SECONDARY MATHEMATICS TEACHERS' INFORMAL INFERENTIAL REASONING: THE ROLE OF KNOWLEDGE STRUCTURES FOR MEASURES OF CENTER, SPREAD AND SHAPE

A Dissertation

Presented to

the Faculty of the Graduate School

University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

CHRISTOPHER ENGLEDOWL

Dr. James E. Tarr, Dissertation Advisor

JULY 2017

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled

SECONDARY MATHEMATICS TEACHERS' INFORMAL INFERENTIAL REASONING: THE ROLE OF KNOWLEDGE STRUCTURES FOR MEASURES OF CENTER, SPREAD AND SHAPE

Presented by Christopher Engledowl,

A candidate for the degree of Doctor of Philosophy,

And hereby certify that in their opinion it is worthy of acceptance.

_____

Professor James E. Tarr

_____

Professor Samuel Otten

_____

Professor Corey Webel

_____

Professor John Lannin

_____

Professor David Bergin

For Gretchen and Leela

ACKNOWLEDGEMENTS

A dissertation is not completed without the aid of a community of supporters. Therefore, this section reflects on those who have made this dissertation possible.

First, I thank my advisor, Dr. James E. Tarr, for his dedication to ensuring that this study was rigorous and of high quality. His continuous support, critical feedback, and thought-provoking questions guided my study from conception, to recruitment of participants, data collection, analysis and into its final form. Moreover, I am grateful for the generosity of Drs. Samuel Otten, Corey Webel, John Lannin, and David Bergin for their feedback and questions that aided in refining my study and results. The committee's engagement significantly strengthened my study.

I would like to also thank my pilot study participants—Susan King, Isaac Townsend, and Matthew Wilson. The time and feedback they provided was crucial to choosing tasks and adapting the interview protocol to sufficiently prepare for any missing contingencies and to refine existing ones. Moreover, a special thank you to the nine teachers that graciously donated their limited time to participate in this study. Their openness and honesty allowed for important insights into their thinking.

Extending my appreciation, I would like to thank Randall Groth and Tim Jacobbe, whose thoughtful conversations provided important refinements to the study's method. Further, I thank all of the instructors of the courses that aided in deepening my thinking and analytic skills, including interactions with other mathematics education faculty members and graduate students at various conferences throughout the last four years.

Without the support of family, perseverance to complete this dissertation would have been much more difficult. I would like to thank my parents, Brian and Barbara Engledowl, for their constant encouragement and for listening when things became

difficult. Moreover, I would like to thank my in-laws, Alan and Miriam Gustaf, for their support over the past 4 years.

Most importantly, I would like to thank my wife, Gretchen, for her continual support, encouragement, and love. Her dedication to her family and this major life change of returning to graduate school has not only pushed me to achieve my goals, but has also grounded me. I also thank my daughter, Leela, who was born just one week before my comprehensive exams. She changed my life forever and incentivized me to work even harder. Thank you, Gretchen and Leela, for your love and support. I owe my success to you.

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study examined middle and secondary mathematics teachers' knowledge structures, informal inferential reasoning (IIR), and pedagogical content knowledge (PCK) for statistics. Using task-based clinical interviews (Goldin, 1997) and cross-case analysis (Creswell, 2013), a stratified purposeful sample (Patton, 2002) of nine teachers responded to the Goals and Outcomes Associated with Learning Statistics (GOALS-2) instrument (Sabbag & Zieffler, 2015), released items from the Levels of Conceptual Understanding in Statistics (LOCUS) assessment (Jacobbe, 2016) and supplemental questions to assess PCK (Watson et al., 2008). Responses were used to construct maps of teachers' knowledge structures for measures of center, spread, and shape (Groth & Bergner, 2013) and knowledge structures were analyzed for common characteristics. Teachers' IIR was coded for the appropriateness of responses (Means & Voss, 1996) and key components of IIR (Makar & Rubin, 2009) were identified. To distinguish teachers' PCK level, descriptions of four hierarchical levels were used (Callingham & Watson, 2011) and knowledge structures were classified as *desirable-connected*, *undesirable-connected*, and *undesirable-disconnected*. Although teachers largely engaged in the *inference* and *data* components of IIR, they rarely referenced the *uncertainty* component. In general, teachers with more connected knowledge structures and fewer undesirable knowledge elements exhibited more *acceptable* forms of IIR and higher PCK levels. Within IIR contexts, teachers struggled to exhibit *acceptable* forms of IIR and demonstrated the lowest levels of PCK, but within non-IIR contexts, they exhibited *acceptable* reasoning more often as well as higher PCK levels. Implications for teacher education are discussed and recommendations for future research are offered.

CHAPTER 1: STATEMENT OF THE PROBLEM AND BACKGROUND

Data-based decision making has become ubiquitous, resulting in a data-dependent society. Websites like Mint (www.mint.com) track every dollar you spend and what you spend it on, and wearable technology (e.g., Fitbit, smart watches) constantly tracks health-related factors to provide on-demand, readily available data to help inform daily personal decisions. Similarly, it is a common expectation that statistical information be consumed and used to make business-related decisions. For instance, the current trending practice among retail stores is to make use of the store's free smartphone application to collect several types of data about shoppers, such as their specific location in the store (and how much time is spent there), and then to offer suggestions through the application to an individual based on observed shopping patterns (see e.g., Siraj Datoo, 2014). This kind of data is important to store employees as they work to maximize sales and to quickly target shoppers who may need assistance or who may appear likely to make use of the store's credit card. Thus, much like the ability to read, statistical literacy (see Ben-Zvi & Garfield, 2005a) has become necessary to daily life.

The Purpose of the Study

Globally, there has been a push for primary and secondary school students to develop statistical literacy (Ben-Zvi & Garfield, 2005a; Franklin et al., 2007) and in the United States, many states have included more statistics content in their curriculum standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Although the increased attention at the K-12 level has been recent, introductory statistics courses have had wide appeal to many college degree programs for multiple decades (GAISE College Report ASA Revision Committee, 2016). Yet, at the college level, there is consistent evidence over the past 20 years that students

completing introductory undergraduate statistics courses are far from being proficient in their statistical reasoning (e.g., delMas, Garfield, Ooms, & Chance, 2007). It is therefore no surprise that although most middle level and secondary mathematics teachers in the U.S. have taken at least one college level statistics course, they do not feel prepared to teach statistics content (Banilower et al., 2013). Perhaps more concerning is that although informal inferential reasoning (IIR) has recently become a major goal of middle and high school grades (Franklin et al., 2007; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), introductory college statistics courses largely focus on formal inferential reasoning (delMas et al., 2007). Furthermore, even courses specifically designed for preservice teachers may not offer opportunities for engaging in informal inferential reasoning (Huey, 2011).

Despite the seemingly frail backgrounds middle level and high school teachers have with statistics content knowledge, some teachers have found ways to deepen their knowledge through professional development and personal study (Peters, 2013). However, even teachers with strong statistics content knowledge often struggle to translate it into pedagogical content knowledge (PCK) (Watson, Callingham, & Donne, 2008)—an important type of knowledge found to be predictive of positive gains in student achievement (e.g., Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; H. C. Hill, Rowan, & Ball, 2005). Such findings about the importance of PCK led the National Mathematics Advisory Panel (2008) to call for research into how differences in teacher knowledge are related to teacher effectiveness. Moreover, for over a decade, research on PCK in statistics has been identified as a critical area of need (Langrall, Makar, Nilsson, & Shaughnessy, 2017; Shaughnessy, 2007).

One way to address these calls for research into connections between teacher knowledge and PCK is to examine differences among teachers' *knowledge structures* (similarly called *cognitive structures*, *mental constructs* or *mental schemes*). By understanding more deeply the misconceptions and contradictory conceptions teachers have, specific interventions could be developed to target the reconstruction of less desirable knowledge structures into more desirable structures that may translate into more productive PCK (Groth & Bergner, 2013; Ron, Dreyfus, & Hershkowitz, 2010). For example, examinations of knowledge structures have been particularly helpful in developing an understanding of children's conceptions of fractions (e.g., L. Steffe, 2001), which in turn has informed studies of elementary teachers' pedagogical content knowledge (e.g., Izsák, 2008).

Given the importance placed on informal inferential reasoning (IIR) with statistics (defined in a later section), the purpose of this study is to examine the knowledge structures of teachers that may support IIR. In particular, structures that make connections among conceptions of center, spread, and shape of distributions are vital for IIR. The *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* includes these three conceptions as important for analyzing data and making inferences even at the lowest developmental level (Level A) of statistical literacy (Franklin et al., 2007), and the National Assessment of Educational Progress (NAEP) places an emphasis on conceptions of center, spread, and shape of distributions (and connections between them) in their assessment framework for all grade levels that are assessed—4th, 8th, and 12th (National Assessment Governing Board & U.S. Department of Education, 2012).

Research Questions

This study is motivated by the calls for more research into the connection between teachers' content knowledge and their PCK (National Mathematics Advisory Panel, 2008), and the recent increased importance placed on teachers for knowing statistics content and how to teach it (Ben-Zvi & Garfield, 2005b; Groth, 2007; H. C. Hill et al., 2005). More specifically, it examines teachers' knowledge structures of statistics content as well as the relation between those structures and teachers' informal inferential reasoning. Accordingly, I proposed the following research questions:

1. What knowledge structures do middle level and secondary mathematics teachers have regarding center, spread, and shape of distributions?
2. How do teachers' knowledge structures support informal inferential reasoning?
3. What is the relationship between teachers' informal inferential reasoning and pedagogical content knowledge?

The next few sections provide an overview of how I studied each of the three research questions. In particular, I first describe my theoretical and conceptual framing, followed by a description of participants, data collection methods, and data analysis methods.

Theoretical Framework

Epistemological Perspective

For this study, I take a *constructivist* perspective that views knowledge as constructed by the individual that learning occurs through integrating newly constructed knowledge elements into one's cognition (Herscovics, 1996). Given this perspective, the term *misconception* is problematic because of its connotation that the conception must be

replaced, whereas a term such as *preconception* implies that the conception can serve as a conduit for more productive ways of thinking (Booker, 1996). Consequently, instead of using the terms *correct* and *incorrect*, I use the terms *desirable* and *undesirable*, implying that an *undesirable* knowledge element contains a perceived flaw that may limit more productive ways of thinking. Therefore, the primary focus of the study is on individuals' knowledge structures and how knowledge elements within those structures might be connected.

Informal Inferential Reasoning

Many studies of informal inferential reasoning have been published over the past decade (e.g., Dolor & Noll, 2015; Garfield, DelMas, & Zieffler, 2012; Leavy, 2010; Makar & Confrey, 2003; Makar, 2014; Pfannkuch, 2011; Stohl Lee, Angotti, & Tarr, 2010), including a special issue on the topic in *Mathematical Thinking and Learning* (Makar, Bakker, & Ben-Zvi, 2011). In an attempt to establish a shared meaning of the term, Zieffler, Garfield, delMas, and Reading (2008) defined IIR as "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (p. 44). Building on this definition, they constructed a framework for research on IIR comprising the following three components:

1. Making judgments, claims, or predictions about populations based on samples, but not using formal statistical procedures and methods (e.g., *p*-value, t tests);
2. Drawing on, utilizing, and integrating prior knowledge (e.g., formal knowledge about foundational concepts, such as distribution or average; informal knowledge about inference such as recognition that a sample may be

surprising given a particular claim; use of statistical language), to the extent that this knowledge is available; and

3. Articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples. (2008, p. 45)

For purposes of this study, I drew on Zieffler and colleagues' (2008) definition, but I revised it to include ideas from Rossman (2008) that expand the definition to include inferences made about causality between explanatory and response variables, in addition to claims made about populations from samples.

Informal Inferential Reasoning Task Characteristics

Although task types are suggested by Zieffler and colleagues (2008), they do not place much attention on the characteristics that would exemplify such tasks. Therefore, I drew on Huey and Jackson's (2015) informal inferential reasoning task framework as a guide to selecting and developing tasks. Within their framework, Huey and Jackson (2015) argued that tasks requiring informal inferential reasoning must (1) require *inference* beyond the data while acknowledging variation, (2) be *ill-structured* (no prescribed solution path), (3) be *open-ended* (multiple solution paths possible), (4) be embedded within a *context* that must be acknowledged in the inference, and (5) contain *visual representations* to be used when making inferences.

Conceptual Framework

Informal Inferential Reasoning

The conceptual framework I drew on for *informal inferential reasoning*, depicted in Figure 1.1, describes the relationship between knowledge elements (e.g., center, spread, shape of distribution) and informal reasoning.

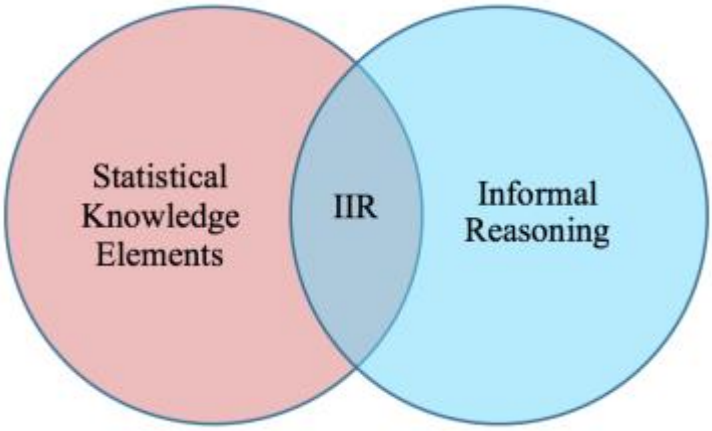


Figure 1.1. Conceptual framework for IIR

While others (e.g., Huey, 2011) have conceptualized that the statistical knowledge elements and informal reasoning abilities are precursors to IIR, the GAISE report (Franklin et al., 2007) does not make this assumption. In the GAISE report, knowledge and reasoning are assumed to develop simultaneously, as seen in the use of the word *and* in the following quote: "Students also should learn how to use basic statistical tools to analyze the data *and* make informal inferences in answering the posed questions" (p. 23, emphasis added). Moreover, there is evidence that even children as young as third grade are able to make informal inferences without formally being introduced to measures of center and spread—even making generalizations beyond their classroom (Makar, 2014). If this is the case, IIR can be observed despite a dearth of knowledge elements. Similarly, in Zieffler and colleagues' (2008) literature review, informal reasoning was found to be not associated with development of content knowledge. Additionally, they found that informal reasoning does not seem to improve with "maturation, education, or life experience" (Zieffler et al., 2008, p. 44) and that even improvement in general intelligence, while positively associated with informal reasoning abilities, results only in "people selectively [using] that intelligence to build their own case rather than to explore

an issue more fully" (p. 44). Therefore, I hypothesize the statistical knowledge elements and informal reasoning abilities as distinct and necessary components of informal inferential reasoning, of which neither are required for the development of the other.

Knowledge Elements. To further explicate the framework in Figure 1, statistical knowledge elements include measures of center, spread, and shape of distribution. Research on students' conceptions of center has shown a range of conceptions extending from (a) basic procedural understanding, to (b) recognizing measures of center as tools for analysis, to (c) understanding that they represent something typical about the data, to (d) the highest level of being able to describe when one measure may be more useful than another (Groth & Bergner, 2006, p. 51). It is important to note that in order to obtain this highest level, it is necessary to integrate notions of shape and spread. It is widely accepted that notions of center, spread, and shape are inherently integrated (e.g., A Bakker, 2004). For instance, according to Noll and Shaughnessy's (2012) developmental framework of hierarchical levels for reasoning with variability (spread), in order to reach the highest level, it is necessary to integrate at least two notions of center, spread, and shape. Lower levels of their framework involved drawing only on frequencies or only drawing on a single measure. Regarding shape, there is evidence that teachers struggle to interpret skewed distributions. For example, Doerr and Jacob (2011) found that teachers tend to think that normal (symmetric) distributions imply less variability and do not attend to how skew effects variability.

