SUBJECT MATTER KNOWLEDGE FOR TEACHING STATISTICAL ASSOCIATION

STEPHANIE A. CASEY
Deerfield High School
scasey@dist113.org

ABSTRACT

This study seeks to describe the subject matter knowledge needed for teaching statistical association at the secondary level. Taking a practice-based qualitative approach, three experienced teachers were observed as they taught statistical association and interviewed immediately following each observation. Records of practice were assembled to create a compilation document to recreate each of the fifty observed class sessions along with related materials including textbook pages and student work. Analysis of the compilation documents focused on the demands upon teachers' subject matter knowledge involved in the practice of teaching. Findings regarding the knowledge required for teaching correlation coefficient are highlighted, including its computation, interpretation, sensitivity, estimation, and related terminology.

Keywords: *Statistics education research; Qualitative research; Teacher knowledge*

1. INTRODUCTION

The nature of the knowledge needed for teaching is largely under-specified and little researched (Ball & Bass, 2003; Ball, Lubienski, & Mewborn, 2001; National Research Council, 2001; RAND Mathematics Study Panel, 2003). Although it is obvious that mathematics teachers need to know mathematics, what is unknown is exactly what aspects of mathematics teachers need to know, how they need to know it, and how and where this mathematics knowledge is used in practice (Ball & Bass, 2003). The need for research is particularly acute for the teaching of statistics. As a result of the National Council of Teachers of Mathematics' (NCTM) *Principles and Standards for School Mathematics* (2000) and its growing importance in today's world (Ben-Zvi & Garfield, 2004), statistics has become an accepted strand of mainstream school mathematics curricula in many countries (e.g., England: Qualifications and Curriculum Authority, 2007; New Zealand: Ministry of Education, 1992; United States: NCTM, 2000). However, many teachers have not studied statistics, and those who have typically experienced a course that emphasized procedural knowledge (Franklin, 2000) rather than the statistical reasoning they will be asked to teach as called for by NCTM (2009) and statistics educators (Ben-Zvi & Garfield, 2004). This has led to concern regarding the knowledge involved in teaching statistics and whether teachers have this knowledge (Conference Board of the Mathematical Sciences [CBMS], 2001; Franklin, 2000; Kettenring, Lindsay, & Siegmund, 2003). This is particularly important for the field as greater knowledge of mathematics by teachers has been linked to higher achievement of students (Hill, Rowan, & Ball, 2005).

The knowledge base regarding statistical knowledge for teaching is thin. Groth (2007) drafted a hypothetical descriptive framework of the knowledge for teaching statistics to encourage related empirical research, but he considered it simply a starting point for the field to begin considering the topic. Two studies (Burgess, 2007; Sorto, 2004) researched the knowledge elementary and middle school teachers need for teaching particular statistical topics, but no research has been done regarding secondary (years 9–12) teachers' knowledge for teaching statistics. Based on his review of current research on statistics learning and reasoning, Shaughnessy (2007) identified the knowledge necessary for teaching statistics as an important area for future research.

The topic of statistical association was identified as one of eight big ideas in statistics (Garfield & Ben-Zvi, 2004) and is a main component in the secondary school curriculum in multiple countries (e.g., England: Qualifications and Curriculum Authority, 2007; New Zealand: Ministry of Education, 1992; USA: NCTM, 2000). It is a foundational concept underpinning the development of statistical reasoning for students (NCTM, 2009) and therefore was the topic I focused upon in my research.

Despite this importance, contemporary policy documents and professional guidelines are vague about teacher knowledge for statistical association. The statements presented in Table 1 illustrate this lack of specification.

Table 1. Statements regarding teacher knowledge for teaching statistical association at the secondary level

Statements	Source
Teachers need experience in using a variety of standard techniques for organizing and displaying data in order to detect patterns and departures from patterns (p. 44)	Conference Board of the Mathematical Sciences (2001). *The Mathematical Education of Teachers.*
[Teacher candidates] design investigations, collect data, and use a variety of ways to display data and interpret data representations that may include bivariate data (p. 6)	National Council for Accreditation of Teacher Education/NCTM (2003). *Program Standards.*
[Accomplished teachers] collect, organize, represent, and reason about data, using a variety of numerical, graphical, and algebraic concepts and procedures, and they look for ways to describe and model patterns in data (p. 29)	National Board for Professional Teaching Standards (2001). *Adolescence and Young Adulthood Mathematics Standards* (2nd ed.).
Teachers at all levels must understand their subject matter [referring to statistics], and at a depth at least somewhat greater than that of the content they actually teach (p. ix)	David Moore (2004). *Foreword.*
Statistics for teaching includes an extremely good statistical knowledge base, knowledge of connections between statistical concepts, and knowledge of applications of statistics.	Patricia Wilson (2004). *GAISE report discussion.*

These broad stroke descriptions leave many questions unanswered, including the following: What knowledge components do teachers need to have in order to deem their knowledge of statistical association extremely good? What depth of understanding do teachers need to have about statistical association? How is this knowledge used in the work of teaching statistics? Descriptions of teacher knowledge need specification, detail, and connections to teaching practice in order to be useful and meaningful for teachers and teacher educators. Thus, I conducted research that used a practice-based approach by studying secondary teachers on the job as they taught about statistical association to define and describe in fine-grained detail the subject matter knowledge needed for teaching statistical association at the secondary level. This type of approach was most appropriate and thorough for understanding the subject matter entailments of teaching and to analyze where subject matter knowledge is used in that work (Ball & Bass, 2000a, 2003; Ball et al., 2001). In this article I describe the entire practice-based methodology used in my research (Casey, 2008), but given the complexity of my findings about the knowledge for teaching statistical association, I have narrowed my findings to focus on one aspect of the topic-the correlation coefficient.

1.1. THEORETICAL PERSPECTIVE

This study used the theoretical construct of teacher knowledge as developed by the Learning Mathematics for Teaching (LMT) project (Hill, Schilling, & Ball, 2004), which has recently been adapted and used in countries around the world (Ball et al., 2009). Ball and Hill (2005) defined mathematical knowledge for teaching as the "mathematical knowledge, skill, [and] habits of mind that are entailed by the work of teaching" (p. 9). For my study the meaning of mathematics is expanded to include the triad of mathematical knowledge, statistical knowledge, and context knowledge, the subject matter knowledge components considered necessary for statistical literacy by Gal's (2004) statistical literacy model, to address the differences between mathematics and statistics as subjects (delMas, 2004).

