition, she computed the probability of getting two reds out of ten and three reds out of ten to justify why there should not be as many low samples. Survey participant #45 also reasoned about the shape and tails of the graphs, but did not coordinate those attributes on the same graph. For instance, on Graph 3 he only reasoned about shape and on Graph 2 he only reasoned about the right tail—expecting fewer nines and tens. The last three rows in Table 8 show the responses of Amanda, Sandy, and Andy. Italics were added to highlight typical language survey respondents used to justify their conclusions. Inter-rater reliability for two coders was 92.6%.

Recall that on the Prediction Task, 82% of TAs used FD or ID arguments to justify their predictions (see Table 7). Yet, only 22% (*n* = 68) of TAs used a FD or ID argument to justify their real/fake identifications (see Table 9).

Table 9. Results of Real/Fake Task (n = 68):
*Correct identifications broken down by reasoning codes**

Number Correct	Center	Tails	Shape	Shape & Tails	ID & FD	Idio-syncratic	Total	%
0	0	0	1	1	0	1	3	4.4
1	1	0	2	0	1	0	4	5.9
2	1	1	10	4	3	4	23	33.8
3	0	0	15	1	2	0	18	26.0
4	0	1	1	8	9	1	20	29.4
Total	2	2	29	14	15	6	68	
%	2.9	2.9	42.6	20.6	22.0	8.8	100.0	

*Inter-rater reliability 92.6%

Of the fifteen responses coded as ID or FD, six coordinated Shape and Center, and the remaining nine coordinated the Tails and the Shape of the graphs or the Tails and Center of the graph. Interestingly, the six who attended to Shape and Center made one, two, or three correct identifications, where as those who attended to Tails in normative ways made four correct identifications. Responses that were coded FD tended to compute the probabilities of the Tails of the different graphs and as a result they often made four correct identifications.

TAs appeared to attend to either or both the Shape (typically with Graph 3) and the Tails (typically with Graphs 1, 2, and 4) more than any other attributes of the distribution. The discussion that follows unpacks typical Shape and Tail responses and provides a characterization for how TAs used these attributes to make their real/fake decisions.

4.3. ATTENTION TO SHAPE

One might consider the 15 TAs who obtained three “correct” real/fake identifications using the single attribute of shape (see Table 9) as a signal of its power for assisting in this particular task. However, it is worth considering that these TAs appeared to eliminate Graph 3 on account of it being “too perfect” to be a real graph and then subsequently concluded that the other three graphs were all possible. Three of the four shape responses provided in Table 8 account for instances where TAs primarily focused on the "perfect" shape of Graph 3 as their reason for identifying it as fake and the other three graphs as real, automatically giving them three correct. Successful navigation of this task requires the ability to coordinate multiple attributes of a distribution and identify reasonable bounds for the spread of the distribution by quantifying the expected variability.

Most survey participants tended to use language of "too perfect," "too smooth," "too expected," "not random enough," without "bumps" or "ups and downs," or "too little variability" to describe Graph 3 and justify why it was likely to be a made up graph. TAs' primary justification for accepting a graph as real was based on finding graphs with "imperfections" in their shapes. TAs appeared to measure imperfections in a graph's shape by looking for unevenness in the frequency of a graph. For example, Graph 1 was often identified as "real" because more handfuls contain four red candies than five red candies, noting that; "the graph goes up at four and drops back down at five," or "there is a dramatic drop in frequency from eight red candies to nine red candies."

TAs' descriptions of variability in relation to Graph 3 are, in some sense, confusing because the word "variability" might refer to the spread and density of the graph or it might refer to the "bumps" and "ups and downs" in the graph. TAs who indicated that they expected variability and randomness in real graphs and concluded Graph 3 was made up because of its lack of this kind of "variability" were coded as Shape, because their descriptions did not appear to attend to statistical variability. The interview data provided additional evidence that these TAs were indeed reasoning about the Shapes of the graphs when they referred to "variability in the graphs"; consider the next excerpt, for example.

Interviewer: When you say Graph 3 is too perfect, what do you mean by too perfect?