Informal Reasoning. As for the informal reasoning component of the framework, I drew on Means and Voss' (1996) work with informal reasoning that described a way to evaluate the quality of students' informal reasoning. In particular, I used the following

two questions form their analytic framework. The other questions they included were not useful because either (a) they did not translate to this context (e.g., counterarguments would not necessarily make sense to offer) and (b) the interview protocol explicitly asks for them so it doesn't make sense to consider that a measure of quality (e.g., all teachers were prompted to consider "multiple sides" in the interview protocol):

1. What claims are being made and what reasons are offered?
2. Is the reason being offered acceptable and does it support the claim?

A reason was identified as acceptable if it drew on *desirable* knowledge elements. For instance, an acceptable reason for a claim might be if a teacher claims that group A is more consistent than group B because group A has a smaller range. In this case, the range is a *desirable* knowledge element to describe variability (spread).

Pedagogical Content Knowledge

When Lee Shulman first coined the term *pedagogical content knowledge* (PCK), he described it as "subject matter knowledge for teaching" that included the knowledge of the "conceptions and preconceptions students bring with them" to the classroom, and "knowledge of strategies" to aid learners in developing desired conceptions of the content (1986, pp. 9–10). Hill and colleagues (2008) further developed, and reconceptualized, a framework of PCK by including three components: knowledge of content and students, knowledge of content and teaching, and knowledge of curriculum. PCK is one of two subsections under the larger domain of *mathematical knowledge for teaching* (MKT)—the other of which is *subject matter knowledge*.

Several frameworks for statistical knowledge for teaching have been developed (e.g., Godino, Ortiz, Roa, & Wilhelmi, 2011; Groth, 2007). In particular, Groth (2007) began an initial discussion around developing a *statistical knowledge for teaching* (SKT)

framework (see Figure 1.2). Within the SKT framework, a major difference between MKT and SKT is made explicit by the inclusion of the components *mathematical* and *nonmathematical* to make explicit that there are fundamental differences between mathematics and statistics and yet statistics does draw on mathematics frequently (e.g., Franklin et al., 2007). To illustrate, consider the knowledge element of the arithmetic mean. Groth describes that "identifying the mathematical properties of the mean that can be difficult for students to comprehend" requires primarily *mathematical* knowledge (2007, p. 430). On the other hand, "realizing that students may compute the arithmetic mean for a data set without regard for the context of the data" requires primarily *nonmathematical* knowledge—that is, knowledge specific to statistics.

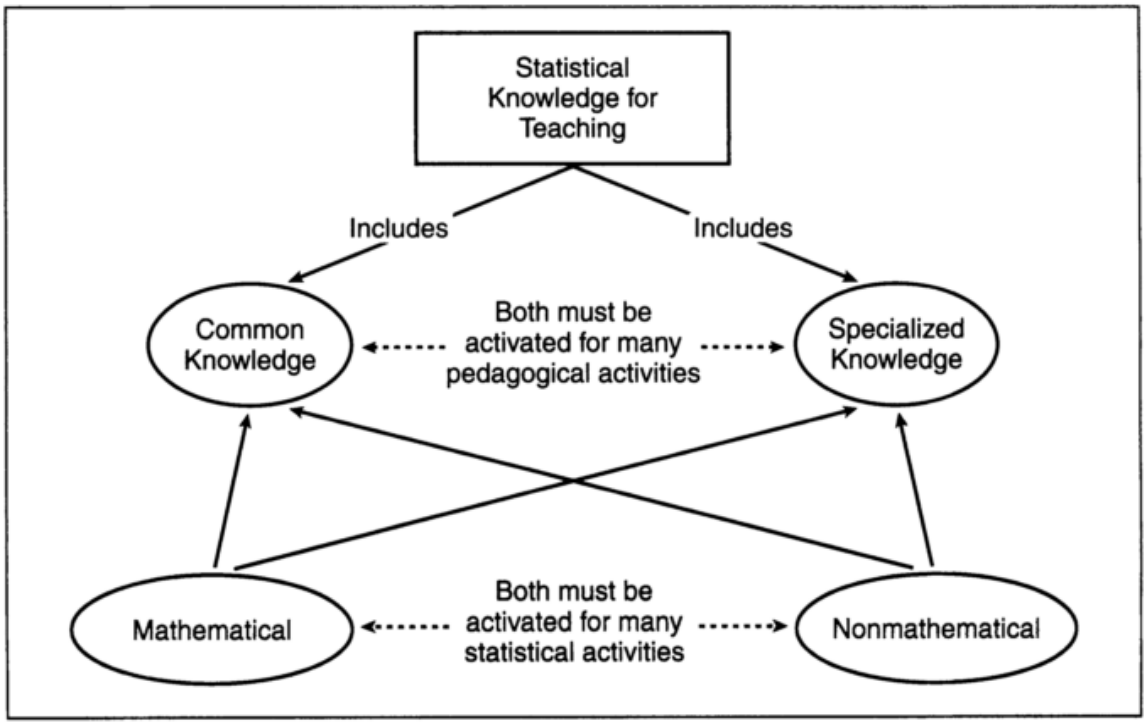


Figure 1.2. Statistical Knowledge for Teaching Framework, taken from Groth (2007)

Groth (2013) updated this framework to include the PCK elements from Hill and colleagues (2008) while retaining that both mathematical and nonmathematical elements

exist. In this study, I focused on teachers' knowledge of content and students and knowledge of content teaching.

To further conceptualize teachers' knowledge of content and students within the context of statistics, I drew on Callingham and Watson (2011), who identified four hierarchical levels of PCK for statistics: *Aware*, *Emerging*, *Competent*, and *Accomplished*. These levels were identified from written responses to two questions about statistics tasks given to teachers in a task-based survey:

1. What are some appropriate and inappropriate student responses?
2. How might you intervene with a student who made an inappropriate response?

Teachers were also provided with student responses to statistical tasks. At the lowest level, *Aware*, teachers struggled to generate student responses and to address student misunderstandings with suggested interventions. At the *Emerging* level, teachers could offer a single appropriate or inappropriate response and make generic intervention suggestions. At the *Competent* level, both types of responses were generated and suggestions for intervention were sometimes content specific, "but only in the context of familiar activities" (Callingham & Watson, 2011, p. 290). At the *Accomplished* level, suggestions for intervention involved an "integration of appropriate statistical content with student-centered intervention strategies" (Callingham & Watson, 2011, p. 290). Therefore, Callingham and Watson's elaboration of these four levels of PCK provides a framework of different types of knowledge of content and students that teachers might have within the context of statistics.

Mapping Cognitive Structures

Aligning with my theoretical perspective that assumes an individual's knowledge is never complete, I drew on literature that examines the ways in which a person's knowledge elements may be connected—otherwise known as knowledge structures. In pursuit of a way to capture a person's knowledge structure, Ron and colleagues (2010) created a method for mapping what they term *partially-correct constructions* (PaCCs). According to Ron and colleagues (2010), PaCCs occur when "knowledge constructs only partially match the underlying mathematics," can lead to both incorrect and correct answers, and can be characterized by "inconsistent student answers or actions" (p. 65). Moreover, their study revealed three types of PaCCs—*missing element*, *incompatible element*, and *disconnected element*. They describe a *missing element* as one that has not been constructed, an *incompatible element* as one that contradicts another, and a *disconnected element* as one that is not connected to another knowledge element. In a subsequent study by Groth and Bergner (2013), a fourth type of PaCC was observed that contains both *missing* and *incompatible* elements. It should be noted that PaCCs are constructed based on the researcher's interpretation of the relationship between observed knowledge elements because an individual's actual knowledge structure is latent.

Ron and colleagues (2010) were further motivated to find a way to map knowledge structures because they observed students making what seemed to be contradictory statements, and arriving at desirable responses to tasks in unexpected (and undesirable) ways. Therefore, they claimed it was important to use a diverse set of modes of data across multiple time points in order to observe *incompatible* knowledge elements and verify *missing* and *disconnected* elements.

To illustrate the use of PaCCs, Groth and Bergner (2013) used PaCCs to study 31 elementary level pre-service teachers during a methods course that included several assignments related to categorical data. As a way to help visualize the knowledge structures, Groth and Bergner (2013) adapted Ron and colleagues' (2010) method of mapping PaCCs to include node-link diagrams. They used a combination of tasks for assessing statistics content knowledge and pedagogical content knowledge to aid in developing a knowledge structure node-link diagram for each participant.

Although the concept of PaCCs were originally designed to trace the construction of knowledge toward intended learning goals, the concept of a PaCC as a *knowledge structure* that represents all *desirable* and *undesirable* knowledge elements—including how they may be connected—is particularly useful for this study.

Overview of the Method

I used a task-based clinical interview design with a cross-case analysis to study nine practicing middle and secondary mathematics teachers' statistics knowledge, informal inferential reasoning, and PCK. In order to participate, they were required to have taught statistics content that included data analysis explicitly using measures of center, spread, and shape of distributions. Moreover, a *stratified purposeful sample* (Patton, 2002) was obtained in order to have representation across all contexts in which statistics is taught in middle and secondary grades. Thus, there were four strata:

- Statistics taught as a unit within middle level mathematics ($N = 3$)
- Statistics taught as a unit within secondary mathematics ($N = 2$)
- Non-Advanced Placement (AP) Statistics ($N = 2$)
- AP Statistics ($N = 2$)

Data were drawn from three different sources and collected across two modes. First, teachers completed the GOALS-2 assessment (Sabbag & Zieffler, 2015) online, which contains 20 forced-choice items related to formal statistics. This served to provide context for the extent of formal experiences teachers had. It also was used as a secondary check on observed knowledge elements, when applicable. Second, four released, constructed-response tasks from the *Levels of Conceptual Understanding of Statistics* (LOCUS) assessment (Jacobbe, 2016) were completed during 60–90 minute video and audio recorded task-based clinical interviews (Goldin, 1997). Responses to LOCUS tasks aided in answering the first and second research questions regarding knowledge elements and IIR. Last, during interviews, after completing each LOCUS task, teachers were asked the following two questions from Watson, Callingham, and Donne (2008):

- What are some inappropriate and appropriate student responses to this task?
- How might you respond to a student who offered one of the inappropriate responses?

Responses to these two questions mainly served to answer research question three regarding PCK. However, in the process of discussing students, sometimes more information was found to be useful in answering the other research questions as well.

The data analysis first involved transcribing teachers' verbal responses to tasks and coding for observed *desirable* and *undesirable* knowledge elements, as outlined in the conceptual framework (see Chapter 3 for details). Then, visual diagramming software was used to construct node-link diagrams to represent each participant's knowledge structure map (Groth & Bergner, 2013). Once knowledge structure maps were

constructed for each participant, a within-case analysis was carried out to confirm each structure and then a cross-case analysis was carried out to identify common themes across structures in an effort to categorize types of knowledge structures.

In order to answer the second research question about IIR, transcript data was coded first according to the questions adapted from Means and Voss' (1996). Next, each observed argument was coded according to the three components of IIR described by Makar and Rubin (2009): generalization beyond the data, data as evidence, and probabilistic language. A within-case analysis was carried out through the use of analytic memos and then a cross-case analysis was used to identify categories of types of IIR. Moreover, types of IIR were compared with types of knowledge structures to characterize how teachers' knowledge structures may support their IIR.

To aid in answering research question three regarding PCK, responses to supplemental questions to the LOCUS assessment were analyzed. First, I identified appropriate and inappropriate student responses, whether generated responses were common responses on the LOCUS, and whether responses were researcher-provided. Then, I coded suggested interventions for whether they addressed the inappropriate nature of the response, drew on the student response, related specifically to the content and context, and whether it was a generic intervention. After looking across all suggestions for each task for each participant, a PCK level (Callingham & Watson, 2011) was distinguished (Aware, Emergent, Competent, Accomplished). A cross-case analysis was then completed to aid in describing types of PCK. Last, these types were then compared with knowledge types and IIR types to describe possible relationships between knowledge and PCK and IIR and PCK.

Significance of the Study

At present, there is a lack of research into teachers' IIR and PCK that would inform a comprehensive professional development program to support teachers' implementation of the statistics content they are required to teach. Franklin describes this as a "critical challenge to the successful implementation of the statistics standards" (2013, p. 3) and *The Mathematical Education of Teachers II* (MET II) report notes the need of such professional development programs since "most new high school teachers will require further coursework to be well prepared to teach … more than basic statistics" (Conference Board of the Mathematical Sciences, 2012, p. 19). Further, *The Statistical Education of Teachers* (SET) report by the American Statistical Association recognizes the need for professional development programs to include development of knowledge about "common student conceptions and thinking patterns" (Franklin et al., 2015, p. 3). More specifically, this study contributes to the statistics education research community by providing a fine-grained look at how teachers' knowledge structures may support their IIR and how their knowledge and IIR relates to their PCK. As professional development programs designed to engage teachers in statistical content through IIR, and that include knowledge about common student conceptions, increases in demand, as recommended in the SET report (Franklin et al., 2015), such information will be indispensable for constructing a successful program.

Organization of This Dissertation

The rest of this dissertation is organized to first provide a more in-depth rationale for the study through a literature review in Chapter 2. Next, the methods of data collection and analysis are explained in detail in Chapter 3. Results of the analysis of data collected are laid out in Chapter 4 and organized by research question. Last, in Chapter 5,

I provide a discussion of results followed by implications for teacher education and recommendations for future research.

CHAPTER 2: LITERATURE REVIEW

Over the next several sections, I establish the development and meaning of key terms used in this study and then proceed to describe what is currently known about 1) teachers' knowledge of measures of center, spread, and shape of distributions and how it might be mapped or measured, 2) teachers' engagement and experience with *informal inferential reasoning*, and 3) the relationship between teachers' statistics knowledge, their *informal inferential reasoning*, and their knowledge of students in these contexts. In doing so, I will establish the need for research that examines the ways in which statistics knowledge supports IIR, and how statistics knowledge and IIR may interact with statistics PCK.

History of Published Use of the Term Informal Inferential Reasoning

A review of the literature indicates that one of the earliest uses of the term *informal inferential reasoning* grew out of research into students' statistical thinking and was included in the conference proceedings of the *5th International Conference on Teaching Statistics* (ICOTS 5). In the conference proceedings, Scheaffer mentioned the term while describing what statistical thinking might look like across K–12 mathematics, stating that in the middle grades "informal inferential reasoning can begin" (1998, p. 22). However, he does not attempt to describe what it means, implying that perhaps it was a commonly known term among those who attend ICOTS. The next mention of the term in published literature appears to be from the conference proceedings at ICOTS 7 by Pfannkuch (2006b), titled simply *Informal Inferential Reasoning*. In the published proceedings, Pfannkuch states that *informal inferential reasoning* "is interconnected to reasoning from distributions, reasoning with measures of centre, and sampling reasoning within an empirical enquiry cycle" (p. 1). She further defined *informal inference* as "the

drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data" (p.1).

A more detailed discussion of *informal inferential reasoning* occurred during Pfannkuch's (2006a) publication in *Statistics Education Research Journal* in which she studied a grade 11 teacher during two episodes of teaching students to compare and interpret boxplots. A main goal of instruction in these teaching episodes was to "make an informal inference about populations when comparing sample distributions and to justify that inference" (Pfannkuch, 2006a, p. 33). She described *informal inferential reasoning* as "in the case of box plots, being able to infer that one group is generally greater than a second group, or that no distinction can be drawn, based mainly on looking at, comparing, and reasoning from box plot distributions" (Pfannkuch, 2006a, p. 28). Moreover, she described the inherent difficulty in studying this phenomenon, stating that "there appears, however, to be … no definitive account of how teachers or students should draw informal inferences" (Pfannkuch, 2006a, p. 29)—indicating that perhaps this study focused on *informal inferential reasoning* was the first of its kind. It should also be noted that this study was part of a larger project that began in 2003 to study students' statistical thinking. During the first year, the researchers realized that students and teachers struggled to make informal inferences, and because there was no research or curricular materials to draw on, Pfannkuch's 2006 study was initiated.