Hill et al. (2004) divide teacher knowledge into two main parts: pedagogical content knowledge and subject matter knowledge. In this study, I limited my attention to teachers' subject matter knowledge due to calls for its focus in mathematics education research (e.g., Ball et al., 2001; National Research Council, 2001; RAND Mathematics Study Panel, 2003) and the lack of research in this area in statistics education. As one component of subject matter knowledge, *common mathematical knowledge* refers to the knowledge of mathematics content one would expect the average person to possess in order to say that he or she knows that content (Ball & Hill, 2005). It is expected that teachers would possess this content knowledge. However, this knowledge alone is insufficient for teachers. Teachers also need another component of subject matter knowledge: a *specialized mathematical knowledge*. This is a unique knowledge of mathematics needed for teaching that is different from the type of knowledge needed in other occupations where mathematics is used. The depth and detail of specialized mathematical knowledge goes well beyond the knowledge needed by persons who only need to carry out mathematical procedures reliably (Ball and Bass, 2003). It also needs to be unpacked, connected both within and across mathematical domains and across time, and held flexibly so that teachers are prepared for the real-time problem solving that teaching requires. Activities exclusive to teaching, such as determining mathematical problems that are productive for student learning, understanding and judging claims made by students, and creating explanations of mathematical concepts that are accurate and useful for students, are occasions when teachers need to utilize their specialized

knowledge. The description of teacher knowledge created as a result of this study includes both common and specialized knowledge.

1.2 REVIEW OF RELATED LITERATURE

Initial steps have been made to define the statistical knowledge for teaching. Groth (2007) proposed a general framework of statistical knowledge for teaching as a model to be tested and adapted following empirical research. Acknowledging the important differences between mathematics and statistics, he proposed that the knowledge for teaching statistics needs its own specialized theory. In his hypothesized framework, statistical knowledge for teaching includes mathematical and nonmathematical knowledge, as well as common and specialized knowledge. A specific framework to describe the components of teacher knowledge regarding statistical thinking and investigating needed by elementary teachers to teach statistics through investigations was developed by Burgess (2007). The framework was presented as a matrix. The columns of the matrix referred to types of knowledge needed by teachers as identified by Ball and Hill (2005). The elements of statistical thinking and empirical inquiry described by Wild and Pfannkuch (1999) defined the rows of the framework. Empirical evidence of the need for all the types of knowledge identified in Burgess' proposed framework was found with the exception of dispositions and the need for data. Both of these frameworks based their structure on the representation of teacher knowledge developed by the LMT project. This supports my choice of this theoretical perspective for my research. However, neither framework has empirical research at the secondary level as its basis. Ball et al. (2001) assert that subject matter knowledge is better manifested in the actual practice of teaching and studied with a practice-based approach, which was the approach taken by the current study.

A key cognitive activity for humans is correlational reasoning (McKenzie & Mikkelsen, 2007). The correlation coefficient, a statistic used to measure the statistical association between two quantitative variables, is a prominent topic in secondary mathematics curricula (e.g., New Zealand: Ministry of Education, 1992; USA: NCTM, 2000) and is the focus of the findings presented in this article. Students' correlational reasoning skills improve as they mature (Moritz, 2004), and secondary school students are at the appropriate maturity level to study the correlation coefficient.

Research regarding the learning of this statistic documents that students come to the topic with misconceptions that need to be addressed through students' learning experiences. Students tend to view data as a series of individual cases rather than as a whole (Bakker, 2004) making it difficult for students to understand what the correlation coefficient is measuring. In Estepa and Batanero's (1996) study, secondary students' pre-instruction strategies for determining association included looking to see whether the pattern held true for each data point (consistently increasing or decreasing), basing their decisions on one part of the data, or relating their decisions to their personal beliefs about whether an association existed between the variables. There also needed to be a strong correlation between the variables before the students detected it. There is also a widespread belief by students that correlation is transitive (Sotos, Vanhoof, Van Den Noortgate, & Onghena, 2009); in other words, if quantitative data sets A and B have a positive correlation, and so do B and C, then students mistakenly believe that A and C must have a positive correlation.

In a teaching experiment, Estepa and Sanchez-Cobo (2003) worked with first-year university students to build their meaning of statistical association. With intervention, these students learned to use the complete data set to determine association and came to

understand that judging association should be done in terms of intensity rather than existence, overcoming the determinist misconception. The students had a difficult time understanding inverse association; they did not understand how to interpret a negative correlation coefficient and failed to understand the concept by the end of the teaching experiment. This may be related to the paucity of examples in secondary school textbooks displaying data sets with a negative correlation (Estepa & Sanchez-Cobo, 1998). Taken together, these findings show the necessity and efficacy of instruction on the correlation coefficient.

2. METHODOLOGY AND RESEARCH DESIGN

2.1. PARTICIPANTS

This research involved case studies of three secondary mathematics teachers as they taught the topic of statistical association. The primary criterion used in the selection of the participants was high quality teaching, defined as that which fosters a view of mathematics as sense making activity (Yackel & Hanna, 2003) and aims for students to learn to think statistically, reason statistically, and become statistically literate citizens. There has been an oft-repeated and intensifying call for educators to emphasize these components in their teaching (Ben-Zvi & Garfield, 2004; NCTM, 2009), and thus in this study I purposefully used them to define the type of teaching I was looking for my participants to engage in. Through conversations with mathematics department chairs and observations of potential participants, I looked for evidence of high quality teaching in potential participants' lessons including asking students for justification of procedures or answers and providing activities where students developed their own methods for solving problems.

The first participant was Mr. Glass (all names used are pseudonyms), a veteran teacher of secondary mathematics with approximately thirty years of teaching experience at a suburban school. He taught the topic of statistical association in his Pre-Calculus/Statistics course composed of year 11 students. Mr. Tablet, another participating teacher, taught the Advanced Placement Statistics course to year 11 and year 12 students at a different suburban secondary school. This course is intended to be equivalent to a semester college-level non-calculus-based introductory statistics course. Mr. Tablet has been teaching secondary mathematics for twenty-three years. The final participating teacher, Ms. Tuck, has taught for five years at an urban school for college-preparatory students. I observed Ms. Tuck's teaching of statistical association to her year 9 and year 10 students.