Amanda: *I expect in a real graph that they're not going to be in order 0, 1, 2, 3, 4 etcetera. Some are going to be higher than the one next to them and some are going to be lower.*

Interviewer: So the even steps up in frequency?

Amanda: *Yes, it goes up very smoothly. It doesn't have any dips in terms of 0* [red candies] *up to the 7, 8* [red candies] *and then from the 7, 8* [red candies] *it's decreasing back down.* … Compare that to Graph 1 where we increase [*in frequency*] from 3 to 4 [*red candies*], but then we decrease from 4 to 5 [*red candies*]. So it increases too smoothly and too evenly. I just don't buy this. … It's just increasing and then just decreasing. Which if I could repeat the experiment to infinity, I might expect to happen.

The italicized utterances suggest that the steady, smooth increase, followed by the steady, smooth decrease is the type of shape Amanda expected to see in theoretical models. Also, worth noting is Amanda's comparison of the smoothness of Graph 3 to the unevenness in Graph 1. Amanda appeared to identify Graph 1 as real, in large part because of the dips in frequency, something she seemed to expect in empirical data.

Interviewer: So is this [Graph 3] also your image of what the ideal graph would look like, sort of in line with your own predictions?

Amanda: Yes.

Interviewer: And experimentally this is …

Amanda: Not going to happen [laughs].

Interviewer: [Laughs]. So Graphs 1 and 4 were real because they have these things that do occur experimentally?

Amanda: *Yeah, all these little quirks*. We have in Graph 1, more piled on 4 [red candies] than we do on 5 [red candies]. [In] Graph 4 we don't have anything at 5 [red candies], these are things that occur in an actual testing situation.

Amanda's focus on, and subsequent comparison of, smooth versus uneven graphs suggests her main criterion for identifying the real versus the fake graphs is based on the "quirky" shapes that she expects to see in empirical situations.

Sandy and Andy also indicated that Graph 3 was "too perfect" to be real, and Graph 3, at least in terms of shape, seemed to match their image of the theoretical distribution.

Interviewer: What do you mean by too perfect [pointing to Graph 3]?
Sandy: Because, *too perfect is like what you expect it to have this shape*. … Because it has, look, [makes a sketch of a left skewed distribution over the Graph 3] *this kind of shape, you see—close to 0 and then mounds up and then decreases.*

Andy: *It's going to have like defects, all sorts of holes, funny anomalies* [makes a gesture—draws a distribution curve in the air with lots of vertical ups and downs]. … Like maybe a case out here [pointing to the 1, 2, and 3 red candies range on graph 1] or maybe like this one [pointing to Graph 1 at the 3, 4, and 5 red candies spots], like this divot [where Graph 1 dips down at 5]. *I mean you could almost draw a nice curve over this* [pointing to Graph 3].

In both of these excerpts, Sandy and Andy made similar comments as those made by Amanda about the qualities of the 'ideal' graph. It seems that Amanda, Sandy, and Andy have strong expectations for the variability in the shapes of empirical distributions of data. The detail of reasoning observed in these interview excerpts supports the conjecture that the survey responses from TAs expecting "more randomness" or "variability" in the distribution of Graph 3 were likely referring to the changes in heights rather than the spread of the distribution. Garfield, delMas, and Chance (2007) observed that undergraduate statistics students expressed a similar propensity for thinking about real data as having variability in vertical frequencies rather than thinking about statistical variation.

4.4. ATTENTION TO THE TAILS OF THE DISTRIBUTION

Analysis of the data suggests that TAs were also drawn to the tails of the distribution in making their real/fake decisions. Some TAs focused on the low end (2, 3, and 4 red candies), the high end (9 and 10 red candies) or both. TAs expected more lows, fewer lows, more highs, or fewer highs. There may be an implicit attention to the population parameter when the expectation for the tails of the distribution is more highs and fewer lows; because the mixture is 75% red, one should expect more highs and fewer lows. Table 8 also illustrates examples of TA responses focused on the tails of the distribution. For example, survey respondent #45 appeared to expect *more* handfuls with 2, 3, and 4 red candies and *fewer* handfuls with 9 or 10 red candies. Given that this TA's expectations for the tails of the distribution were opposite of what one should expect, it is not surprising that he made 0 correct identifications. Perhaps this TA is also not coordinating the center and spread when reasoning about the distribution.