One year later, 2007, the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* was published and endorsed by the American Statistics Association. This document, designed to complement the Data Analysis and Probability standards in *Principles and Standards for School Mathematics* (National Council of

Teachers of Mathematics, 2000) and with an eye toward statistical literacy by grade 12, included a statement that students should make "informal inferences" at the lowest level of statistical literacy (Franklin et al., 2007, p. 23). That same year, Michael Shaughnessy's (2007) chapter on statistics and probability in the *Second Handbook of Research on Mathematics Teaching and Learning* contained a section dedicated to informal inference (pp. 957–1009). Although both of these highly influential publications did not use the term *informal inferential reasoning* precisely, they both discussed informal inference in the context of attention to students' reasoning.

By 2008, the term *informal inferential reasoning* had become prevalent within the statistics education community, and *Statistics Education Research Journal* published a special issue on the topic that included eight articles (Pratt & Ainley, 2008). Three years later, *Mathematical Thinking and Learning* also published a special issue on the topic including eight articles (Makar & Ben-Zvi, 2011), thus bringing *informal inferential reasoning* into the broader mathematics education community's view. However, the overwhelming majority of articles published in both of these special issues were developed during the 5th and 6th International Forum on Statistical Reasoning, Thinking and Literacy—held in the United Kingdom and Australia in 2007 and 2009, respectively. The impact and importance of research regarding *informal inferential reasoning* continues to grow, gaining an explicit section dedicated to it in the newly published chapter on probability and statistics in the *Compendium for Research in Mathematics Education* (Langrall et al., 2017)—colloquially known as the third handbook, following Lester and colleagues' (2007) second handbook.

In subsequent sections, I review the literature on *informal inferential reasoning*, its importance to developing formal inferential reasoning, and what gaps my study seeks to fill in an effort to aid the field in moving forward.

Defining Key Terms

Before reviewing the literature relevant to this study, a common understanding of terms used in this study is needed. In this section, the terms *informal inferential reasoning*, and what is meant by *desirable knowledge elements* are laid out.

Defining Informal Inferential Reasoning

As indicated by Langrall and colleagues (2017), there have been many ways of describing the term *informal inferential reasoning* and it seems the field has not yet settled on a common definition—although Langrall and colleagues point out that some commonalities exist across them. Before expressing the definition this study builds on, I first provide some background on the terms *informal knowledge* and *informal reasoning*, to support an understanding of the way I am defining *informal inferential reasoning*.

Informal knowledge. What is meant by *informal*? The term is used often in literature without sufficient explanation of what precisely is meant by this term. Moreover, it is sometimes used interchangeably with the term *intuitive* without drawing a distinction (see Amerom, 2003; Jones, 1995) and even Farmaki and Paschos employ the hyphenated *intuitive-informal* (2007, p. 356), thereby making it more difficult to separate the two notions. To further confuse the two terms, several publications have set the term *intuitive* in contrast to *formal*, making it easy to assume that *informal* and *intuitive* have the same meaning.

Fischbein and Schnarch (1997) have written about the term *intuition* and describe it as "a cognition that appears subjectively as self-evident, directly acceptable, holistic,

coercive, and extrapolative" and furthermore can be characterized by "the feeling of obviousness, of intrinsic certainty" (p. 96). Using this more specified description of *intuition*, a distinction can be made from *informal*.

To aid in drawing a distinction, within the broader field of mathematics education, Lave (e.g., 1984) can be cited for his work in understanding the use of *informal* mathematics outside of the school context. Similar work has also been done within statistics education (e.g., Arthur Bakker, Kent, Derry, Noss, & Hoyles, 2008). Moreover, Zieffler and colleagues (2008) append to this notion of informal knowledge as "real-world" knowledge, that informal knowledge is that which has resulted from a "less formalized knowledge" constructed from prior formal instruction (p. 42). In this way, *informal* knowledge is distinct from *intuition*—intuition does not draw on such prior formal instruction. Moreover, Zieffler and colleagues go on to summarize studies comparing experts and novices that imply that informal knowledge is important for students to have in order to recreate formal knowledge.

For the purposes of this study, I follow Zieffler and colleagues' (2008) description of *informal knowledge* that describes it as involving both non-school related experiences and knowledge resulting from prior formal instruction (p. 43). Moreover, it is important to note that Zieffler and colleagues also found in their literature review that informal knowledge was important for developing formal knowledge—this claim will be discussed further in a later section.

Informal reasoning. Examining the literature on *informal reasoning*, as noted by Zieffler and colleagues (2008), there is no commonly accepted definition. However, for this study, I use Zieffler and colleagues' description of the term, in which they draw on

an edited book by Voss, Perkins and Segal (1991) and the chapter by Perkins, Farady, and Bushey (1991). From these sources, they first draw on Voss and colleagues (1991) to define *informal reasoning* as that which occurs "in non-deductive situations, such as decision making, that is employed in everyday life" (Zieffler et al., 2008, p. 43). Expanding this further, Zieffler and colleagues quote Perkins and colleagues (1991, p. 85), claiming *informal reasoning* is exemplified by a person considering all facets of a situation and then using "common sense, causal, and intentional principles" to consider alternative solutions (Zieffler et al., 2008, p. 43).

To further express the nature of *informal reasoning*, Zieffler and colleagues' (2008) literature review on this topic resulted in a couple notable conclusions. First, content knowledge, experience, and maturity do not always relate positively with the quality of informal reasoning (p. 45). Therefore, it cannot be measured indirectly through these other means. Second, they found that although general intelligence appears to influence higher quality of informal reasoning, such intelligence is typically not used to fully explore an argument. Instead, people tend to "use that intelligence to build their own case" (p. 45). Lastly, and most importantly, Zieffler and colleagues found that despite the lack of positive interaction between *informal reasoning* and other domains that might be expected to improve it, *informal reasoning* can be improved through instruction (p. 45).

Informal inferential reasoning. After a thorough review of the literature, in an attempt to define and provide a framework for research on *informal inferential reasoning*, Zieffler and colleagues (2008) defined it as "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (p. 44). Rossman (2008) argues for two distinct

forms of inference. The first is the one indicated by Zieffler and colleagues (2008), namely, inferring something about a larger population from a sample. The second form of inference that Rossman (2008) adds to Zieffler and colleagues' definition is the concept of inferring causality between explanatory and response variables when there is random assignment, which also requires recognition of variability between samples. Therefore, this study employs the use of both possibilities by defining *informal inferential reasoning* as:

> the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (Zieffler et al., 2008, p. 44) or to support inferences
> "by drawing a more profound conclusion about the relationship between the variables (e.g., that the explanatory variable causes a change in the response). (Rossman, 2008, p. 5)

To further specify this term, I also use Zieffler and colleagues' elaboration that it is a process containing the following three components, which have been adapted based on Rossman's extension on the term *inference*:

1. Making judgments, claims, or predictions about populations based on samples [or about causality between variables (Rossman, 2008)], but not using formal statistical procedures and methods (e.g., p-value, t tests);
2. Drawing on, utilizing, and integrating prior knowledge (e.g., formal knowledge about foundational concepts, such as distribution or average; informal knowledge about inference such as recognition that a sample may be

surprising given a particular claim; use of statistical language), to the extent that this knowledge is available; and

3. Articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples [or about causality between variables (Rossman, 2008)]. (Zieffler et al., 2008, p. 45)

Desirable Knowledge Elements

Theoretical stance on knowledge. I use the term *desirable* knowledge elements instead of *correct* knowledge elements because of my theoretical perspective. Ontologically, I take the stance that, in the realm of mathematics and statistics, there is no absolute truth that is to be discovered, but rather that mathematics and statistics were created by humans in order to make sense of the world. This is evident in Piaget's work (1964/1997) that describes a child playing with pebbles who places a structure on the pebbles to organize them, thus abstracting an organization from the pebbles. The child did not discover an absolute truth about mathematics regarding the way the pebbles should be organized, but rather the child organized them in a way that aided in making sense of the pebbles at that time. Epistemologically, I take a *constructivist* stance that knowledge is constructed and learning occurs through integration of newly constructed knowledge elements into one's cognition (Herscovics, 1996). Thus, existing knowledge structures, including all its inconsistencies and perceived flaws, are viewed as guiding the *extension* of new ways of thinking (Duit, 1992, as cited in Booker, 1996).

Defining knowledge as desirable. Because of my theoretical stance on knowledge, terms such as *correct* and *incorrect* are inherently problematic. Therefore, I will use the terms *desirable* and *undesirable* to describe knowledge that may support or limit broader conceptions. Not only does this way of describing knowledge align with my

theoretical stance, it also keeps knowledge from being defined dichotomously. Although defining knowledge as *correct* and *incorrect* seems more parsimonious, consider a teacher who has constructed a knowledge element that the term *average* implies *arithmetic mean*. If this type of knowledge were to be viewed as *incorrect*, because *average* can also imply geometric mean, median, or mode, among others, then it ignores the times in which viewing *average* as *mean* is a useful way of thinking. By defining it as *undesirable*, the connotation is that there is something about the knowledge that may limit other ways of thinking about measures of center. A more detailed discussion of how *undesirable* and *desirable* knowledge elements are identified can be found in Chapter 3.

Teachers' Statistics Content Knowledge

Over the next few sections, I review literature regarding the broad overview of teachers' statistics content knowledge as well as the types of experiences teachers can be expected to have had with *informal inferential reasoning*. Moreover, this study focuses on knowledge related to measures of center, spread, and shape of distributions for three reasons. First, these are main knowledge elements assessed in the *National Assessment of Educational Progress* (National Assessment Governing Board & U.S. Department of Education, 2012), they are given extensive focus in the GAISE report (Franklin et al., 2007) and the CCSSM (NGA & CCSSO, 2010), and variation and distribution (which inherently incorporate ideas of center and shape) are two of the seven fundamental ideas in statistics listed by Burrill and Biehler (2011). Second, collecting data on advanced statistical topics from middle level and secondary mathematics teachers would likely be unproductive, given their anticipated limited background knowledge. Third, there is evidence that elementary school children can engage in *informal inferential reasoning* by drawing on broad, non-formalized, notions of center, spread, and shape of distributions

(e.g., Makar, 2014). Therefore, is reasonable to focus on these three types of knowledge as important aspects for making inferential statements, and teachers can be expected to have had some experience with each of these, even if it is informal rather than formal knowledge (e.g., typical vs mean, spread out vs standard deviation, bell-shaped vs normal). After reviewing literature on teachers' knowledge of center, spread, and shape of distributions, a discussion follows of how such knowledge might be measured and analyzed.

Statistics Content Knowledge

Over at least the past 20 years, researchers have consistently found that after completing a college-level, introductory statistics course, most students have not constructed desirable ways of thinking (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007; delMas et al., 2007; Falk & Greenbaum, 1995). Moreover, some have been found to be confident in their errors (Castro Sotos, Vanhoof, Noortgate, & Onghena, 2009), while others have developed new misconceptions by the end of the course that were not evident to begin with (delMas et al., 2007). Perhaps more compelling, Haller and Krauss (2002) found that among 30 instructors of introductory statistics courses across psychology departments at six German universities, 80% demonstrated at least one of the same errors as their students, echoing an earlier study by Oakes (1986), indicating that the landscape had not changed much over the past 16 years.

Given that degree plans for most prospective mathematics teachers require at least one course in statistics, as recommended by the National Mathematics Advisory Panel (2008), it is clear from these broader studies of introductory statistics courses that teachers likely do not have a sufficient understanding of statistics. Furthermore, few

studies have reported specifically on the statistical backgrounds of teachers and preservice teachers.

Perhaps the most convincing evidence of teachers' weak background in statistics comes from the *2012 National Survey of Science and Mathematics Education* (Banilower et al., 2013), which boasts a nationally representative sample of K–12 mathematics teachers in the United States. Results from this survey showed that although 69% of middle school mathematics teachers, and 83% of secondary mathematics teachers have taken at least one college level course in statistics, only 48% of middle school mathematics teachers and 30% of secondary mathematics teachers feel confident to teach statistics content. Despite this survey not being able to infer the cause for such a difference, it is reasonable to assume that a lack of understanding of statistics content (as described previously) is a main source of their lack of confidence.

To further support the argument that teachers do not have sufficient background in statistics, it is important to note that a sufficient understanding of the formal statistics typically taught in college-level introductory courses is not enough. The two most recent reform documents for statistics standards for students across grades 6–12 have been complementary in calling for students to engage in *informal inferential reasoning* with an expectation that by the end of 12th grade, students will have developed the formal counterpart. In the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) report, Franklin and colleagues (2007) claim that "every high school graduate should be able to use sound statistical reasoning to cope with the requirements of citizenship, employment, and family and to be prepared for a healthy, happy, and productive life" (p. 1), and they include informal inference as an important component of

a student's statistics education. Moreover, the *Common Core State Standards for Mathematics* (CCSSM) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) explicitly include drawing "informal comparative inferences" at grade 7 (7.SP.3 and 7.SP.4, p. 50) and it is implied in multiple standards where inference is expected but formal ways of testing significance have not yet been introduced. While one might claim that experiences with *informal inferential reasoning* likely occur within the context of courses designed for teachers, rather than the more general introductory statistics course, there is evidence that such courses (in the rare event that one exists specific to statistics) may not provide such experiences (Huey, 2011).

On a more positive note, Peters (2011) found that teachers are able to develop a deep understanding of statistics when self-motivated to seek ways to improve. Her study also points to an underlying issue, that these teachers largely had to look outside of their own districts for help in further developing their own statistical conceptions. This implies that teachers likely do not have adequate professional development opportunities to aid them in extending their background knowledge, experience with *informal inferential reasoning*, and in increasing their confidence in teaching statistics. There is some evidence that this is indeed the case, as found by Whitaker (2016) who studied the professional statistics teaching identities of 12 high school statistics teachers. Whitaker (2016) similarly found that teachers struggled to develop their identities as teachers of statistics because they had not been trained to be statistics teachers. Furthermore, teachers in this study claimed having to "[engage] with colleagues beyond their local school" (p. 997) in order to develop statistics teaching identities. This provides further evidence of

the sense of urgency for developing teachers' facility with and for teaching statistics that the *Statistics Education of Teachers* document lays out (Franklin et al., 2015).

Center

Research regarding teachers' conceptions of measures of center has been primarily conducted with pre-service elementary teachers. However, there has been some empirical study at all grade bands and with in-service teachers as well. In a literature review conducted by Jacobbe and Carvalho (2011), teachers and preservice teachers were generally found to 1) have a procedural understanding of measures of center and to prefer algorithms when asked about the center, 2) struggle to make connections between different measures of center, 3) prefer the mean over other measures, and 4) struggle to flexibly apply computations and consider effects of the outliers, distribution, and context on the result. Moreover, as noted by Jacobbe and Carvalho (2011), Leavy and O'Loughlin's (2006) study of preservice teachers' conceptions of mean found that a majority do not distinguish between the words "average" and "mean."

Despite these deficit findings, at least two studies described a range of conception types and one separated them into hierarchical levels to help understand the types of conceptions teachers may have and what limitations they inherently impose. Gfeller, Niess, and Lederman (1999) found that when solving problems involving the arithmetic mean, mathematics and science preservice teachers ($N = 19$) made use of either a procedural understanding or a "fair share" understanding (which the GAISE report places at the lowest level of understanding), which involved balancing deviations. Perhaps more useful in distinguishing conceptions, Groth and Bergner (2006) used the SOLO Taxonomy (Biggs & Collis, 1982) to describe preservice elementary teachers' thinking, when comparing mean, median, and mode, across four hierarchical levels:

- Unistructural: procedural understanding only;
- Multistructural: procedural with a recognition that all are analysis tools (but with no specification of what they measure);
- Relational: indicates that all are measures of center or represent what is typical for a dataset;
- Extended abstract: includes description of when each measure might be more useful than another (Groth & Bergner, 2006, p. 51)

Groth and Bergner's (2006) findings align with others in that most of the teachers were at the multistructural and relational level, indicating that the vast majority were not able to describe salient differences in measures of center. These findings are consistent with Jacobbe and Carvalho's (2011) suggestion that "research concerning teachers' understanding of average must involve more than understanding the arithmetic mean" (p. 207). Moreover, it is important to note that understanding when, for instance, the median might be more useful than the mean requires relating conceptions of center and shape (e.g., skewed, symmetric). Moreover, as we turn our attention to teachers' knowledge of variability, conceptions of center are an important component of variability.