With only three participating teachers in the study, the observed practice of teaching was not comprehensive with respect to all teaching contexts. It is certainly possible that secondary teachers teaching statistical association could need subject matter knowledge that was not identified by this study. However, through the purposeful selection of three teachers using different curricula to teach statistical association in varying school years and settings in a total of fifty class sessions, the findings become more compelling.

2.2. METHODS

Procedures In accordance with a practice-based approach, I collected data by observing the participating teachers whenever their class sessions pertained to the topic of statistical association as determined by each class' curriculum. I observed sixteen days with Mr. Glass, twenty-six days with Mr. Tablet, and eight days with Ms. Tuck. During

the observations, which were audio recorded, I took fieldnotes documenting what was happening in the class, including notes of things that the audio recording would not document such as notes written on the board. My attention focused upon the teacher, but I also noted student comments, questions, solutions, and claims, because these student-teacher interactions called on the teacher to apply his or her knowledge of the subject. In addition to the audiotapes and fieldnotes of observed class sessions, I collected additional records of practice, which included handouts distributed during the lesson; presentation materials used by the teacher, such as computer files or overhead slides; assessment tasks; and copies of student work. Copies of the relevant textbook pages for each lesson were also made.

Following each observed class session, I interviewed the participating teacher about the statistics knowledge he or she drew upon when planning and teaching the lesson, and with regard to the assessment of student learning. I used the interviews to work towards understanding the subject matter knowledge used in teaching statistical association from the participating teachers' perspectives rather than mine (Merriam, 1998). All interviews were audio recorded and followed a semi-structured format, which involved a mix of predetermined questions and exploratory questions created at the time of the interview. Some of the predetermined questions dealt with the planning of the lesson, such as "How did you choose the definitions of terms and their symbols used in the lesson?" Others dealt with assessment, asking questions related to the subject matter knowledge used by teachers in the design and evaluation of assessments. The predetermined questions were supplemented with questions regarding specific occurrences in the classroom session observed. This format allowed me to respond to the situation observed and to new ideas on the topic. During the interview I read through my observation fieldnotes and formulated questions for the teacher regarding what they were thinking about from a subject matter perspective related to a specific incident or document they had used in class. To illustrate, I asked questions during interviews regarding a teacher's choice of a data set and what a teacher was thinking about in responding to a student.

Data Sources The audiotapes from all observed class sessions and interviews were transcribed. My fieldnotes from the observations were a helpful resource during the creation of the transcripts, providing information regarding a setting or observed activity that the audiotape did not include. The audiotapes, transcripts, and fieldnotes taken together represent a data triangulation (Denzin, 1978) regarding what occurred during each observed class session. Utilizing multiple sources helps to provide cross-data validity checks (Patton, 1990). Additionally, each teacher reviewed the transcripts from his or her class sessions and interviews for the member checking process (Stake, 1995) to test the accuracy of the data collected. Following this process, modifications to the transcripts and notes were made based on the participants' feedback with the intent of creating an accurate account of the classroom and interview sessions.

The transcript and relevant records of practice for each observed class session were brought together to create a compilation document. Each document included the full transcript, essentially a written re-creation of the class period observed, followed by the interview transcript and copies of the relevant textbook pages. Assessment materials were the next component of each compilation document. Most often this was a homework assignment, and in this case the textbook pages or handouts containing the assignment were included. Other types of assessments included quizzes, tests, or projects, copies of which were included in the compilation document for any days that assessment was referred to. If the teacher needed different subject matter knowledge to assess an assignment than that used in the observed class period, I asked the participating teacher

for anonymous copies of student work. The student work was added as the final component of the compilation documents. Lastly, the lines were numbered in each compilation document to provide a point of reference during the analysis process. Together a total of fifty compilation documents were created, corresponding to the fifty class sessions observed for this study.

Analysis Process The analysis process began with my study of all of the compilation documents. As I read each compilation document, I broke apart the sequence of events or items into teaching incidents. A teaching incident refers to a teaching activity or objects that can be extracted from the rest of the compilation document and still make sense standing alone. This could be any incident involved in teaching, including a student question, an explanation provided by the teacher, student work on a homework problem, or a test question. Then, for each teaching incident, I analyzed the subject matter knowledge needed by a teacher in that situation. The teachers' knowledge could not be documented directly through the data sources, nor was the purpose to analyze any particular teacher's knowledge. Instead, I used the data sources as a catalyst for developing conjectures regarding the subject matter knowledge a particular teaching incident might entail (Ball & Bass, 2000b). The data sources grounded the analysis in the practice of teaching, helping me to generate claims regarding the subject matter knowledge needed by teachers involved in the tasks of their profession. When writing the knowledge descriptors, I aimed to explain the particular understanding of the topic that the teacher needed, with an end goal of creating a fertile description that would be useable and useful to teachers, educators, and researchers.

A piece of knowledge does not stand alone; there exists a learning hierarchy for each topic which describes the progression of prerequisite skills needed to master the topic (Gagné, 1985). A limit had to be set regarding how far back or forward to go in describing the knowledge needed for understanding a topic. I decided that the description of knowledge needed should generally go no further than two classes or grade levels, either previous or future, from the current class. This allowed me to unpack the knowledge needed for a topic so that a rich description would be provided without getting bogged down in the process by having to go all the way back to the most elemental prerequisite knowledge components.

I created a written report, called an analysis report, for each compilation document. It contained knowledge descriptors for each teaching incident and the line numbers of the compilation document referring to the teaching incident. This provided an explicit correspondence between the practice of teaching and the subject matter knowledge that is needed in that practice, supporting the validity of the findings. Each knowledge descriptor included provided enough detail so that an unambiguous explanation of the concept that the teacher needed to know was created.

Additional Analysts Twelve of the compilation documents were analyzed by two additional persons: a statistician and a statistics education expert. This constituted a type of triangulation known as triangulating analysts, which enhanced the quality and credibility of the analysis and its findings (Patton, 1990). Due to the complex and dynamic phenomena of teaching and learning mathematics, where much remains hidden and needs interpretation and analysis, it was important to use an interdisciplinary group of experts to analyze the data (Ball & Bass, 2003). I merged the analysis reports created by the statistician and statistics education expert with the reports I constructed, thus creating a meta-analysis report for these twelve compilation documents.