Amanda also gravitated to expecting more lows and fewer highs. In the next excerpt, she discusses why she believes Graph 2 is fraudulent.

Amanda: I'm expecting to see something down here [referring to the 2, 3, and 4 red candies spots on Graph 2]. Especially taken in conjunction with the fact that I have, how many are piled here on 9? A lot. Twelve, oh, 11 in the 9 slots on Graph 2, and 6 in the 10 slot. And I feel like this is a little disproportionate. I've got nothing here [in 2s, 3s and 4s] and a lot going on at 9 and 10. And I would feel more comfortable. Watch this. This is just going to be awful. *If I removed some off 9 and 10 and moved them over here* [to the 2, 3 and 4 red candy slots] *so that it looked more like Graph 3. The theoretical one* [laughs] *that I think is implausible … . I'm having a battle in my head about theoretically what I expect to happen, which would look like Graph 3, and reasonably in practice what I have seen happen*. … I've spent many, many hours drawing samples on a computer and seeing what they look like. … and they're always a little quirky.

Amanda clearly recognized that she was grappling with how to reconcile her theoretical expectations with her knowledge for variability in a sampling environment. The previous excerpt suggests that her primary criterion for variability continues to be rooted in the shape of the graph. In addition, she appeared to believe that there should be more outcomes at 2, 3, and 4 red candies (i.e., she expects more lows and fewer highs). This belief provides some evidence that Amanda expected greater statistical variability than is actually likely in this scenario. Despite the fact that the population proportion is 75% red, and thus, it is more likely one would draw handfuls with 9 or 10 reds than with 2, 3, or 4 reds, she appeared to think otherwise. Her expectation may go back to her "mind's eye visualization" that she expressed during the Prediction task and which she acknowledged may not be how the "actual" probabilities work out.

The survey and interview evidence suggest that many of these TAs either ignored statistical variation altogether or experienced difficulty estimating the likely distance a collection of sample statistics would fall from the population parameter. In many instances, TAs expected a wider range of outcomes than is likely to occur and expected more outcomes in the left tail of the distribution, perhaps an indication that they did not coordinate the population proportion with their expectations of variability.

5. DISCUSSION

Table 10 summarizes the key findings from the survey data and case studies. The remainder of the paper synthesizes the important findings, discusses distinctions between common and specialized content knowledge in light of the findings and task design, discusses limitations of the research, and concludes with directions for future research.

The survey and interview tasks provided a context to better understand how TAs use their statistical knowledge and apply their understanding of theoretical models to make sense of empirical data. With only a few exceptions, the survey and interview participants demonstrated considerable knowledge of the underlying probability structure in the candy jar context and appeared capable of attending to multiple aspects of the distribution. This is not surprising given their statistical background. However, the survey responses to the Real/Fake Task and the follow up interview questions to this task indicated that several of them experienced difficulty as they attempted to make inferences using the empirical

Table 10. Summary of key findings

Task	Typical Reasoning Approaches	Key Difficulties
Prediction Task	Applying hypergeometric or binomial probability distribution	Making predictions that are inconsistent with their reasoning—in particular wide predictions
Real/Fake Task	Shape • Emphasis that empirical distributions should be bumpy, not smooth Tails • More or fewer high • More or fewer low	Evaluating and quantifying expected spread • Expect greater spread than is likely to occur or more in the low end of the distribution than is likely to occur • Focus on Shape only and ignore issues of spread when evaluating graphs

sampling distributions. The source of this difficulty appeared rooted in their expectations of variability in empirical sampling situations. Most of the TAs in this study did not coordinate multiple attributes of the distributions when reasoning about the Real/Fake Task. Their survey responses suggest that most often they focused on Shape. In particular, the focus was on variability in the frequencies of the different outcomes for each sampling distribution, rather than on statistical variation. The three case studies provided additional evidence that the language “random,” “variation,” and “idealized” were in reference to variation in the heights of columns in the graphs of the sampling distributions and not in terms of the range or spread of the data. Garfield et al. (2007) also noted undergraduate students’ attention to variation in heights of graphs at the expense of attention toward statistical variation. Each of the interview case studies suggested Graph 3, in the Real/Fake Task, was the ideal graph based solely on its shape. The fact that Graph 3 has an unusually high number of outcomes placed at 2, 3, and 4 red candies did not appear to bother the respondents.