Variability

Teachers' conceptions of variability have only recently begun to be studied. In contrast, students' conceptions of variability have been a persistently researched topic over the past 20 years. For instance, Shaughnessy (2007) presented the following list of eight different conceptions of variability that had been identified in empirical studies of students dating as far back as 1996:

1. Variability in *particular values*, including *extremes* or *outliers*

2. Variability as *change over time*
3. Variability as *whole range*
4. Variability as the *likely range* of a sample
5. Variability as *distance or difference from some fixed point*
6. Variability as the *sum of residuals*
7. Variation as *covariation or association*
8. Variation as *distribution* (pp. 984–985, emphasis in original)

In the most recent handbook chapter on statistics, Langrall and colleagues (2017) claimed that the eight conceptions listed by Shaughnessy had remained prevalent over the previous ten years, and they presented four different frameworks for the development of the conception of variability and distribution (p. 494, Figure 18.1)—as noted by Shaughnessy (2007), distribution and variation are inherently connected. One commonality across the four frameworks that Langrall and colleagues (2017) identified was that increased sophistication was characterized by "the integration of concepts and the ability to attend to and coordinate multiple aspects of the data being examined" (p. 495).

To illustrate a framework for variability, Noll and Shaughnessy's (2012) study involved ten grades 6–12 mathematics teachers and their students over a two-year period. Two task-based surveys (N = 272, 236) were administered to all students twice during the first year, and 24 students participated in two task-based interviews following the first survey and one following the second survey (during the second year). The developmental framework constructed from students' responses showed a hierarchy of reasoning progressing from additive (e.g., comparing frequencies) to proportional, which they claim

may involve a focus only on center or only on shape or only on variability, and then to the highest level, distributional reasoning, which they claim demands integrating ideas of at least two of center, shape and variability. A typical student response coded at the distributional reasoning level that was more informal in nature was "on average, the number is going to be around 7.5. However, that's just the average, meaning it's not going to happen every time. Sometimes it will be a bit higher, sometimes a little lower" (p. 528). In this response, there is informal mention of spread around a measure of center. This is in contrast to, for instance, only mentioning a measure of spread, such as the range or the standard deviation, without mention of the center itself.

Although the studies described so far have dealt with student conceptions of variability, it is reasonable to assume that teachers may have similar ways of describing variability given their expected limited background in statistics and preparation to teach it. However, there has been limited attention given to teachers' conceptions of variability. A study by Makar and Confrey (2005) involving three secondary mathematics teachers and one secondary mathematics pre-service teacher at the end of a 6-month professional development program focused, in part, on statistics content both at the middle and secondary level and extending into content traditionally found in a formal introductory tertiary course. Makar and Confrey (2005) found that when examining variability between two groups, teachers did not struggle to describe variability *within* each dataset, but struggled to describe variability *between* datasets. For instance, teachers used descriptions of shape, outliers, range, standard deviation, and more informal language such as "spread out" and "tighter" to aid in describing variation *within* a dataset (p. 368). However, when describing *between* variation, teachers mostly compared ranges and

standard deviations, and sometimes compared means or shapes. This is not evidence of higher levels of thinking about *between* variation because there was no evidence that the teachers were considering whether differences were meaningful. Moreover, these types of responses would largely be classified as *developing* according to Reid and Reading's (2008) framework constructed from tertiary student responses to tasks requiring comparing variation between two groups. Reid and Reading (2008) described a *developing* response as one that did not link *within* and *between* variation.

Another study designed to understand a hierarchy of possible levels of reasoning with variability among teachers was described by Sánchez, Silva, and Coutinho (2011). In their literature review, they described the four hierarchical levels of reasoning, paraphrased below, about variability found by Silva and Coutinho (2008):

- Level 1 (idiosyncratic): calculating mean and standard deviation without indicating knowledge of what they mean
- Level 2 (verbal reasoning): "perception of variation; understanding the standard deviation as the measure of the difference between the values of the data; the idea that low standard deviation is better and recognizing there are quantities of values within one standard deviation interval from the mean." (p. 3)
- Level 3 (transitional reasoning): using at least two summary statistics, such as "maximum and minimum values, mode and the graph representation itself", many times leaving out standard deviation and not completely integrating the ideas (p. 4)

- Level 4: (procedural reasoning): standard deviation as measure of deviation from the mean without estimating "a percentage of the observations in this interval" (p. 5)

In Silva and Coutinho's (2008) study with nine secondary mathematics teachers, the majority were identified at level two. It is noteworthy that in the section on teacher knowledge in Langrall and colleagues (2017), there were *no studies* mentioned that focused on teachers' knowledge of variability, and they referred readers to the book *Teaching statistics in school mathematics-Challenges for teaching and teacher education: A Joint ICME/IASE Study* (Batanero, Burrill, & Reading, 2011). Perhaps the lack of attention and recommendation was a result of choice of focus for the chapter. However, Sánchez, Silva, and Coutinho (2011) wrote the literature review chapter in this volume on variability, titled *Teacher's Understanding of Variation*, and cited a total of four studies explicitly involving teachers—the most relevant of which I have just summarized above. This is evidence of the incredible lack of research in this area, especially in the context of teaching, as noted by Sánchez, Silva, and Coutinho (2011) who argue, "the study of teachers' professional knowledge and teachers' practices while teaching variation is an urgent need" (p. 219).

Considering the dearth of studies on teachers' understanding of variability, merely one further study appears useful in understanding different conceptions of variability teachers might have. Peters (2011) provided a framework to describe a robust understanding of statistics. Extending from her dissertation work with 16 AP Statistics teachers who were identified as teacher leaders, Peters' framework, titled *Elements and Reasoning Indicative of Robust Understanding of Variation,* includes three perspectives

across four elements of variation. The three perspectives are *design*, *data-centric*, and *modeling*, and the four elements are described as *variational disposition*, *variability in data for contextual variables*, *variability and relationships among data and variables*, and *effects of sample size on variability* (Peters, 2011, p. 67)—see Figure 2.1.

Perspective / Element	Design Perspective	Data-centric Perspective	Modeling Perspective
Variational disposition	DP1: Acknowledging the existence of variability and the need for study design	DCP1: Anticipating reasonable variability in data	MP1: Anticipating and allowing for reasonable variability in data when using models
Variability in data for contextual variables	DP2: Using context to consider sources and types of variability to inform study design or to critique study design	DCP2: Describing and measuring variability in data for contextual variables as part of exploratory data analysis	MP2: Identifying the pattern of variability in data or the expected pattern of variability for contextual variables
Variability and relationships among data and variables	DP3: Controlling variability when designing studies or critiquing the extent to which variability was controlled in studies	DCP3: Exploring controlled and random variability to infer relationships among data and variables	MP3: Modeling controlled or random variability in data, transformed data, or sample statistics
Effects of sample size on variability	DP4: Anticipating the effects of sample size when designing a study or critiquing a study design	DCP4: Examining the effects of sample size through the creation, use, or interpretation of data-based graphical or numerical representations	MP4: Anticipating the effects of sample size on the variability of a sampling distribution

Figure 2.1. Elements and Reasoning Indicative of Robust Understanding of Variation (Peters, 2011)

Although this framework provides a useful way of understanding what kinds of knowledge teachers have, and presents a target to aim toward in teacher development courses, there is need for further research across these various frameworks in order to build a common set of expectations to work toward.

Shape of Distributions

Before reviewing the literature regarding teachers' knowledge of shape of distributions, it is important to note that much of the research regarding shape is included in studies of variability. In fact, in Langrall and colleagues (2017), the sub-heading to review literature on students' conceptions of variability is titled *Variability and*

Distributions (p. 492). This pairing is also evident in Shaughnessy's (2007) list of conceptions of variability as the eighth conception is "variability as distribution" (p. 985). In an effort to not repeat information in the previous section, this section will attempt to focus on studies that more explicitly conceptualize shape, as opposed to studies in the previous section that described it as an important feature more broadly.

As a concept integral to thinking about distribution, shape tends to involve integration of center and spread (Arthur Bakker & Gravemeijer, 2004; Pfannkuch & Reading, 2006; Reading & Canada, 2011). For instance, a symmetric distribution with a high peak might be described as "tight", indicating a narrow spread as well as a general location of multiple measures of center. For this reason, ideas of distribution cannot easily be isolated. This is one of the reasons that the editorial for the special issue of *Statistics Education Research Journal* dedicated to *Reasoning about Distribution*, was titled *Reasoning about Distribution: A Complex Process* (Pfannkuch & Reading, 2006). Moreover, the notion of shape is just one of many conceptions included in the literature on distribution, and is widely accepted to involve "features such as centre, spread, density, skewness, and outliers but also involves other ideas such as sampling, population, causality and chance" (p. 4), and thus distribution has been referred to as being located within a "web of statistical knowledge" (Reading & Canada, 2011, p. 224). However, some ideas relating to shape (as well as center and spread) from the literature on teachers' conceptions of distribution are helpful to inform this study.

One common misunderstanding students and teachers appear to have is that skewed distributions will be more variable than symmetric distributions (delMas & Liu, 2005; Doerr & Jacob, 2011). For example, Doerr and Jacob (2011) studied a group of

secondary practicing and prospective mathematics teachers, as well as mathematics education doctoral students. Some common misunderstandings they claimed teachers exhibited regarding variability and shape were:

1. not attending to how skewness impacts standard deviation when comparing two skewed distributions in which one is more strongly skewed,
2. judging symmetric distributions to have a smaller standard deviation than a non-symmetric distribution with a much smaller range and similar mean, and
3. more clearly indicating that symmetry "minimizes standard deviation" (pp. 780–781).

These findings led Doerr and Jacob to conclude that "some teachers had difficulty interpreting skewed distributions and tended to inappropriately choose symmetric normal distributions" (p. 784).

Summary of Teachers' Knowledge

After reviewing the literature relating to teachers' conceptions of center, spread, and shape, a couple of important ideas should be made explicit. First, all three knowledge element types (center, spread, shape) cannot exist in isolation from another. The various frameworks that exist clearly indicate that in order to develop higher levels of cognition, it is required that multiple ideas are integrated together. Therefore, it is important that research be conducted that attempts to account for the integrated knowledge that teachers have when investigating how their statistics knowledge might be used, for example, to make inferences or in planning and teaching contexts. Second, the studies reviewed in the previous sections on teachers' knowledge were largely focused only on teachers' knowledge. There were few attempts to connect teachers' knowledge and ways of reasoning about center, variability, and shape to how that knowledge supported their

inferential reasoning. Moreover, almost none of these studies attempted to connect to teachers' statistics pedagogical content knowledge, and Langrall and colleagues (2017) noted this as a "critical area for future research" (p. 517). This study attempts to address both of these needs.

Measuring and Mapping Knowledge Structures

A particularly useful way of measuring knowledge is through examining knowledge structures. First, prior literature has shown that knowledge of center, spread, and shape are inherently integrated. Second, my theoretical stance on knowledge and learning encourages paying attention to knowledge structures individuals may have, regardless of whether I believe they are correct. Therefore, it is vital that teachers' knowledge structures are examined as they are observed to exist.

In the field of mathematics education, the idea of structures of knowledge (also called *cognitive structures*, *mental structures*, and *mental schemes*) goes back to Piaget's work with elementary level students (1964/1997). Piaget's methods that stemmed from his theoretical perspective of *constructivism* led to many important contributions in the field of mathematics education. For instance, considering students' conceptions of fractions by analyzing their cognitive structures has led to great advances in the field in terms of what is known about how students construct and develop fractional reasoning (e.g., L. Steffe, 2001). In turn, this understanding, paired with the same method of analyzing cognitive structures, has led to greater understanding of teachers' mathematical knowledge for teaching related to fractions (e.g., Izsák, 2008).

As noted by Izsák (2008), identifying teachers' knowledge structures is not a simple task. For instance, a teacher may not reveal knowledge he or she has constructed, even in an interview setting, because it may seem less relevant than some other

knowledge at the time. Therefore, just because a knowledge element is not observed, does not mean it does not exist. Thus, an observed knowledge structure can never represent a person's true knowledge structure. It is for this reason that the literature review now takes a slight turn.

Motivated by the notion that knowledge structures can never be complete, I began to see other ways knowledge structures have been examined. One concept that has received much attention over the past 25 years is Smith, diSessa, and Roschelle's (1993) *Knowledge in Pieces*. This construct perceives, literally, that knowledge is made up of pieces—some of which may be viewed as misconceptions. However, the central notion of *Knowledge in Pieces* is that all pieces of knowledge are used in the construction of other pieces of knowledge and should therefore not be disregarded. Placed within the context of the basics of *constructivism*, new knowledge is believed to occur through the process of adaptation and assimilation (Piaget, 1964/1997).Therefore, Smith, diSessa, and Roschelle (1993) posited that knowledge pieces currently perceived as a misconception may yet be the knowledge that supports a later perceived correct conception, thus turning that misconception into a preconception.

Building somewhat on the principles of *Knowledge in Pieces*, Ron, Dreyfus, and Hershkowitz (2010) created the perspective of *partially-correct constructions* (PaCCs). Ron, Dreyfus and Hershkowitz were confronted with a situation in which student participants in their study were exhibiting unexpected contradictions in their observed knowledge. For instance, students were coming to the same correct response, but getting there in contradictory ways, and across time, students they had believed to have constructed the intended knowledge were later observed to have constructed

contradictory knowledge. They defined a *partially-correct construct* as "a construct that only partially matches a corresponding intended knowledge element" (p. 69). In this way, they identified all possible constructions and used a content analysis to identify *intended* knowledge elements from which to decide if an observed knowledge element was constructed in the intended way or not. Moreover, these knowledge elements are intended to have observed connections between them—in part because the original purpose was to trace learning over time and thus understand the process of knowledge construction.

Because *intended* knowledge elements were agreed upon a priori, and connections between elements were recorded, Ron, Dreyfus, and Hershkowitz (2010) identified three different types of PaCCs: *missing element*, *incompatible element*, and *disconnected element*. A *missing element* PaCC was observed to no include one or more *intended* knowledge elements. An *incompatible element* PaCC was observed to have at least one non-intended knowledge element that contradicts an *intended* element. A *disconnected element* PaCC was observed to have an *intended* element that was not observed to be connected to, or contribute to, any other *intended* element.

In the context of studying elementary preservice teachers' statistical knowledge for teaching nominal categorical data analysis, Groth and Bergner (2013) used this method of mapping knowledge structures in terms of PaCCs and made some enhancements to it. First, they formed the structures into visual node-link diagrams (see Chapter 3 for examples). Second, they identified a fourth PaCC called *missing and incompatible element* in which both types of knowledge elements were observed in the knowledge structure.

Through the identification of PaCCs, these researchers were able to more deeply understand how participants' knowledge may have been constructed through identification of links among elements that may be contradictory—which Groth and Bergner claim "may be difficult to detect with quantitative instruments" (p. 262). Moreover, such identification allows for a finer-grained analysis, and understanding, of teachers' knowledge (Groth & Bergner, 2013; Ron et al., 2010).

Although this study is not attempting to trace the construction of pre-determined knowledge over time, these studies provide a useful way of conceptualizing how knowledge structures might be recorded and analyzed. The next phase turns away from thinking about teachers' knowledge (RQ1), and turns attention to *informal inferential reasoning* (IIR).