As not all of the compilation documents were analyzed by the three analysts, it is possible that the data in the documents analyzed only by myself could have resulted in different knowledge descriptors written by the statistician and statistics education expert. However, the overall agreement of the findings of the multiple analysts for the twelve selected compilation documents (73% of the knowledge descriptors were listed by multiple analysts) indicates that my findings were consistent with theirs. It is also possible that there are aspects of the subject matter knowledge that one or all three analysts missed entirely, but this was minimized through a review of the findings by the participating teachers and analysts.

Illustrative teaching incidents To illustrate the analysis process, I have selected two teaching incidents from Mr. Glass' class when he introduced the topic of correlation coefficient and the resulting knowledge components identified in the meta-analysis report by the three analysts. On this day, Mr. Glass asked his class to find the regression line for a set of data using their graphing calculator and the value for *r*, the correlation coefficient, was included in the calculator output. He used this as the starting point for discussing the correlation coefficient: "OK. So right now, the only models we're going to be worried about in this section are linear models. OK. And that's where this value comes in. It's called the correlation coefficient *r*." Mr. Glass continued on to describe the values *r* can take. In the second teaching incident, which occurred just after Mr. Glass' description, a student asked "What is a correlation coefficient? I know how to find it and everything, I just don't know what it represents." When analyzing these teaching incidents, the three analysts used the context of these situations as a means for considering what the teacher needed to know to meet the demands of high quality teaching.

Table 2 gives the knowledge descriptors from the meta-analysis report related to these teaching incidents with the line numbers removed.

Table 2. Excerpt from meta-analysis report

Analyst A:	*r*: Correlation coefficient—possible values. Relationship between correlation coefficient (positive or negative) and the gradient of the graph. Relative strength of relationship as *r* is closer to 1. Reliability of prediction. Interpretation for different values of *r* (e.g., -1, 0, 1).
Analyst B:	*r*: Correlation coefficient, statistic that measures the strength and direction of a linear relationship; only works for linear relationships (or those that can be transformed to be linear like exponential) because its calculation is based on *z*-scores, which are linear functions of the data values*;* number sense: range is -1 to 1, no relationship is 0, perfect linear relationships are 1 and -1, strengths relative to values between 0 and 1 (and 0 and -1), reasonable ranges for *r* to label as weak, moderate, and strong associations based on the context; *r* is positive when the data have a positive association; calculation of *r*: formula, how formula works (e.g., why *r* is positive when the data have a positive association; why *r* can't be larger than 1); slope of LSRL = $r\times(s_y/s_x)$, therefore *r* is not the slope but does vary directly with the slope.
Analyst C:	Correlation coefficient (symbol = *r*); calculation of; essentially the sum of product of *z*-scores so it is positive/negative when "most" of the data are in quadrants I and III/ II and IV. Measure of strength and direction of linear association.

These knowledge descriptors provide many things that the statements from Table 1 do not: rich, thick descriptions of the knowledge component; drilled-down descriptions of knowledge focusing on the correlation coefficient; connections between multiple knowledge components of statistical association; and linkage to specific instances of teaching practice that called for this type of knowledge. As these knowledge descriptors emanated from documented teaching activities, they can speak to where and how teacher knowledge is used in practice as well as what aspects of content knowledge teachers need to know and how they need to know it.

Some portions of these knowledge descriptors, such as knowing *r* ranges from -1 to 1 and that it is positive when there is a positive association, are facts any user of this statistic should know and thus would be considered common content knowledge. Other portions, like understanding how the formula works, address the specialized knowledge a teacher needs that a person using this statistic in practice may not. An advantage of the practice-based approach for studying teacher knowledge is that it results in findings that inherently include both common and specialized knowledge descriptors.

Synthesis In the next phase of analysis, I synthesized the data from the twelve meta-analysis reports and the thirty-eight analysis reports that I constructed alone. This was the first time the data from all three participating teachers were brought together so that all of the documented teaching incidents could be included. Categories or tags of the knowledge descriptors were made using the constant comparative method (Glaser & Strauss, 1967). I began by creating a tag that identified the subject matter topic for each knowledge descriptor. Knowledge descriptors received multiple tags if they linked together multiple ideas. For example, all three knowledge descriptors in Table 2 were given the tag "r" to denote that they dealt with the correlation coefficient. The last sentence in Analyst B's knowledge descriptor was also given the tag "LSRL" because it described the relationship between the correlation coefficient and the slope of the least-squares regression line. As I went through this process I minimized the number of tags that I used, always using a previously identified tag if possible.

Once tagged, the knowledge descriptors were re-sorted into groupings based on their tags. All the knowledge descriptors for each tag were placed in a table together. If a knowledge descriptor had multiple tags, then it was listed in the table for each tag. Every knowledge descriptor had an accompanying compilation document number with line numbers to maintain the explicit connection between the descriptors and the teaching incidents from which they originated. At the end of this process, twenty-nine categories were defined, nine of which dealt with pre-requisite knowledge and were therefore removed. The remaining twenty categories are listed in Table 3 along with the frequency of their use in teaching measured by the number of teaching incidents that referenced that topic.

The topic of the findings in Section 3.2, the correlation coefficient, was the second-most referenced topic in my research (see Table 3) and is interrelated with many of the other knowledge components.

3. RESULTS

3.1. KNOWLEDGE FOR TEACHING STATISTICAL ASSOCIATION

The overall findings of my research (Casey, 2008) provide a comprehensive, detailed description of the subject matter knowledge needed by teachers to teach statistical association at the secondary level. It is beyond the scope of this article to report the entire

Table 3. Knowledge descriptor categories for statistical association

Category	Number of referenced teaching incidents
Best Fit Line (including LSRL)	116
Correlation Coefficient	100
Chi-square Statistic and Test (general)	96
Mathematical Modeling (excluding Best Fit Line)	83
Chi-square Test for Independence	64
Residual Plot	61
Calculator	60
Residual	54
Predictions	52
Coefficient of Determination	46
t test for Slope of Regression Line	38
Association	37
Scatterplot	36
Context	35
Terminology	26
Data	21
Computer	19
Two-Way Tables	15
Confidence Interval for Slope of Regression Line	10
Outliers & Influential Points	10

twenty-seven page description, but the concept map in Figure 1 visually displays the knowledge components' groupings and their relationships. All of the knowledge components described in Section 3.2 fall into at least one of these groupings.