In some instances, TAs’ survey responses did attend to statistical measures of variability. For instance, many responses focused on the tails of the distribution. Yet, many of these TAs expected *more* at the *low* end of the distribution and *fewer* at the *high* end. For example, Amanda believed there should be more outcomes with 2, 3, or 4 red candies than with 9 or 10 red candies. Also, in some instances TAs provided wider predictions than is actually likely to occur in the experiment.

This study provides new insights into the ways in which TAs reason about foundational topics in the introductory statistics curriculum. There is evidence from both the survey and interview data that in novel, empirical situations, TAs may reason in similar ways to college students, high school students, and/or high school mathematics teachers. Despite their strong statistical knowledge of formal probability distributions, many of the TA survey participants, as well as the interviewees, did not appear to apply their statistical knowledge when making predictions using empirical sampling distributions. The research presented in this paper suggests that these TAs may have compartmentalized their theoretical knowledge of statistics and have difficulty applying that knowledge when working with empirical data. It could also be that these TAs simply do not have multiple approaches for thinking about these problems and/or cannot easily articulate the concepts behind the theory with which they are familiar.

This study also raises questions about SKT and distinctions between *common* and *specialized* forms of content knowledge. Is the knowledge that these TAs seemed to lack a form of *specialized content knowledge*, *unique* to teaching introductory statistics courses, or *common content knowledge* needed in other professions that require statistics? The fact that many of the TAs expressed strong knowledge of the formal probability model in the Prediction task may indicate that they have common content knowledge and that their difficulty in discussing conceptual ideas of distribution in terms of center, shape, and variability in the Prediction and Real/Fake, as well as their difficulty deciding on criteria for assessing the graphs in the Real/Fake task may be an indication of weak specialized content knowledge (i.e., *statistical knowledge unique to teaching* introductory statistics). On the one hand, the ability to explain concepts, to explain why a particular procedure works, or to have different ways of representing problems is in part how Ball et al. (2008) defined specialized content knowledge. In addition, one could argue that the Real/Fake Task is set up as a pedagogical task with particular introductory statistics concepts in mind. On the other hand, one could easily place the Real/Fake Task in a different context where a statistician is programming a computer simulation looking for glitches in the program. She may look at outputs of sampling distributions for unusual results in order to determine whether the program is functioning properly. Further, statistics is very much an applied field and requires working with uncertainty and making inferences set in a particular context with a given set of data. In both the Prediction and Real/Fake Tasks TAs had access to the underlying theoretical model, but often this is not the case and only empirical data are available. How would TAs reason when only the empirical data are available? This is certainly a situation that statisticians need to deal with on a daily basis. Thus, one could also argue some of the types of reasoning with which these TAs struggled are in fact part of *common content knowledge* not necessarily unique to teaching introductory statistics. It is clear that both these knowledge forms are necessary for teaching introductory statistics in line with current curricular reforms, but it may also be that developing such habits of mind as described above would also serve TAs well as practicing statisticians.

It is not all together surprising that TAs, novice teachers of statistics, and apprentice statisticians would experience difficulty resolving differences between theoretical models and empirical data and struggle with their expectations for variability in sampling problems. It is likely that TAs need more experiences with empirical data, as well as mentoring opportunities with expert statistics educators, in order to support their evolution toward constructing better intuition, more sophisticated reasoning strategies, and the integration of multiple attributes of a distribution into their thinking about empirical data.

The primary limitation of this study is that the participants comprise a small convenience sample. Thus, the results presented here do not necessarily generalize to the larger population of statistics TAs. The survey sample contained a small number of international students, which may not be representative of the larger population. In addition, the data collected in the surveys are limited in scope because in most cases the written responses were brief and it is difficult to make definitive conclusions on TA reasoning on the basis of those written responses. The interviews provide more detail on what some TAs were thinking when they discussed issues of shape and variability in the graphs of empirical sampling distributions. The sample size for the interviews was small and all the interviewees were TAs from the same institution, where there is no mentoring or training for teaching introductory statistics. The results may have been different if interviewees came from an institution where statistics TAs do receive training in teaching introductory statistics. However, the details in the interviews provide some likely roadmaps for the thinking of the TAs in the survey.