Teachers' Informal Inferential Reasoning

The Value of Informal Inferential Reasoning

The importance of informal knowledge and informal reasoning. As previously mentioned, Zieffler and colleagues (2008) make a strong case for IIR as an important pre-cursor for developing the formal inferential reasoning that both the *GAISE* report and the *CCSSM* claim as end goals of students' high school careers. Zieffler and colleagues' argument rests on their findings from a review of the literature on both informal knowledge and informal reasoning. Their most significant findings come first from Smith, diSessa, and Rochelle (1993), who said that incorrect informal knowledge of novices should not be disregarded because it is similar in many ways to the knowledge of experts but has merely been used in an incorrect manner.

To illustrate this difference in how knowledge might be used incorrectly, in Smith and colleagues' (1993) study of expert and novice reasoning with comparing fractions,

novices would incorrectly claim that the larger fraction was the one with the smallest denominator. However, "their strategy was conceptually correct if incomplete (smaller denominators do indicate larger parts and therefore tend to increase fraction size), but they had not restricted its application to equal-numerator situations" as the experts had (p. 137). Therefore, it seems that informal knowledge is important to build on if this is, in fact, the case. Zieffler and colleagues provided more evidence by drawing on Gravemeijer and Doorman (1999) who claimed that informal knowledge is important for students to build on to recreate formal knowledge and further advised that "students should be given the opportunity to ground their understanding in their own informal knowledge (p. 115). Thus, informal knowledge can be built upon in order to gain access to formal knowledge.

Regarding informal reasoning, Zieffler and colleagues (2008) drew frequently on Perkins and colleagues (1991), among others, who found that increases in content knowledge, age or maturity, and motivation does not improve informal reasoning, but it does improve with instruction. Because reasoning from evidence is an important component of informal reasoning and for making decisions within statistics, informal reasoning is an important component of informal inferential reasoning that helps build towards the ultimate goal of formal inferential reasoning.

Widespread issues with formal inferential reasoning. One might argue at this point that formal inferential reasoning can be taught directly and that developing informal reasoning may not be necessary. However, looking back across at least the past 20 years has shown a consistent pattern of students not making significant gains in their understanding of statistical concepts or in their reasoning abilities (e.g., Castro Sotos et

al., 2007; delMas et al., 2007). Moreover, researchers have made several arguments about why students struggle to understand formal statistical concepts and to successfully reason with them. Possible matters of concern include lack of confidence (Castro Sotos et al., 2009), the logic of hypothesis testing being inconsistent with the logic used in mathematics (Castro Sotos et al., 2007; Falk & Greenbaum, 1995), and both teachers (Haller & Krauss, 2002) and textbooks (Falk & Greenbaum, 1995) passing along incorrect information, or placing too much emphasis on computations (delMas et al., 2007). Regardless of what or whom is to blame for the vast array of issues, of which only a few of the most common are mentioned here, it is obvious that there are major issues, globally, in the teaching and learning of formal statistics and that a change is needed to address them.

Informal inferential reasoning as a path to success. One way to address the issue is to draw on the research by Smith and colleagues (1993), Gravemeijer and Doorman (1999), and Zieffler and colleagues (2008) and place initial emphasis in the learning of statistics on informal knowledge and informal reasoning to aid in the development of IIR. Although the notion of IIR is fairly new, and making its way into state standards is even more recent, some success has been found in translating IIR into formal inferential reasoning.

Evidence of the most success is found in Garfield, delMas and Zieffler's (2012) study that used design principles to build an introductory statistics course at the post-secondary level that used the National Science Foundation funded curriculum project called *Change Agents for Teaching and Learning Statistics* (CATALST). The course was designed with a focus on statistical reasoning and modeling, and incorporated a

summative assessment adapted from the *Comprehensive Assessment of Outcomes in Statistics* (CAOS) assessment (delMas et al., 2007) called the *Goals and Outcomes Associated with Learning Statistics* (GOALS) assessment, as well as one designed to model students' thinking. Furthermore, classroom discourse and testing conjectures were important components of the course (Garfield et al., 2012). They found that students could not only reason as well or better than students in a traditional setting, but that most students had positive affect toward statistics and the course, despite it being radically different than the norm (Garfield et al., 2012). It is noteworthy that despite students performing weakly on items referring to interpretation of *p*-value (a commonly misunderstood topic), they still out-performed students in a traditional setting.

Success with IIR has been documented in studies not intending to develop formal reasoning among elementary students (e.g., Makar, 2014; Paparistodemou & Meletiou-Mavrotheris, 2008), middle level students (e.g., Gil & Ben-Zvi, 2011; Lavigne & Lajoie, 2007; Stohl Lee et al., 2010), secondary students (e.g., Dierdorp, Bakker, van Maanen, & Eijkelhof, 2012; Pfannkuch, 2011), and preservice and inservice teachers (e.g., Dolor & Noll, 2015; A. M. Leavy, 2010). These successes provide further evidence that a focus on building on students' informal knowledge and informal reasoning can be successful for developing their *informal inferential reasoning* in hopes that it will allow students to overcome the historical plague of failure in developing formal inferential reasoning. Because the formal inferential reasoning that makes use of tools such as hypothesis testing and confidence intervals is prevalent in the larger global society (Castro Sotos et al., 2007), working towards the formal inferential reasoning required to appropriately use

those tools is an important goal of IIR, of which there is a wide range of evidence of success.

Teachers and Informal Inferential Reasoning Contexts

The literature on teachers engaged in *informal inferential reasoning* (IIR) is scarce. Moreover, the majority of the literature that does exist is focused on preservice teachers rather than practicing teachers. However, this literature is still relevant given the evidence that practicing teachers likely do not have any more background in statistics (perhaps even less) than preservice teachers. This section will document a few of the important findings from these studies.

In the context of a prospective elementary mathematics methods course that Leavy (2010) taught, she included learning goals targeted at developing statistics content, engaging in IIR, and planning and enacting a lesson designed for IIR (the latter part will be discussed in a later section). Leavy found that the elementary prospective teachers struggled to connect the ideas of IIR with their prior formal experiences with statistics, and they struggled to understand the meaning of IIR from the two-week unit they spent developing the ideas.

In a similar way, within the context of content courses designed for preservice and inservice teachers, Dolor and Noll also found teachers to struggle with IIR, despite largely being successful. During the 10-week non-introductory statistics course, taught by the researchers, Dolor and Noll (2015) engaged students in developing an informal hypothesis test for categorical data and making informal inferences along the way. They found that students were successful in reinventing a hypothesis test, although analysis of class discussions leading up to that point indicated a large amount of struggle due to students' lack of experience with simulations (a major component of the course).

Perhaps corroborating preservice teachers' struggle to improve in their reasoning, Huey (2011) studied the development and transition of middle and secondary preservice teachers' reasoning from informal to formal inferential reasoning across the semester. It is of note that although the course did include some tasks related to IIR, the course was not intentionally designed to develop IIR. Huey gave students a set of tasks requiring IIR as a pre-test, with the post-test involving formal inferential reasoning, and carried out mid-semester task-based interviews focused on IIR. She found that students generally improved in their IIR abilities, but most did not progress beyond using a single measure of center plus a global characteristic such as "range, spread and variance" (p. 165). Moreover, about a third of the students stayed at the lowest level of reasoning, consistently only drawing on measures of center for making inferential statements. This finding is striking because this occurred despite the inherent opportunity to learn to use measures of center, spread, and shape within the context of informal inference.

Relating more to the context of teaching, one final relevant study focused on a single secondary teacher of a non-AP statistics course. Pfannkuch (2006a) studied this teacher during teaching instances in which boxplots were being compared and informal inferences were being made. She found, similar to Huey (2011), that the teacher felt more comfortable with discussing measures of center and struggled when attempting to draw on concepts of variability to support inferential statements. In fact, Pfannkuch (2006a) stated that it wasn't clear to the students "how comparing spreads helped in making an inference" (p. 41). Another major finding was that among a set of reasoning elements, Pfannkuch claimed that the evaluative element (weighing the evidence) was critical for making an informal inference and moderated all other reasoning elements. She further

claimed “qualitative judgments on the whole must be made to ascertain whether one is prepared or not prepared to state Group A is greater than Group B, on average” (p. 42).

Although there is not much research on teachers’ engagement in IIR, the little that exists aligns with studies of teacher knowledge that indicate that in general, teachers struggle to move into higher levels of reasoning that require integrating multiple conceptions of center, spread, and shape when making an informal inference. Moreover, as described by Huey (2011), more coursework in statistics does not necessarily lead to development of higher levels of reasoning.

Measuring IIR

A few different frameworks have been used to study *informal inferential reasoning* (IIR). Zieffler and colleagues’ (2008) framework (described earlier) was explicitly designed for research on IIR and considers it as a process involving (a) making predictions, (b) drawing on prior knowledge, and (c) making a judgment based on evidence (p. 45). In a similar way, Makar and Rubin (2009) described a framework that includes the components of (a) generalizing beyond the data, (b) using data as evidence, and (c) using probabilistic language recognizing a degree of uncertainty in the inference (p. 85). Note that these two frameworks contain fairly similar features except that Makar and Rubin include attention to language that indicates a non-deterministic statement.

A foundation of the basic structure of IIR leads to a discussion of the types of tasks that are needed in order to elicit engagement in IIR. Zieffler and colleagues provided some perspective on this by including examples of three types of tasks they claimed aligned with their framework:

1. Estimate and draw a graph of a population based on a sample;

2. Compare two or more samples of data to infer whether there is a real difference between the populations from which they were sampled, and
3. Judge which of two competing models or statements is more likely to be true. (p. 47)

Given that Zieffler and colleagues' (2008) framework built on the work of Means and Voss' (1996) work with informal reasoning more generally, it is instructive to note that they claimed it was important that tasks designed to elicit informal reasoning be ill-structured and open-ended. Building on all of these frameworks, as well as Means and Voss' work, Huey and Jackson (2015) constructed a task framework claiming that a task designed for IIR should: (a) necessitate an inference beyond the data, (b) be ill-structured, (c) be open-ended, (d) require the consideration of context, and (d) contain or provide the means to create a visual representation (p. 432). Therefore, as described more in the Method section, I drew on this framework when choosing tasks for this study.

Pedagogical Content Knowledge

Mathematical Knowledge for Teaching

Since the mid-1980's, there has been a wealth of research on teachers' content knowledge, and what has become known as *pedagogical content knowledge* (PCK) (Shulman, 1986). When Shulman first introduced his framework, it included "(a) subject matter content knowledge, (b) pedagogical content knowledge, and (c) curricular knowledge" (p. 9). However, this framework was designed to cut across content domains. Building on Shulman's framework, Hill, Ball, and Schilling (2008) constructed a framework for *Mathematical Knowledge for Teaching* (see Figure 2.2) that contained three sub-categories of PCK: (a) knowledge of content and students, (b) knowledge of content and teaching, and (c), knowledge of curriculum (p. 377).

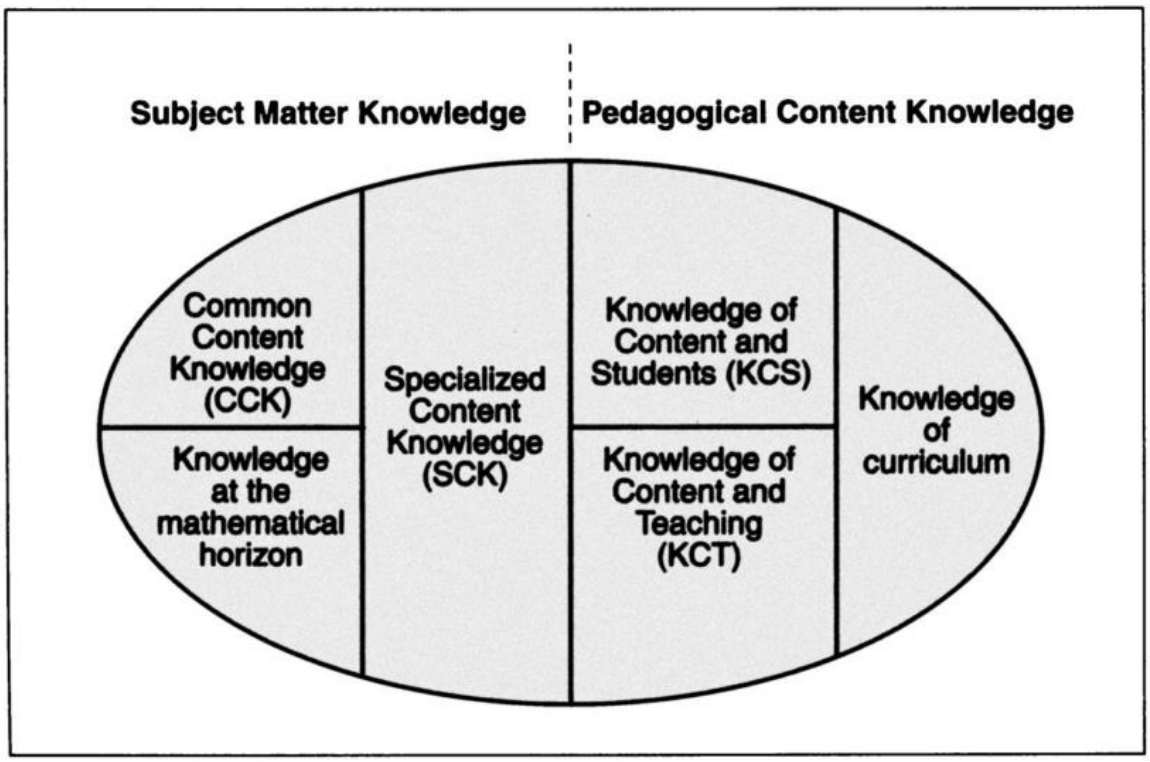


Figure 2.2. Mathematical Knowledge for Teaching Framework (Hill et al., 2008).

In Hill and colleagues' framework, knowledge of content and teaching was meant to include knowledge of teaching moves, while knowledge of content and students was meant to include knowledge of students' thinking.

Since this time, and including the work of Hill and colleagues, the majority of research regarding MKT and PCK has been done at the elementary level. For instance, in a related study, Hill and colleagues (2008) designed a widely-used instrument to measure MKT and its sub-categories, including PCK, at the elementary level. No such instrument exists for grades 6–12, and there is a current debate at the high school and undergraduate levels about how well Hill and colleagues MKT framework holds (Speer, King, & Howell, 2015). Narrowing the focus to statistics, PCK is a much less researched area, despite being identified as a critical research need a decade ago (Shaughnessy, 2007), and

it remains in critical need (Langrall et al., 2017). However, there has been some recent progress in this area.

Statistical Knowledge for Teaching

Because statistics is different from mathematics in meaningful ways (Ben-Zvi & Garfield, 2005b; Franklin et al., 2007), it is important for the field of statistics education to incorporate these differences into a framework for *Statistical Knowledge for Teaching* (SKT). Randall Groth (2007) conceptualized such a framework (see Figure 2.3), drawing on the work of Hill and colleagues (2004) as well as the GAISE report (Franklin et al., 2007). However, Groth's (2007) model only included ideas from the subject matter knowledge portion of Hill and colleagues' (2004) conceptualization and introduced a way of thinking about how mathematics and statistics content knowledge and specialized content knowledge may interact. He later added pedagogical content knowledge (see Figure 2.4), maintaining the three broad categories from Hill and colleagues (2008) but adding specification to indicate possible unique descriptions, or traits as he called them, to statistics (Groth, 2013). In this updated framework, both mathematics and statistics subject matter content knowledge and pedagogical content knowledge are necessary—as indicated in his 2007 version.