At the top of the concept map are three foundational knowledge components for teaching statistical association: meaning, terminology, and knowledge of context. Next the map breaks into two branches, quantitative and categorical, corresponding to the types of data being studied for association. Within each of these branches, the map outlines ways to analyze the data for association as well as technology for viewing the analysis process. The map can be read like a sentence from the top oval down to any of the ovals below. For example, reading along one of the threads of the map from top to bottom the sentence "Statistical association between categorical data sets can be analyzed descriptively including graphically by a segmented bar graph" can be created. Although the knowledge components' groupings are linearly related and separated on the map, in practice they are interconnected, both within themselves and with other mathematical and statistical knowledge, and interact during the process of teaching.

3.2. TEACHING ABOUT THE CORRELATION COEFFICIENT

This section presents a narrative of what teachers need to know when teaching about the correlation coefficient at the secondary level. The correlation coefficient is one of multiple statistics included in the "Statistics" category at the bottom middle of the concept map in Figure 1. Included in this section are knowledge descriptors for the teaching of correlation coefficient as well as excerpts from the records of practice representing situations in which teachers needed this type of knowledge.

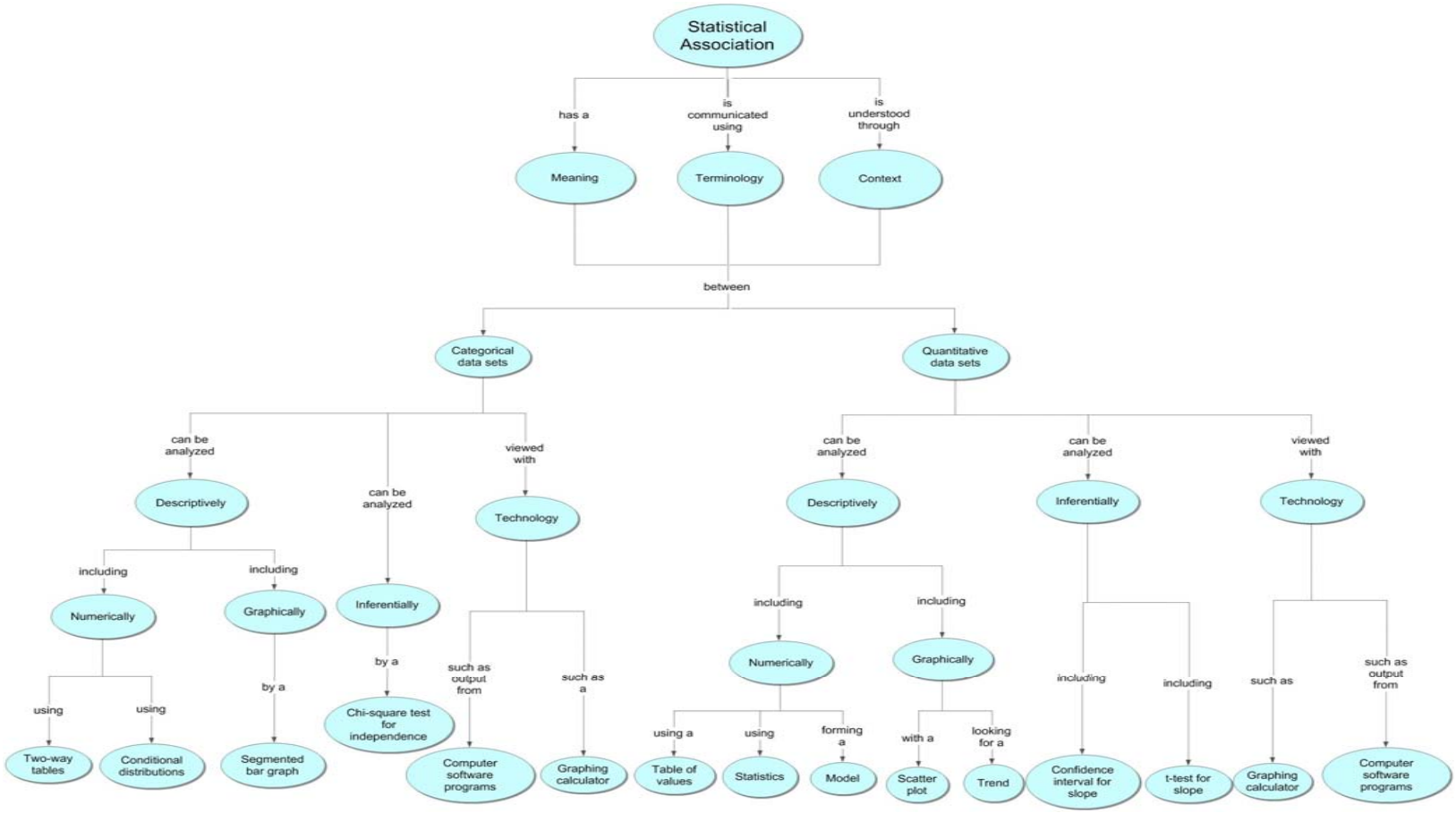

Figure 1. Concept map of knowledge for teaching statistical association

The description begins with knowledge about the computation of *r* which leads into understanding its properties. This is followed by a piece on interpreting the values of *r*, including when it is equal to zero. The next two segments address the sensitivity of the correlation coefficient and estimating its value. The section ends with a discussion about the use of appropriate terminology when teaching the topic of correlation coefficient.

Computation The correlation coefficient is a statistic which measures the direction and strength of a linear association between quantitative variables. In order to teach about the correlation coefficient, teachers must know how and why it is calculated as it is in order to measure linear association. The teaching incidents cited in Section 2.2 illustrated the need for this type of knowledge. One equation for the correlation coefficient involves the product of *z*-scores:

(1) $$r = \frac{\sum z_x z_y}{n-1}$$

More than just knowledge of the equation, teachers need to understand how it works to create a statistic that describes the strength and direction of a linear association between the *x*- and *y*-variables. One way to gain this understanding is to analyze a scatterplot by creating four quadrants through the graphing of the *x*-bar and *y*-bar lines. Those points in the resulting first and third quadrants will create positive products of *z*-scores, whereas those points in the resulting second and fourth quadrants will create negative products. Therefore, summing all of these products gives a value that calculates whether or not the data have a positive or negative association overall.