Despite the obvious limitations of the study, this research does provide a first glimpse into how TAs reason in empirical sampling situations. The conceptual framework provides information on TAs' cognitive development in the context of sampling and has potential implications for graduate education in statistics. The author's conceptual framework serves as a useful tool for thinking about TAs' reasoning in a sampling context and continued work in this area may bring about further refinements to the framework.

The fact that many of these TAs experienced difficulty accessing and applying their knowledge of distributions in an empirical context suggests that TAs may be limited in their ability to teach their students how to connect core statistical concepts to empirical contexts. Graduate statistics courses, focused on theory, are not designed to provide graduate students with the needed experiences in working with empirical data sets to prepare them adequately for teaching sampling concepts. Large-scale empirical studies are needed in order to further understand and model TAs' statistical reasoning about sampling and sampling distributions. Based on the results of this study, I recommend research that investigates the impact of a mentoring program on TAs' statistical content knowledge. In particular, a mentoring program designed to provide TAs with experience (1) working with empirical data sets, (2) solving novel statistical tasks, (3) responding to hypothetical student solution strategies, and (4) thinking about reform statistics curricula may prove useful in allowing TAs to rethink and refine their understanding of introductory statistics material. There are several universities that have such mentoring programs in place (see for example, Froelich et al., 2005; Gelman, 2005; Harkness & Rosenberger, 2005), but empirical research needs to assess the impact of mentoring and training programs. In addition, the statistics education research community may want to consider design research similar to the work of Ball and her colleagues (2008) in order to tease apart common and specialized forms of statistical content knowledge. That is, if we want to better understand what types of statistical knowledge components are unique for teaching introductory statistics courses, so that we can design research-based professional development and mentoring programs, we need to conduct research that would allow us to identify particular statistical knowledge needed for teaching introductory statistics. The directions for future research highlighted here will likely become important considerations as institutions of higher education become more accountable for improving student learning.

ACKNOWLEDGEMENTS

The author is extremely grateful to J. Michael Shaughnessy and Karen Marrongelle for their helpful suggestions and comments.

REFERENCES

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer Academic.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). New York: Macmillan.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it specials? *Journal of Teacher Education, 59*(5), 389–407.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of* *developing statistical literacy, reasoning and thinking* (pp. 3–15). Dordrecht, The Netherlands: Kluwer.

Canada, D. (2004). *Pre-service elementary teachers' conceptions of variability*. Unpublished doctoral dissertation, Portland State University, Portland, OR. [Online: www.stat.auckland.ac.nz/~iase/publications/dissertations/04.Canada.Dissertation.pdf]

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295–324). Dordrecht, The Netherlands: Kluwer Academic.

Cobb, G. W. (1993). Reconsidering statistics education: A National Science Foundation conference. *Journal of Statistics Education, 1*(1). [Online: http://www.amstat.org/publications/jse/v1n1/cobb.html]

Cobb, G. W. (1998). The objective-format question in statistics: Dead horse, old bath water, or overlooked baby? *Paper presented at the annual meeting of the American Educational Research Association*, San Diego, CA.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801-823.

Franklin, C., & Garfield, J. (2006). The GAISE Project: Developing statistics education guidelines for pre K-12 and college courses. In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth annual yearbook of the National Council of Teachers of Mathematics* (pp. 345-375). Reston, VA: National Council of Teachers of Mathematics.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A preK-12 curriculum framework*. Alexandria, VA: American Statistical Association. [Online: http://www.amstat.org/education/gaise/GAISEPreK-12_Full.pdf]

Froelich, A. G., Duckworth, W. M., & Stephenson, W. R. (2005). Training statistics teachers at Iowa State University. *The American Statistician 55*(1), 8-10.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science and Business Media B.V.