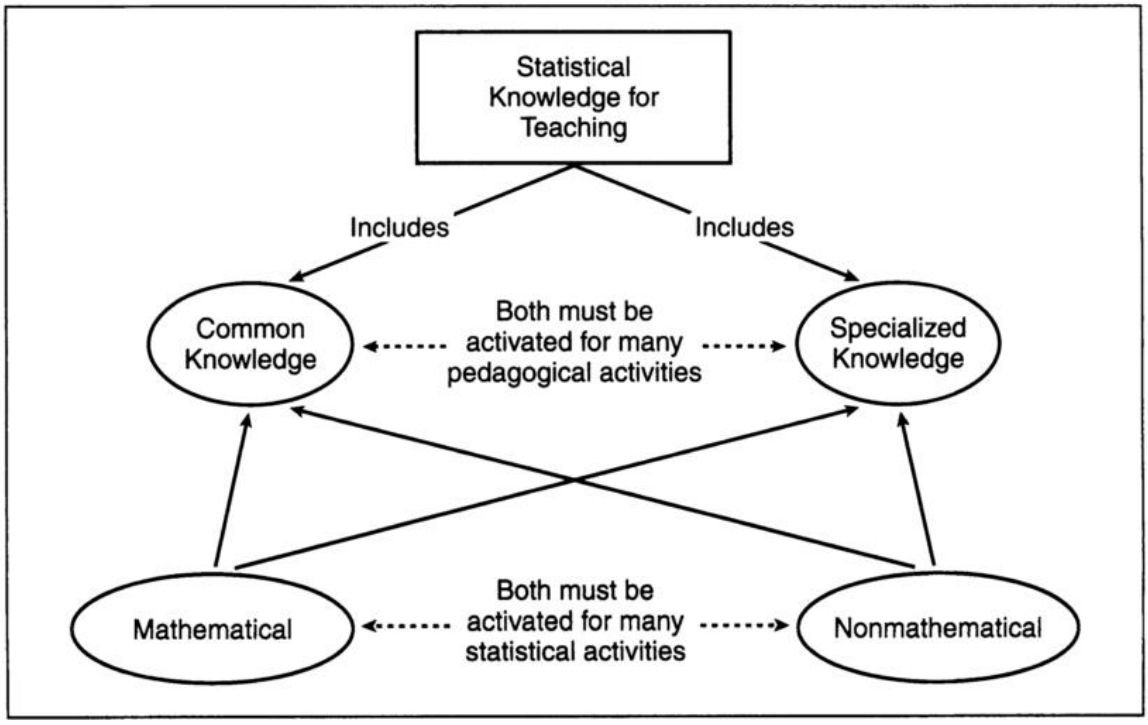


Figure 2.3. Statistical Knowledge for Teaching Framework (Groth, 2007)

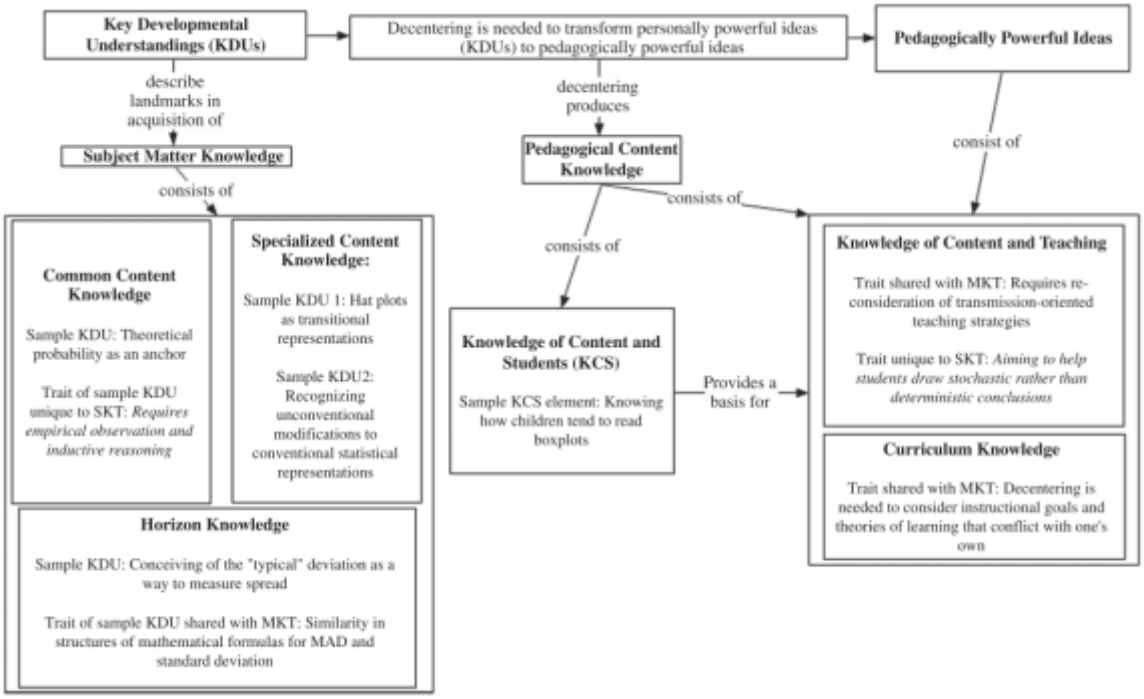


Figure 2.4. Statistical Knowledge for Teaching Framework (Groth, 2013)

In an attempt to distinguish between different types of PCK for statistics, Watson, Callingham, and Donne (2008) conducted a study with a group of elementary teachers. For this study, their primary main aim was to measure PCK in terms of "teachers' [statistics] content knowledge, its reflection in knowledge of their students' content knowledge, and their PCK in using student responses to devise teaching intervention" (Watson et al., 2008, p. 1). Note that although there was some analysis that compared teachers' responses to that of a sample of student responses, I report here only about the teachers' responses. Teachers responded to task-based surveys that included tasks designed for elementary level students. For each task, they were asked to (1) generate some possible appropriate and inappropriate student responses and (2) to describe how they might intervene with a student who provided one of the inappropriate responses. They were also provided sample student responses to respond to. Quantitative analysis of the responses indicated three hierarchical levels of statistics pedagogical content knowledge. At the *low* level, teachers were described as having been "partially successful" at suggesting student responses and generally having no success responding to the provided student responses (Watson et al., 2008, p. 4). At the *middle* level, teachers were able to suggest both appropriate and inappropriate student response and were able to suggest single, generic ideas for intervention. At the *high* level, teachers were able to suggestion both appropriate and inappropriate student responses and struggled to suggest ideas for intervention, but their suggestions related specifically to the content.

Three years later, they repeated the same design, again with elementary teachers, and were able to distinguish four levels of statistics PCK (Callingham & Watson, 2011). At the *Aware* level, teachers were able to generate either inappropriate or appropriate

student responses, and suggestions for intervention did not address student misunderstandings. At the *Emerging* level, teachers were able to generate multiple inappropriate or appropriate student responses, and suggestions for intervention were generic and not content specific. At the *Competent* level, both inappropriate and appropriate student responses were generated, and "some statistically appropriate interventions [were] suggested but only in the context of familiar activities" (Callingham & Watson, 2011, p. 290). At the *Accomplished* level, both inappropriate and appropriate student responses were generated, and teachers exhibited an "integration of appropriate statistical content with student-centered intervention strategies" (Callingham & Watson, 2011, p. 290).

A couple of other studies have also attempted to measure teachers' statistics PCK that are worth mentioning here. When working with elementary preservice teachers in the context of a mathematics methods course, some evidence exists that correct knowledge structures (including both content and PCK) do not necessarily lead to successful intervention with students (Groth & Bergner, 2013), and students teaching IIR-focused lessons struggle to direct students' attention toward using data as evidence, yet engagement in lesson study generally improved their PCK (A. M. Leavy, 2010).

Studies of teachers' PCK for statistics at the secondary level are essentially non-existent, thus a decade of calls for research in this area have gone unmet and are of critical need (Langrall et al., 2017; Shaughnessy, 2007). This study attempts to address this call by examining possible relations between content, IIR, and statistics PCK (RQ3).

Summary

The previous sections reviewed the literature regarding teachers' statistical knowledge related to center, spread, and shape of distributions, teachers' informal

inferential reasoning (IIR), and their pedagogical content knowledge (PCK) for statistics. This review has shown that 1) there is an extreme lack of research in all three topic areas that focuses on teachers, 2) research on teachers' knowledge that draws connections between knowledge elements mostly focuses on different levels of correct, or *desirable*, conceptions, and 3) the research that does exist largely does not attempt to describe possible connections between knowledge, IIR, and PCK. This study is an attempt to address each of these three gaps in the literature through 1) examining teachers' knowledge structures (as they are observed to exist), 2) characterizing possible ways that their knowledge structures may support their IIR, and 3) how teachers' PCK for statistics may be related to their knowledge structures and IIR.

CHAPTER 3: RESEARCH DESIGN AND METHOD

In this chapter, I address the research design and method in three sections. In the first section, I provide my background in statistics and teaching statistics and discuss the participants, data sources, and data collection. In the second section on data analysis, I provide information on the: (a) method used to map knowledge structures, (b) framework for characterizing informal inferential reasoning (IIR), (c) framework used to distinguish levels of statistics PCK, and (d) validity and reliability of both the data sources and the data analysis. In the final section, I provide a summary of the chapter.

Research Design

To answer the research questions for this study, I began by recruiting a stratified purposeful sample (Patton, 2002) in order to maximize the various settings in which statistics is taught by middle level and secondary mathematics teachers. A task-based clinical interview design (Goldin, 1997) was then used to collect data, followed by a cross-case analysis (e.g., Groth & Bergner, 2013) to identify patterns of knowledge structures (RQ1), knowledge structures as support for IIR (RQ2), and possible relationships between knowledge structures, IIR, and PCK (RQ3). Analysis of teachers' PCK involved coding teachers' responses from an a priori coding framework (Callingham & Watson, 2011; Watson & Callingham, 2014). The next sections will provide depth on how each of these pieces of the design were enacted.

Research Method

Researcher Background

This study involved a task-based clinical interview design (Goldin, 1997) and decades of research have shown that teachers and students alike have widespread misconceptions about statistics (e.g., Castro Sotos et al., 2007) and that completing a

tertiary course in statistics does not necessarily mean that any sound statistical reasoning was developed (delMas et al., 2007). Therefore, it is important that I provide the extent of my background in both learning and teaching statistics to provide a sense of faith that I have the necessary skills to 1) engage practicing teachers in deep discussion during interviews around statistics content and the teaching of that content, and 2) carry out the qualitative data analysis of teachers' responses.

My training in statistics has consisted of two tertiary courses in statistics, one of which was calculus based, one Advanced Placement (AP) summer institute offered by The College Board that included four days of intense training, and nine graduate-level courses in educational statistics and quantitative research methods including Multivariate Analysis, Structural Equation Modeling, and Hierarchical Linear Modeling. I have a Bachelor's degree in Mathematics and a Master of Arts degree in Teaching Secondary Mathematics. Prior to entering my doctoral program, I taught Advanced Placement (AP) Statistics to 11^{th} and 12^{th} grade students for two years. During this time, I continued my professional development for statistics by attending specific sessions designed for AP Statistics teachers at two *Teaching Teachers with Technology (T^3) International Conferences* and by participating in conversations on the AP Statistics Teacher Community listserv hosted by The College Board. Regarding my students' performance, 75–80% of my students earned a score of at least a 3 on the AP Exam, which is the minimum required by most institutions of higher education to earn college credit for an introductory statistics course. Moreover, during my doctoral program, I taught an undergraduate mathematics methods course for prospective middle school teachers that included a focus on statistics.

Participants

Selection. Recruitment was targeted at mathematics teachers across all grades spanning 6–12 in order to obtain a stratified purposeful sample (Patton, 2002) across four strata: AP Statistics, non-AP Statistics, a unit of statistics in a middle level mathematics course, and unit of statistics within a secondary level mathematics course. More specifically, teachers were recruited who had taught statistics content that explicitly involved data analysis, such as calculating summary statistics, to make decisions about data. Recruitment efforts were made through mathematics department heads at middle level and secondary schools in a large school district in the Midwest. Individual teachers were then contacted directly. Due to one teacher dropping out of the study from a lack of time to participate in the interviews, an additional participant was added who was recommended by a participant in an initial pilot study.

Background. Recruitment resulted in nine middle level and secondary teachers of mathematics. Their backgrounds in teaching statistics are described in Table 3.1. As can be seen in the table, the sample was uniformly distributed across the four targeted strata, with at least 2 teachers in each stratum. All participants had completed a Bachelor's degree in Mathematics or Mathematics Education, a Master's degree in an education related field (e.g., Curriculum and Instruction, Secondary Education), and had completed at least one tertiary course in statistics.

Table 3.1

Study Participant Backgrounds

Teacher	Course Statistics Taught In Most Recently	Grade Level of Course	Years Teaching Mathematics	Years Teaching Statistics	Undergraduate Statistics Courses	Graduate Statistics Courses	Statistics Specific PD
Kathy	6th Grade Mathematics	6	4	4	1	1	Several
Ruby	Pre-Algebra & Algebra I	8	10	10	2	0	None
Amalia	Algebra I	9	4	3	1	1	None
Harrison	Pre-Algebra & Algebra I	8-9	12	5	1	0	None
Ellie	Pre-Algebra & Honors Geometry	8	3	3	2	0	None
Machaela	Non-AP Statistics	12	9	1 (3 as unit)	1	0	None
Mike	AP Statistics	11-12	13	2 (7 as unit)	2	1	Once
Rosalynn	AP Statistics	10-12	11	4 (7 as unit)	1	10	Extensive
Tim	Non-AP Statistics	12	14	5 (10 as unit)	1	0	None

Background in statistics content was relatively similar across participants, with three teachers having one graduate-level statistics course and one teacher, Rosalynn, having a graduate certificate in educational statistics. Although both AP Statistics teachers reported having attended professional development specifically related to statistics, most teachers in this study reported having no professional development for statistics.

Across the sample, there was considerable variability in mathematics teaching experience, ranging from three to 14 years, and teachers of stand-alone statistics courses had more experience on average than teachers who taught statistics as a component of a yearlong mathematics curriculum. However, there was much less variability in statistics teaching experience, ranging from three to 10 years teaching statistics as a unit, and among teachers of stand-alone statistics courses, experience teaching such a course ranged from one to five years. Moreover, neither the most nor the least experienced at teaching statistics were clustered at a particular grade band or stratum.

Data Sources

Three data sources were used to develop knowledge structures for each teacher regarding center, spread, and shape of distributions, to describe teachers' informal inferential reasoning (IIR) and how their knowledge structures provide support for their IIR, and to describe relationships between teachers' IIR and their knowledge of students as a sub-domain of pedagogical content knowledge (PCK). These data sources are listed in Table 3.2 and are mapped to the corresponding research question(s) they were intended to support. Over the next few sections, I describe each of the data sources in detail.

Table 3.2
Data Sources

Data Source	Research Question
Written responses to Goals and Outcomes Associated with Learning Statistics (GOALS-2) assessment (Sabbag & Zieffler, 2015)	RQ1: Knowledge Structures
Written/verbal responses to Levels of Conceptual Understanding in Statistics tasks (Jacobbe, 2016) - 4 constructed response released tasks	RQ1 & RQ2: Knowledge Structures and IIR
Written/verbal responses to Supplemental questions per LOCUS task: - What are some appropriate and inappropriate student responses? - How might you respond to a student who offers an inappropriate response? (Watson, Callingham, & Donne, 2008)	RQ1, RQ2 & RQ3 Knowledge Structures, IIR & PCK

GOALS-2 Assessment. The GOALS-2 assessment (Sabbag & Zieffler, 2015) was developed with funding from the National Science Foundation (NSF) to assess the effectiveness of the NSF-funded Change Agents for Teaching and Learning Statistics (CATALST) curriculum, which was designed for a tertiary statistics course (Garfield et al., 2012). It was developed from items from the Comprehensive Assessment of Outcomes in Statistics (CAOS) assessment (delMas et al., 2007) and is designed to assess students' statistical reasoning after completing a first course in statistics at the undergraduate level. GOALS-2 includes 20 forced-choice items related to covariation, samples and sampling, hypothesis testing (e.g., confidence intervals, *p*-values), and study design. Therefore, because this assessment largely measures formal knowledge, it was selected as a measure of participants' background experience with formal statistics and to aid in confirming some of the knowledge elements (RQ1) found during analysis of other data sources. Participants completed the GOALS-2 online prior to task-based clinical interviews and I retrieved item-level data of their responses afterward.