A thorough understanding of equation (1) can facilitate the understanding of other knowledge components teachers need. For example, the correlation coefficient does not change when the units of measurement of *x*, *y*, or both are changed because it uses standardized z-scores in its calculations. This gives the correlation coefficient the added advantage of being able to be compared across various data sets. Teachers can also understand why the correlation coefficient is sensitive to outlying points by understanding it uses the mean and standard deviation in its calculation. The formula also explains numerically why the correlation coefficient of *y* on *x* is the same as the correlation coefficient of *x* on *y* and why it only applies to bivariate quantitative data.

Interpretation Interpreting the value of the correlation coefficient involves numerous knowledge components, including understanding the reasons it ranges from -1 to +1, its sign indicates the direction of the association, association between the variables is considered stronger the closer the absolute value of *r* gets to 1, and understanding it is only appropriate for linear data. A particularly interesting case is when the correlation coefficient has a value of zero as illustrated by the following three scenarios.

Mr. Glass assigned his students a homework problem that presented a scatterplot of points following a horizontal line then asked the students to identify and explain whether the correlation coefficient was 1, 0, or -1. A task such as this is a sense making activity for the students and thus is in line with the definition of high quality teaching. Some of the students who chose zero gave the following explanations: there's no line of best fit, the slope is zero, because the points don't go positive or negative, there is no change in direction. Although these students are choosing the correct answer of zero, not all of them are justifying it correctly. In order to assess these students' work, a teacher needs to know why statisticians decided to define *r* as zero when the points follow a horizontal line. If one conceptualizes correlation as a measure of how much help *x* gives you in predicting *y*, then you can explain that because all values of *x* produce the same value for *y*, correlation should be zero in this case. Other justifications include the direct relationship between *r*

and the slope of the least-squares regression line or its relationship as the square root of the coefficient of determination. Knowledge of all of these reasons provides a rigorous knowledge foundation upon which teachers can base their work.

Another scenario where the value of *r* is zero is when the points follow a non-linear pattern such as a quadratic function. It is important that teachers know and help their students to learn why the correlation coefficient should be used only to measure association that is linear in nature. Students in Mr. Tablet's class did a homework problem that gave them a data set that followed a quadratic pattern. The students were asked to show and explain why the correlation is zero for this data set even though there is a strong relationship between the two variables, which is another example of a sense making activity for the students. One student answered because part goes up and part goes down making not one or the other. Another responded that $r = +1$ when you group the first half of the points together, $r = -1$ when you group the second half of the points together, which gives a sum of 0. With thorough knowledge of the correlation coefficient, grounded in understanding of its computation and its direct relationship with the slope of the best fit line, a teacher can aptly evaluate these student responses.

If a scatterplot presents a random scatter of points with no association, once again the correlation coefficient should be approximately zero. Teachers should know why this happens from a conceptual as well as formulaic perspective. From a conceptual standpoint, because knowing the value of *x* provides no predictive power, the correlation coefficient should be zero. From the formula for *r*, the positive *z*-score products will be approximately the same value as the negative *z*-score products, resulting in a sum of the *z*-score products near zero and therefore an *r* close to zero as well. The direct relationship between *r* and the slope of the regression line also reinforces the fact that zero is the value of *r* in this scenario. All of these knowledge components help provide the knowledge basis needed for teaching about interpreting values of *r*, including how its computation works in the case of no association. This is related to the *t* test for the slope of the regression line as well. This inference procedure is included in the required curriculum for the Advanced Placement Statistics course (College Board, 2008) and was taught by Mr. Tablet in his class. He explained that testing whether the slope of the regression line was zero or not is equivalent to testing to whether or not the correlation is zero. A student said "Like if the *r* value was like for example 0.2 or something." The teacher responded "Well again we have to look into the output. It's a good question because you're asking the question—well still how do we know if 0.2, 0.5, 0.6, 0.7, if any of this is going to be significant." Here the teacher needed to know that whether or not a value of *r* is significantly different from zero depends not only on the magnitude of the difference between *r* and zero but also on the number of points in the data set. Teachers also need to know that if *r* is deemed statistically significantly different from zero, then the two variables are considered to have a significant association.

Sensitivity Related to interpreting the value of *r* is an understanding of how the alteration of points in a data set changes its number. Mr. Tablet asked his class "What's going to make the *r* value bigger according to this formula [equation (1)]?" A student replied "When the points are closer together." The real-time demands of teaching require that teachers be able to judge a student's response and reply appropriately at that moment. Here the teacher needed to have a robust understanding of the calculation of *r* in order to determine whether the student's statement was correct, including knowledge that *r* is only used to assess how well the data points follow a line as opposed to any closer grouping of the points. Teachers also need to understand the effect of adding or removing outliers or influential points to a data set, as textbooks including those used by Mr. Glass (Senk et

al., 1998) and Mr. Tablet (Yates, Moore, & McCabe, 1999) and standardized tests (e.g., College Board, 2003) often include such problems. In particular, they need to understand the effect on *r* of adding or removing a point that follows the pattern but has a much larger or smaller *x*-value than the rest of the points, a point away from the pattern with a much larger or smaller *x*-value than the rest of the points, and a point in the middle of the domain with a much larger or smaller *y*-value than the other points. This understanding can again be gained by a thorough understanding of the computation of *r*. For example, adding a point that has a much larger or smaller *x*-value than the rest of the points yet is near the best fit line will result in a stronger correlation because that point's *z*-score product will be large due to the large differences between its coordinates and the values of the means of the *x*'s and *y*'s.

Estimating Assessments also ask for students to estimate the value of the correlation coefficient given a scatterplot, and both Mr. Glass and Mr. Tablet assigned homework problems related to this skill. In order to develop this skill in students, teachers need a complete understanding of the computation of the correlation coefficient and the impact unusual observations can have on its value. A teaching activity that calls for such knowledge is the creation of example data sets for students to use when developing and practicing this skill, as seen in Mr. Glass' teaching when he had to create scatterplots for his students to use in class.