Garfield, J. B., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett and P. Shah (Eds.), *Thinking with Data (Proceedings of the 33rd Carnegie Symposium on Cognition)* (pp. 117–147). New York: Erlbaum.

Gelman, A. (2005). A course on teaching statistics at the university level. *The American Statistician 55*(1), 4–7.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427–437.

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal*, *3*(2), 7–16. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3%282%29_Gould.pdf]

Green, J. L. (2010). Teaching highs and lows: Exploring university teaching assistants' experiences, *Statistics Education Research Journal*, *9*(2), 108–122. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ9%282%29_Green.pdf]

Harkness, W. L., & Rosenberger, J. L. (2005). Training graduate students at Penn State University in teaching statistics. *The American Statistician 55*(1), 11–13.

Heid, K., Perkinson, D., Peters, S., & Fratto, C. (2005). Making and managing distinctions–The case of sampling distributions. In G. M. Lloyd, M. Wilson, J. L. M.

Wilkins, & S. L. Behm (Eds.), *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* [CD]. Blacksburg, VA: Virginia Polytechnic Institute and State.

Hill, H. C., Ball, D. L., Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

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

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259–289.

Liu, Y., & Thompson, P. (2005). Understandings of margin of error. In G. M. Lloyd, M. Wilson, J. L. M. Wilkins, & S. L. Behm (Eds.), *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* [CD]. Blacksburg, VA: Virginia Polytechnic Institute and State.

Lutzer, D. J., Rodi, S. B., Kirkman, E. E., & Maxwell, J. W. (2007). *Statistical abstract of undergraduate programs in mathematical sciences in the United States: Fall 2005 CBMS Survey*. Providence, RI: American Mathematical Society.
[Online: http://www.ams.org/profession/data/cbms-survey/cbms2005]

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353–374). Dordrecht, The Netherlands: Kluwer Academic.

Moore, D. S. (2005). Preparing graduate students to teach statistics: Introduction. *The American Statistician, 59*(1), 1–3.
[Online: http://www.stat.purdue.edu/~dsmoore/articles/TeachPrep.pdf]

National Council on Education and the Disciplines. (2001). *Mathematics and democracy: The case for quantitative literacy*. The Woodrow Wilson National Fellowship Foundation.

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (267–294). The Netherlands: Kluwer Academic.

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal, 5*(2), 4–9.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ5%282%29_GuestEd.pdf]

Pfannkuch, M., & Wild, C. J. (2004). Towards an understanding of statistical thinking. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–46). Dordrecht, The Netherlands: Kluwer Academic.

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

Rossman, A., Chance, B., & Medina, E. (2006). Some key comparisons between statistics and mathematics, and why teachers should care. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth annual yearbook of the National Council of Teachers of Mathematics* (pp. 323-333). Reston, VA: NCTM.

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.

Saldanha, L., & Thompson, P. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257–270.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Charlotte NC: National Council of Teachers of Mathematics.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004a). Types of student reasoning on sampling tasks. In M. Johnsen HØines & A. Berit Fuglestad (Eds.). *Proceedings of the 28th meeting of the International Group for Psychology and Mathematics Education* (Vol. 4, pp. 177–184). Bergen, Norway: Bergen University College Press.

Shaughnessy, J. M., Ciancetta, M., Best, K., & Canada, D. (2004b, April). *Students' attention to variability when comparing distributions*. Paper presented at the Research Pre-session of the 82nd annual meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.

Shaughnessy, J. M., Ciancetta, M., Best, K., & Noll, J. (2005, April). *Secondary and middle school students' attention to variability when comparing data sets*. Paper presented at the Research Pre-session of the 83rd annual meeting of the National Council of Teachers of Mathematics, Anaheim, CA.

Shulman, L. S. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd edition). New York: Macmillan.

Tall D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics 12*(2), 151–169.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10–26.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ5%282%29_Wild.pdf]

Zieffler, A., Garfield, J., delMas, R., & Reading. C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40–58.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ7%282%29_Zieffler.pdf]

JENNIFER NOLL
Portland State University
Fariborz Maseeh Department of Mathematics and Statistics
PO Box 751
Portland, OR 97207
USA