LOCUS Tasks. To aid in developing maps of knowledge structures and in describing teachers' IIR, four constructed response released items from the *Levels of Conceptual Understanding in Statistics* (LOCUS) assessment (Jacobbe, 2016) were

selected. Developed through NSF funding, the LOCUS assessment is designed to assess middle and secondary level students' conceptual understanding of statistics (Whitaker, Foti, & Jacobbe, 2015). More specifically, items were designed to align with 1) the *Common Core State Standards for Mathematics* (CCSS-M) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), and 2) the three hierarchical developmental levels across the four statistical problem solving process components described in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) *Report* (Franklin et al., 2007), which has been officially endorsed by the *American Statistics Association* (ASA) and was funded by the ASA/NCTM Joint Committee. Moreover, the items included on the LOCUS assessment, including those that are released items, were thoroughly vetted by experts in the field, including multiple authors of the GAISE report, and were given final acceptance by the ASA/NCTM Joint Committee on Curriculum in Statistics and Probability, as well as the NSF project's advisory board and two test development committees (Jacobbe, Case, Whitaker, & Foti, 2014).

The LOCUS assessment contains both forced-choice and constructed-response items, and because RQ2 is focused on inference, the four selected items were categorized under the statistical problem solving process components of *data analysis* and *interpreting results*, as outlined in the GAISE report. Similarly, I drew on Huey and Jackson's (2015) IIR task framework. With this framework, Huey and Jackson argue that IIR tasks should be ill-structured and open-ended, echoing the recommendations of Means and Voss (1996) for encouraging informal reasoning more generally, and should be embedded within a context that is necessary in order to respond to the task. Therefore,

all four selected items were constructed-response tasks that contained these features. Additionally, all four items require knowledge elements related to measures of center, spread, or shape, and either ask for an inferential statement beyond the data or ask for a critique of an inferential statement beyond the data—as also suggested by Huey and Jackson (2015). One final feature from Huey and Jackson's framework that was included in three of the four tasks was the use of a visual representation of the data to shift participants' "thinking away from local attributes or summary statistics towards global characteristics and relationships" (p. 435).

To illustrate the presence of these task features, the New Year's Day Race task (see Figure 3.1) prompted participants: "Jaron predicted that the mile times of runners in the 5K race would be more consistent than the mile times of runners in the half-marathon. Do these data support Jaron's statement? Explain why or why not" (Jacobbe, 2016). This task is ill-structured because it does not dictate a particular solution method, it is open-ended because it allows for multiple types of responses, it includes a visual representation of the data, and it requires some type of knowledge of measures of spread—and with the visual representation, there is no need for knowledge of more formal measures such as standard deviation.

A later sub-part of this task prompted participants: "Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K?" (Jacobbe, 2016). This question retains the ill-structured and open-ended features of the task, and also requires a critique of an *inferential* statement—to critique the study design as limiting the possibility of such an inferential statement because runners were able to choose their race. Moreover, as

recommended by Huey and Jackson (2015) the context of the data is an essential component to responding to this item because without understanding who the data represents and how the data was collected from them, it is not possible to critique such an *inferential* statement. The four tasks selected were: New Year's Day Race, Tomatoes and Fertilizer, Extended School Day, and Jumping Distance (Jacobbe, 2016; See Appendix F for all four tasks).

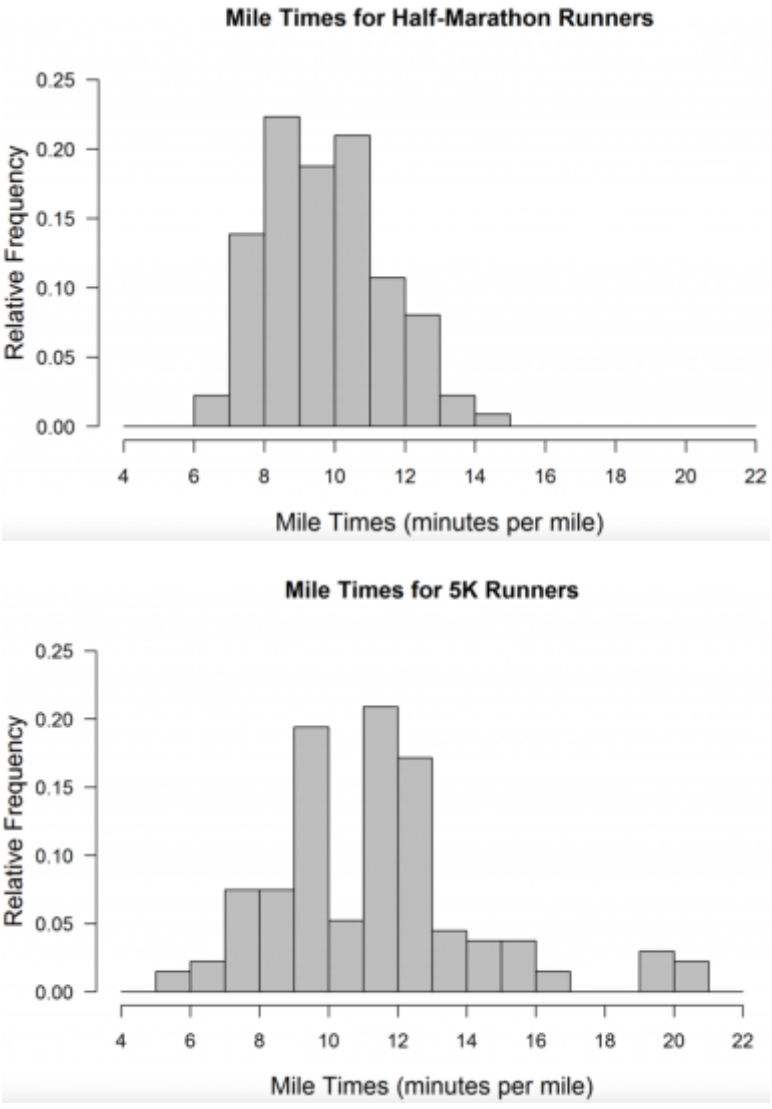


Figure 3.1. Graphs presented in the New Year's Day Race task. Released item from LOCUS assessment (Jacobbe, 2016)

In order to address the aims of the study, some of the selected LOCUS tasks were adapted by removing a sub-question(s). On the Tomatoes and Fertilizer task, part (a) was removed because this study was not attempting to analyze teachers' knowledge of randomization methods and part (c) was removed because it required *formal* inferential reasoning and this study only aims to describe teachers' *informal* inferential reasoning. On the Extended School Day task, the stem of part (a) was included in the main stem of the task, and the question itself was removed and added as a contingency to the interview protocol. All participants made mention of the content of part (a) without having to prompt for it. On the Jumping Distance task, part (a) was removed because this study was not attempting to analyze teachers' knowledge of randomization methods.

Supplemental questions about students per LOCUS task. To aid in answering RQ3 regarding teachers' knowledge of students, I draw on Callingham and Watson (2011) and Watson and Callingham (2014). In their 2011 study, Callingham and Watson distinguished between four hierarchical levels of pedagogical content knowledge (PCK) for statistics. Both studies observed teachers' PCK through prompts asking teachers to describe some appropriate and inappropriate student responses to a selection of statistics tasks, and to describe how they would intervene with a student for the inappropriate responses. Therefore, I have used the same method, and for each task, teachers were asked:

1. What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

2. Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Although the phrase "appropriate and inappropriate" could result in multiple interpretations, teachers responded to these items by interpreting it to mean "correct and incorrect" rather than, for instance, thinking a student response could be appropriate because it might be a pedagogically useful response (despite being an incorrect one). However, it is of note that teachers sometimes also included this information in addition to stating whether the student response was correct or incorrect, such as Rosalynn who said, "The incorrect conclusion that 'yes you can, but because of', I love the discussion … So it would be inappropriate-ish. Appropriate with development" (Tomatoes and Fertilizer, Line 26).

In addition to these questions, in preparation for a teacher who might struggle to generate student responses, or for a teacher who might offer student responses that are more generic (e.g., I don't know) or relate to surface errors (e.g., make a mistake when calculating the mean), I gathered examples of responses that exhibited common errors that students made, that were specific to the statistical concepts being measured, on each sub-part for each LOCUS item (see Jacobbe, 2016) to offer teachers during interviews, when necessary. For example, on the New Year's Day Race task (see Figure 3.1), the most common student error when describing which group of runners had more consistent run times was to place attention on the heights of the bars as in the following student responses:

- "while looking at the frequency, it varies dramatically from a high frequency to a low frequency" (Jacobbe, 2016)

- "there are more spikes in the graph for the 5K, and less in the graph for the half marathon" (Whitaker & Jacobbe, 2014, p. 14)

Although the primary purpose of these supplemental questions was to analyze teachers' PCK (RQ3), responses were also used to aid in answering the other two research questions. For instance, when constructing Rosalynn's knowledge map (see Figure 3.2), the connection between the knowledge elements of "Median" and a broader notion of center as "the middle" did not occur clearly until she offered an intervention for an inappropriate student response she identified for the New Year's Day Race task. Similarly, when examining teachers' IIR, sometimes it was not until discussing student responses that a teacher would decide on which way of reasoning they believed was "correct." On the other hand, sometimes discussing student responses would reveal contradictory ways of reasoning. This sometimes happened because of a student response I offered when teachers were struggling, and sometimes it was a result of the teacher having a second opportunity to consider the task. For example, when generating student responses on the Extended School Day task, Kathy changed her response to part (a), claiming she misread the question, and then changed her response to part (b) after considering a new perspective on the data. In contrast, when Amalia indicated she could not think of any further student responses to the Jumping Distance task, I realized that she had not offered one of her own ways of reasoning, so I offered it as an example student response to investigate whether she had reconsidered it as an appropriate way of reasoning or would exhibit a contradiction without realizing it. She claimed it was inappropriate, indicating a contradiction in her reasoning.

Task-based clinical interviews. The LOCUS tasks and supplemental questions about students were administered during task-based clinical interviews (Goldin, 1997). Each participant completed two LOCUS tasks with the supplemental questions about students during each of two task-based clinical interviews that lasted between 60 and 90 minutes. Each interview video and audio recorded, with video capturing only what was written and being gestured toward in order to clarify vague language (e.g., "this graph is more consistent", "what I wrote over here"). Task-based clinical interviews were used, as opposed to only written responses, in order to validate data and strengthen evidence of participants' observed responses (Ginsburg, 1981). Following Ginsburg, the tasks were used to focus participants on specific types of knowledge and the interview protocol was designed to use follow-up questions to require reflection, determine the seriousness of responses, confirm that participants understood the question, and to evaluate participants' strength of belief in their responses by challenging their responses (*right* or *wrong*). Moreover, consistent with Goldin's (1997) recommendations, I used "non-directive follow-up questions" (p. 45) and anticipated as many possible contingencies for participant responses to each task as possible by drawing on the common errors students had made and from 3 teachers responses to the tasks and interview protocol questions during a pilot study.

One anticipated contingency included in the interview protocol for the New Year's Day Race task, for example, was that a participant might make the observed common student error of attending to the bar heights by talking about "spikes" in the histograms, thereby concluding that Jaron's statement that the 5k mile times are more consistent is not supported by the data. Therefore, the first question in the interview

protocol for this contingency was to ask "What do you mean by 'spikes' (or whatever term is used)?" This question does not point the participant in any specific direction and was exploratory in nature (Goldin, 1997), while simultaneously requiring reflection (Ginsburg, 1981). The second pre-planned question for this contingency was "Why do you think the 'spikes' (or other term) are appropriate to use?" This question was designed to determine the strength of belief of the response (Ginsburg, 1981). The final pre-planned question, designed to further challenge the response, as recommended by Ginsburg, was "Is there any possibility that the data *do* support Jaron's statement?"

Moreover, for the questions about student responses, one pre-planned question was "What is it that makes those student responses appropriate or inappropriate?" If the teacher was unable to generate any student responses, the interview protocol included asking, "Do you think students might respond in a similar way that you did?" In the contingency that a teacher could not generate an inappropriate response to suggest an intervention to, or if the inappropriate response(s) was vague or related to surface error (e.g., minor mistake when calculating the mean), I would ask, "What would you do if a student offered ____ (one of the common student errors) response?", followed by a clarifying question if it wasn't apparent if the provided response was believed to be appropriate or inappropriate.

Data Collection

After receiving signed consent to participate in the study, participants were sent Web links via email to complete (1) a background survey and (2) the GOALS assessment. Task-based clinical interviews were then scheduled with each participant. In order to balance the time for each of the two interviews, the first interview included the

New Year's Day Race and Tomatoes and Fertilizer tasks and the second interview included the Extended School Day and Jumping Distances tasks.

Data Analysis

Data analysis was first carried out by constructing maps of teachers' knowledge structures related to measures of center, spread, and shape of distributions through analysis of teachers' responses to tasks and GOALS-2. A cross-case analysis of knowledge maps was conducted to identify categories of knowledge structures (RQ1). Next, teachers' IIR was characterized through Means and Voss's (1996) informal reasoning framework and Makar and Rubin's (2009) IIR components framework through analysis of responses to tasks. Types of IIR were identified through a cross-case analysis (Creswell, 2013) and compared with knowledge structures to identify ways knowledge may support IIR (RQ2). Finally, teachers' responses to supplemental questions about student responses to LOCUS tasks were analyzed using methods described by Watson and Callingham (2008) and Callingham and Watson (2011). The PCK level of each teacher within each task was distinguished using Watson and Callingham's (2014) list of characteristics for each level, as well as by constructing look-fors from their narrative descriptions to aid in identifying PCK levels. Cross-case analysis was then carried out to identify possible types of PCK and then possible relationships were examined with both knowledge structures and IIR (RQ3).

GOALS Assessment

Because the tasks being used to understand teachers' knowledge were created to assess middle level and secondary students' conceptual understandings of statistics (Jacobbe, 2016), the GOALS assessment (Sabbag & Zieffler, 2015) was used to gain further depth of the formal knowledge that teachers might have. Although background

information was collected through a survey, this assessment provides further contextual information regarding the extent of experience teachers have with statistics content. Results are reported as z-scores (see Table 3.3) because the intent is not to explain their formal knowledge, but to explain how teachers might be more similar or different from one another, rather than any perceived deficits.

Table 3.3
GOALS Results

Teacher	GOALS Z-Score
Kathy	-1.17
Ruby	-0.75
Amalia	-0.75
Harrison	-0.54
Ellie	-0.33
Michaela	-0.33
Mike	1.15
Rosalynn	1.36
Tim	1.36

Note from Table 3.3 that there was a clear grouping in which Mike, Rosalynn, and Tim all scored close to one another, and then the other six teachers scored more similar to one another. This is perhaps unsurprising since Mike, Rosalynn, and Tim had all taught stand-alone statistics courses for multiple years. Michaela had also taught a non-AP statistics course, but she had only taught it one year, two years prior to the study. Despite her experience teaching a stand-alone statistics course (which included teaching formal statistics content), her formal statistics knowledge was more similar to those who had never taught such a course.