Terminology A teacher needs to be fluent in the language of statistics in order to teach statistical association with high quality, particularly for the development of students' statistical literacy. As in most fields, statistics has its own terminology which teachers must use accurately and appropriately in their work with students. This includes use of terminology during all-class instruction, when answering student questions, and when writing and evaluating assessments for students. Some statistical terms related to correlation coefficient that teachers need to have a thorough understanding of include random, outlier, influential point, explanatory variable, response variable, independence, association, and regression. An instance that called for such knowledge was when a student in Mr. Tablet's class asked him "What's the difference between correlation, association, and regression?" In order to respond to such a question, a teacher needs to know not only the meaning of each term but the complex relationships amongst them.

4. DISCUSSION

4.1. USING A PRACTICE-BASED APPROACH IN RESEARCH

My research is unique because it is the first to use a practice-based approach to study the subject matter knowledge needed for teaching statistics at the secondary level. In the literature, there has been a call for research on teacher knowledge that studies it in the context where it is used (Ball & Bass, 2000a; Ball et al., 2001; Sorto, 2004). This study has made a significant contribution to the field by using a methodology to study the knowledge for teaching in action. The conclusions regarding teacher knowledge are grounded in the work of high quality teaching and thus better able to improve teacher practice and policy (Ball et al., 2001). This methodology makes it possible to answer the important questions of exactly what aspects of mathematics teachers need to know, how they need to know it, and where and how it is used in practice (Ball & Bass, 2003).

4.2. USING SUBJECT MATTER KNOWLEDGE TO INFORM THE FIELD

This study contributes to the research base regarding the subject matter knowledge for teaching statistics, an area of need identified in the literature (Burgess, 2007; Groth, 2007; Shaughnessy, 2007). The findings can mold the formation of a general framework regarding the statistical knowledge for teaching. For example, the findings support Groth's inclusion of mathematical and nonmathematical knowledge for teaching, but call into question whether his framework includes components such as knowledge of technology.

The findings presented in this paper showed that teachers need a substantial knowledge base for teaching the specific topic of correlation coefficient. It is necessary for teachers to know the how *and* why of concepts to meet the demands of teaching. For example, it is not enough for teachers to be able to calculate the value of the correlation coefficient. They also need to know how it is computed, why it is computed that way, and the implications of its computation such as non-resistance. Professionals who use the computation of r in their work should know that the correlation coefficient of a data set that follows a horizontal line is zero, but they probably do not need to know why it has been defined that way and how that connects to the slope of the regression line, definition of R^2, and the conceptual meaning of correlation as teachers do. The linking of the how and why of topics may be something that sets apart specialized mathematical knowledge for teaching statistical association from common mathematical knowledge of statistical association. My findings maintain the existence of these two types of knowledge, as theorized by Hill et al. (2004), for teachers of statistics.

The practice-based description of the subject matter knowledge needed for teaching statistical association at the secondary level has uses for several communities: secondary statistics teachers, post-secondary statistics teachers, teacher educators, developers of curriculum for teachers, policy makers, and assessment developers. The uses of the findings of my research will be explained by addressing each of these communities in turn.

Secondary statistics teachers can use the results both as a learning tool and an informal assessment tool. The findings may be used as a resource for secondary statistics teachers to learn more about statistical association. They can also be used for teachers to gauge their own knowledge in the field, determine areas of strengths and weaknesses, and determine any follow-up actions that may be needed.

New kinds of statistics courses need to be created in order to help teachers develop their knowledge for teaching (Groth, 2007). These practice-based findings can help inform the faculty responsible for designing and implementing the curriculum for such courses in two ways. First, the description of the knowledge needed by teachers designates content that should be taught in these courses. Second, excerpts from the records of practice can be used in the courses to simultaneously develop teachers' subject matter knowledge and pedagogical content knowledge, an approach proven effective at doing so (e.g., Groth, 2006), and keep teachers' learning experiences grounded in situations that arise in teaching.

Policy documents (e.g., CBMS, 2001; NCTM, 2005) have emphasized the importance of a strong subject matter knowledge base for teaching, but none have described this knowledge base at the level of depth or detail contained in the findings from this study. These results can contribute towards an expanded description of the subject matter knowledge for teaching statistics and help policy makers make decisions regarding the requirements for teaching certification. In this same vein, the findings of this study can

speak to the content that should be contained in assessments for teachers used in the certification process.

4.3. RECOMMENDATIONS FOR FUTURE RESEARCH

There are many directions that future research regarding the subject matter knowledge for teaching statistics may take. One area of research could focus upon efforts to help teachers develop the knowledge needed for teaching statistical association as identified by this study. Work needs to be done to determine how best to help teachers not only possess this knowledge but possess it in a way that is readily accessible and ready to be used in the unpredictable arena of the classroom (Ball & Bass, 2000a). This should include learning opportunities for both prospective and in-service teachers through offerings like courses or teacher study groups.

The research of Hill et al. (2005) showed that greater knowledge of mathematics by teachers was linked to higher achievement of students. Another line of research could investigate whether this is true for teachers and students in the area of statistical association, using the description of the knowledge needed for teaching statistical association from the present study as the foundation. Research that seeks evidence for the effects of greater teacher knowledge is important to justify the work that is done in teacher education and to attempt to establish empirically the relationship between teacher knowledge and student achievement (Ball & Hill, 2005).

ACKNOWLEDGMENTS

The research reported in this article was part of my dissertation study conducted at Illinois State University under the direction of Dr. Cynthia Langrall. I also acknowledge the work of statistician Julie Clark and statistics education expert Tim Burgess for serving as analysts for this study. Thank you to the editor and reviewers for your helpful comments to improve the article.

REFERENCES

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ3%282%29_Bakker.pdf]

Ball, D. L., & Bass, H. (2000a). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83–104). Westport, CT: Ablex.

Ball, D. L., & Bass, H. (2000b). Making believe: The collective construction of public mathematical knowledge in the elementary classroom. In D. Phillips (Ed.), *Yearbook of the National Society for the Study of Education, Constructivism in Education* (pp. 193–224). Chicago: University of Chicago Press.

Ball, D. L., & Bass, H. (2003). Toward a practice-based theory of mathematical knowledge for teaching. In B. Davis & E. Simmt (Eds.), *Proceedings of the 2002 Annual Meeting of the Canadian Mathematics Education Study Group* (pp. 3–14). Edmonton, AB: Canadian Mathematics Education Study Group.

Ball, D., Cole, Y., Delaney, S., Fauskanger, J., Kwon, M., Mosvold, R., & Ng, D. (2009, April). *Adapting and using U.S. measures of mathematical knowledge for teaching in*

other countries: Lessons and challenges. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

Ball, D. L., & Hill, H. C. (2005, April). *Knowing mathematics as a teacher*. Presentation at the Annual Meeting of the National Council of Teachers of Mathematics, Anaheim, CA.

Ball, D. L., Lubienski, S., & Mewborn, D. (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.

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

Burgess, T. A. (2007). *Investigating the nature of teacher knowledge needed and used in teaching statistics* (Unpublished doctoral dissertation). Massey University, Auckland, New Zealand.

Casey, S. (2008). *Subject matter knowledge for teaching statistical association* (Unpublished doctoral dissertation). Illinois State University, Normal, IL.

College Board (2003). *AP statistics 2003 free-response questions form B.* New York: Author.
[Online: www.collegeboard.com/prod_downloads/ap/students/statistics/b_statistics_frq_03.pdf]

College Board (2008). *AP Statistics course description.* New York: Author.
[Online: http://apcentral.collegeboard.com/apc/public/repository/ap08_statistics_coursedesc.pdf]

Conference Board of the Mathematical Sciences (2001). *The mathematical education of teachers.* Washington, DC: Mathematical Association of America; Providence, RI: American Mathematical Society.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79–95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed.). New York: McGraw-Hill.

Estepa, A., & Batanero, C. (1996). Judgments of correlation in scatterplots: Students' intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education, 4*, 25–41.

Estepa, A., & Sanchez-Cobo, F. T. (1998). Correlation and regression in secondary school textbooks. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on the Teaching of Statistics* (vol. 2, pp. 671–676). Voorburg, The Netherlands: International Statistical Institute.

Estepa, A., & Sanchez-Cobo, F. T. (2003). Evaluacion de la comprension de la correlacion y regression a partir de la resolucion de problemas. *Statistics Education Research Journal, 2*(1), 54–68.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ2%281%29.pdf]

Franklin, C. (2000). Are our teachers prepared to provide instruction in statistics at the K-12 levels? *National Council of Teachers of Mathematics Education Dialogues, 10.*
[Online: http://www.nctm.org/resources/content.aspx?id=1776]

Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 47–78). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Gagné, R. M. (1985). *The conditions of learning.* New York: Holt, Rinehart, & Winston.

Garfield, J. & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The*

challenge of developing statistical literacy, reasoning, and thinking (pp. 397–409). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Groth, R. E. (2006). Analysis of an online case discussion about teaching stochastics. *Mathematics Teacher Education and Development, 7*, 53–71.

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

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *Elementary School Journal, 105*(1), 11–30.

Kettenring, J., Lindsay, B., & Siegmund, D. (Eds.) (2003). *Statistics: Challenges and opportunities for the 21st century.* National Science Foundation Report. [Online: http://www.stat.psu.edu/~bgl/nsf_report.pdf]

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54*(1), 33–61.

Merriam, S. B. (1998). *Qualitative research and case study applications in education.* San Francisco: Jossey-Bass Publishers.

Ministry of Education. (1992). *Mathematics in the New Zealand curriculum.* Wellington, NZ: Author.

Moore, D. (2004). Foreword. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. ix–x). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Moritz, J. (2004). Reasoning about covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 227–255). Dordrecht, The Netherlands: Kluwer Academic Publishers.

National Board for Professional Teaching Standards. (2001). *Adolescence and Young Adulthood Mathematics Standards* (2nd ed.). Arlington, VA: Author. [Online: http://www.nbpts.org/userfiles/File/aya_math_standards.pdf]

National Council for Accreditation of Teacher Education/NCTM. (2003). *Program Standards.* Reston, VA: Authors. [Online: http://www.nctm.org/uploadedFiles/Math_Standards/NCTMSECONStandards.pdf]

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2005). Highly qualified teachers: A position of the National Council of Teachers of Mathematics. *NCTM News Bulletin, 42*(3), 4.

National Council of Teachers of Mathematics. (2009). *Focus in high school mathematics: Reasoning and sense making.* Reston, VA: Author.

National Research Council. (2001). *Knowing and learning mathematics for teaching: Proceedings of a workshop.* Washington, DC: National Academy Press.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Qualifications and Curriculum Authority. (2007). *The National Curriculum 2007.* Earlsdon Park, Coventry: Author. [Online: http://curriculum.qcda.gov.uk]

RAND Mathematics Study Panel. (2003). *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education.*

Washington, DC: Office of Education Research and Improvement, U.S. Department of Education.

Senk, S., Viktora, S., Usisking, Z., Ahbel, N., Highstone, V., Witonsky, D., … Schultz, J. E. (1998). *Functions, Statistics, and Trigonometry* (2nd ed.). Glenview, IL: Scott Foresman and Company.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Reston, VA: NCTM.

Sorto, M. A. (2004). *Prospective middle school teachers' knowledge about data analysis and its application to teaching* (Unpublished doctoral dissertation). Michigan State University, Ann Arbor, MI.

Sotos, A. E. C., Vanhoof, S., Van Den Noortgate, W., & Onghena, P. (2009). The transitivity misconception of Pearson's correlation coefficient. *Statistics Education Research Journal*, *8*(2), 33–55.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ8%282%29_Sotos.pdf]

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage Publications.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–265.

Wilson, P. S. (2004, August). *GAISE report discussion.* Presentation at the Joint Statistical Meetings, Toronto, Canada.
[Online: http://www.amstat.org/education/gaise/WilsonJSM04.ppt]

Yackel, E., & Hanna, G. (2003). Reasoning and proof. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to the principles and standards for school mathematics* (pp. 227–236). Reston, VA: NCTM.

Yates, D. S., Moore, D. S., & McCabe, G. P. (1999). *The Practice of Statistics: TI-83 Graphing Calculator Enhanced.* New York: W.H. Freeman.

STEPHANIE A. CASEY
Deerfield High School
1959 North Waukegan Road
Deerfield, Illinois 60015
USA