Mapping Knowledge Structures

To aid in answering the first research question, constructed knowledge elements related to measures of center, spread, and shape of distributions, and connections between them, were identified through open coding of written and verbal responses to each task. Video and audio recordings of participant responses were transcribed and coded using MAXQDA software. Responses to the GOALS-2 assessment were used as supplemental evidence when interview-based evidence of a knowledge element was weak. Because I take a constructivist approach to knowledge, I followed a similar method to that of Ron and colleagues (2010) and modified by Groth and Bergner (2013) that involves mapping Partially-Correct Constructions (PaCCs) for each participant. Knowledge maps are called *partially-correct* because they represent all evidenced knowledge constructions, regardless of whether they represent "impoverished conceptual and procedural knowledge" (Groth & Bergner, 2013, p. 250). I employed Groth and Bergner's modification to this method by using node-link diagrams to visually represent knowledge elements and connections between them. Although these PaCCs were originally intended to trace knowledge element construction across an intended learning trajectory, this method can also be useful for visually representing knowledge elements at a single point in time. Moreover, this method was initially theorized by Ron and colleagues (2010) because they noticed inconsistent and contradictory responses from students and sought a way to describe how this may have occurred. In a similar way, this study seeks to characterize teachers' knowledge with all the inconsistent and contradictory elements and connections intact in order to better understand the relationship it might have with their reasoning (RQ2).

Through the work of both Ron and colleagues (2010) and Groth and Bergner (2013), PaCCs have been identified as *missing element*, *incompatible element*, *disconnected element*, or some combination. Because this study did not intend on aiding participants in constructing knowledge elements, unless participants explicitly claimed they had not constructed a particular knowledge element, it was impossible to identify missing elements. For the same reason, this study could not identify which elements were disconnected—a lack of evidence does not mean the participant had not constructed a connection to the element, but it could imply that connections were not strong enough to warrant making a connection—nor which elements *contributed to* the construction of other elements, as found in Groth and Bergner's (2013) study. My study is therefore limited in this capacity, and therefore, I use the term *knowledge structures* instead of PaCCs to distinguish my knowledge maps from Groth and Bergner's.

For efficient construction of knowledge maps, visual diagramming software was used (Inspiration, 2014). Each element identified through open coding was visually represented as *desirable*, indicated visually by a rectangle, or *undesirable*[1], indicated visually by a rectangle with rounded edges. A *desirable* knowledge element is an element that aligns with current understandings of the content in the field of statistics that is believed to support conceptual development of other knowledge elements and types. On the other hand, an *undesirable* knowledge element is an element the does not align with current understandings of the content in the field of statistics, or would not allow for a

[1] These value-laden terms should not be misconstrued as an indication of a teacher's a lack of expertise because an *undesirable* knowledge element does not necessarily imply that it is inherently incorrect. Moreover, a comprehensive exploration of teachers' knowledge was beyond the scope this study.

desirable knowledge element to be constructed from it. When missing elements were explicitly identified, they were indicated by a rectangle with an X over it. See Figure 3.2 for an example of a knowledge map.

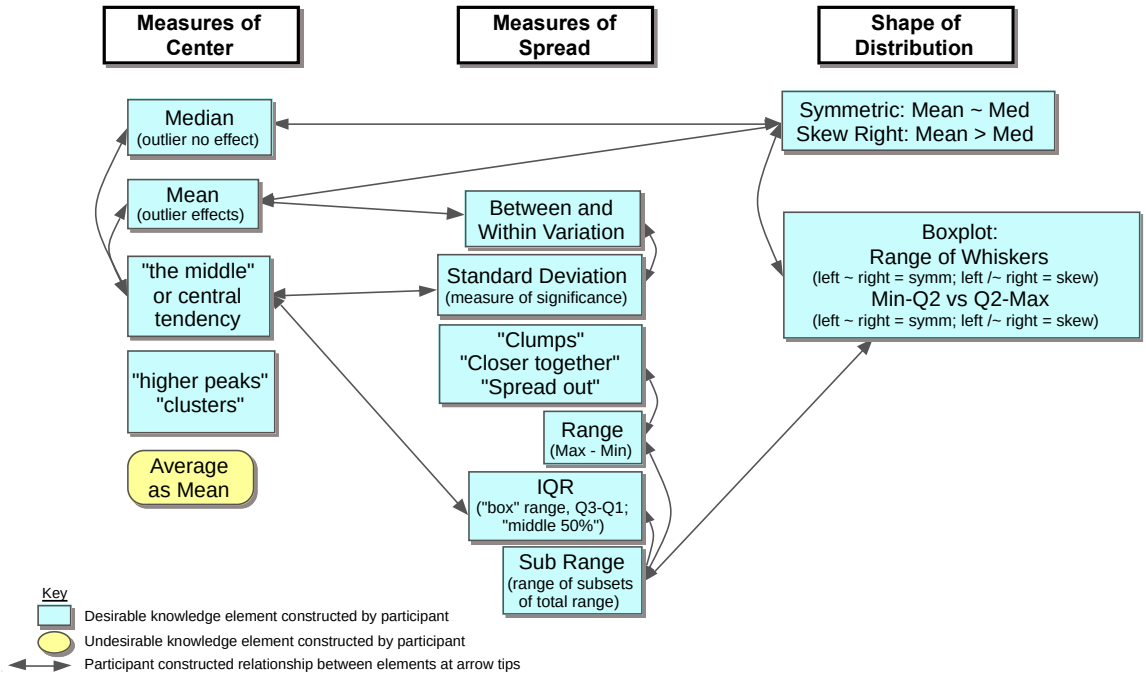


Figure 3.2. Depiction of Rosalynn's knowledge structures.

To illustrate how knowledge elements and connections were decided upon, an *undesirable* knowledge element would be considering the word "average" to imply the arithmetic mean. While an association between these is not necessarily *undesirable* in and of itself, it is *undesirable* when "average" is consistently replaced with "mean." The word "average" is a more general term that could be applied to anything considered "typical" and therefore, the unidirectional meaning that "average" implies "mean" is undesirable because it can inhibit more general ways of thinking in terms of what is "typical." Therefore, it would be visually represented by a yellow rectangle with rounded edges. On the other hand, a *desirable* knowledge element would be that the arithmetic mean is sensitive to extreme values, or outliers. This is considered a *desirable* knowledge

element because it is a statistically sound knowledge construction that is consistently used in the field of statistics. An understanding of the mean in this way allows for a more integrated knowledge base by supporting connections between different conceptions of center (i.e., median, mode) and when each would be most appropriate, while also allowing for connections to be drawn between the respective positions of different measures of center based on the shape of the distribution. Such connections have been found to be hierarchically higher knowledge types (Groth & Bergner, 2006). Accordingly, this element was visually represented as a blue rectangle.

To further illustrate how knowledge elements were placed into maps, consider someone who reasons about the jumping distance task by arguing, "because the IQR of those with a target is smaller than the IQR of those without a target, then those with a target have less variability in jumping distances." In this case, the knowledge element of interquartile range (IQR) is a *desirable* knowledge element because IQR is a measure of spread. Assuming that the response is also evident from a provided display or a calculated result, this statement would be represented by a blue rectangle, labeled with the term IQR. The rectangle would include additional information to provide specifics about the participant's reference to IQR. For instance, if the participant were to gesture to the length of the middle 50% box on a boxplot, a remark such as "length of the box" would be included in the blue rectangle of the knowledge structure map. Similarly, if a calculation was observed, the phrase "Q3-Q1" would be included. Through these additional details, a clearer understanding of participants' knowledge can be represented.

It is of further note that knowledge construction was inferred when participants referenced a particular statistical measure. For instance, when a participant claimed

standard deviation as a possible way to measure spread, it was included in their knowledge structure because it was identified as a measure of spread. It is worth noting, however, that inclusion of an element does not necessarily imply a solid understanding of the concept.

When evidence was provided, connections were drawn between knowledge elements to indicate that the participant had constructed some kind of connection between the two elements. For instance, suppose a teacher stated that in comparing the two boxplots of jumping distances, the distances for those with a target to jump toward are less symmetric and possibly skewed left because the left whisker is longer than the right whisker. Moreover, the teacher might notice that the two whiskers for those without a target to jump toward are fairly similar, indicating a more symmetric distribution. Also, because the distances for those with a target to jump toward are skewed left, a teacher might claim that it implies that the mean will be pulled to the left of the median. Therefore, since the distances for those without a target to jump toward are symmetric, the mean and median will be approximately the same.

Such a possible description from a teacher would provide evidence that the teacher had constructed the *desirable* knowledge elements of "mean" and "median" as measures of center (although more evidence would be sought in other responses to confirm this), and the *desirable* knowledge element that for symmetric distributions the mean and median are similar, whereas the mean is less than the median for skewed-left distributions. This last knowledge element relates the shape of a distribution to measures of center, and therefore, a double-headed arrow would be drawn to connect both the mean and the median elements to the shape element. However, such connections can also be

made between *desirable* and *undesirable* elements. Suppose a teacher constructed the *desirable* knowledge element that in general, spread can be described in terms of "consistency", and also the *undesirable* knowledge element that less consistency means the distribution will have a symmetric shape. This last *undesirable* shape element would then be connected to the *desirable* spread element with a double-headed arrow. Moreover, the shape element would be considered *undesirable* because there are cases in which a distribution can have more spread than another and still maintain its symmetric shape. Furthermore, as with knowledge elements, an arrow was drawn when a participant made an explicit connection. Including the arrows on the map does not imply that the connection is strong or that the participant has meaningfully integrated the elements.

Once knowledge maps were constructed for each participant, I completed a cross-case analysis of the knowledge structures and looked for common patterns in structures that might indicate different categories of knowledge structures. These categories were then described and used to understand differences in support for teachers' IIR (RQ2).

Characterizing Informal Inferential Reasoning

Informal reasoning. Teachers' informal inferential reasoning (IIR) was analyzed by initially attending to the arguments they constructed. Building on the work of Means and Voss (1996), I first identified teachers' claims and reasons for those claims. Continuing to apply Means and Voss' method for analyzing informal reasoning quality, I then identified whether the reason offered was acceptable[2] and supported the claim being made. A teacher's reason (or support) for a claim was determined to be *acceptable* if it

[2] There is a potential for these value-laden terms to be misconstrued as an indication of a lack of reasoning ability. This is problematic because an argument coded as *unacceptable* does not imply that a teacher is incapable of reasoning in *acceptable* ways. It may instead indicate that the response did not address the main issue of the task.

was related to the content, addressed the key features of the particular task item, and drew on the *desirable* knowledge elements as identified by the descriptions of "ideal responses", "sample responses indicating a solid understanding" and "common misunderstandings" provided by Jacobbe (2016). For this reason, the threshold for being categorized as *acceptable* was not uniform across tasks or task sub-parts.

To illustrate this process, on the New Year's Day Race task (see Figure 3.1), a teacher may state that the data displayed in the histograms do not support the given claim that the 5k runners' mile times were more consistent because the range of the 5K runners' times is larger than the range of the half-marathon runners' times. Because a reference to a measure of spread was necessary to be considered an "ideal response" (Jacobbe, 2016), the range is a *desirable* knowledge element of spread, and the difference in the ranges is used to compare consistency, the reasoning to support the claim is considered *appropriate*. On the other hand, suppose a teacher were to reason that if the two bars located to the far right on the 5k histogram (perhaps outliers) were removed, then the mile times for the two races would be similarly consistent because the ranges of the center mass of each plot would be the same. This reason would be identified as *unacceptable* and *not* supporting the claim. Although focusing on a small subrange of the center of the distribution is a *desirable* knowledge element for spread (e.g., IQR is spread of the middle 50%), the subranges of the center mass of each plot would be the same regardless of whether outliers are considered or not. Therefore, this would be identified as *unacceptable* and *not* supporting the claim.

Informal inferential reasoning components. Analysis then focused on which components of IIR were observable within these arguments. I drew on Makar and

Rubin's (2009) framework that includes three components as necessary for engagement in IIR. I adapted Makar and Rubin's framework with Rossman's (2008) inclusion of statements of causality. The following questions were used to identify these components:

- Does the statement move beyond the data or attempt to link a cause?
- Does the statement use data as evidence?
- Does the statement make use of probabilistic language to attach a level of uncertainty to the inference (thus acknowledging the variation)?

Although most arguments could be linked to some components of IIR, such as using data as evidence, some subparts of each task provided more opportunity to observe inferential statements—the main focus. For instance, part (a) of the New Year's Day Race task asked participants to identify which mile times were more consistent. Although the open-ended nature of the question does not limit a statement that attempts to move beyond the data, it does not explicitly encourage it. On the other hand, part (b) stated:

> Recall that individual runners chose to run only one of the two races. Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K? Explain why or why not. (Jacobbe, 2016)

The prompt *made* a generalized statement beyond the data and asked participants to respond to it. Although teachers were not making such statements themselves, this item allowed for observing whether they recognized that this statement *does* move beyond the data and that it is not reasonable to do so. Moreover, this had implications for what types of arguments were coded as *acceptable*. In this specific case, to be coded as *acceptable*, a response had to recognize that an inference was not plausible due to the

design flaw of allowing runners to choose their race—or a lack of random assignment. Furthermore, the response had to explicitly identify that the self-selection was the problem. For instance, stating that the inferential statement was inappropriate because the two groups of runners are different does not necessarily mean that the participant recognized the design flaw. Even if random assignment was used, the two groups would be comprised of different people. It is inadequate to argue that the runners who chose to run the half-marathon were likely more in shape and therefore it is not an appropriate inference. Such a statement might suggest that a participant believes that the inference (as stated in the prompt) should be stated differently for each group of runners—that 5K runners' times would increase if they ran a half-marathon and half-marathon runners' times would decrease if they ran a 5K. Therefore, it was necessary for the participant to explicitly identify the design flaw that prohibits an inferential statement. Across all tasks there were opportunities to observe all three components of IIR (Makar & Rubin, 2009).

A *within-case analysis* (Creswell, 2013) was then carried out across responses to all tasks for each participant through analytic memos. Afterward, a *cross-case analysis* (Creswell, 2013) was performed to locate patterns of IIR while simultaneously describing how knowledge structures may support qualitatively different types of IIR.

Distinguishing Levels PCK

To aid in answering the third research question, regarding teachers' knowledge of students as a sub-domain of statistics PCK, I draw on Callingham and Watson's (2011) study that distinguished among four hierarchical levels of statistics PCK through analyzing responses to the supplemental questions asked of each teacher per LOCUS task. Following their method, and adding information about the common errors students made from the LOCUS assessment, the following questions were used to code responses:

- Was the student response identified as appropriate or inappropriate?
- Was the teacher's identification aligned with typical responses on LOCUS?
- Was the student response generated by the teacher?
- For each suggested intervention for a student response:
    - Does the teacher recognize the inappropriate nature of the response?
    - Did the teacher draw on the student response?
    - Did the suggested intervention relate to statistics content?
    - Did the suggested intervention relate to the context of the task?
    - Did the teacher suggest a generic intervention?

After identifying these characteristics for each response, a *within-case analysis* was completed for each participant by taking a holistic view across all responses to each task. In particular, the frequency of both appropriate and inappropriate responses was tabulated, along with how many times they aligned with the common correct responses and common errors students made on the LOCUS assessment. I then turned to Watson and Callingham (2014) who utilized the same method, and who provided a more succinct description of each hierarchical level of statistics PCK (see Table 3.4).

Table 3.4

Levels of PCK (adapted from Watson and Callingham, 2014, p. 267)

Hierarchical Level	Characteristics
Accomplished	Teacher responses suggest a range of appropriate and inappropriate student responses and an integration of appropriate statistical and mathematical content with student-centered intervention strategies.
Competent	Teacher responses address more traditional and familiar topics in the school statistics curriculum. Some statistically appropriate interventions are suggested but only in the context of familiar classroom activities.
Emerging	Teachers use some statistical knowledge to suggest several either appropriate or inappropriate responses for students. Generic rather than content-specific strategies are suggested for classroom intervention, which implies good teaching but not necessarily in the context of statistics.
Aware	Teachers may suggest a single appropriate or inappropriate student response to the items. They display little broader statistical understanding, and do not make suitable suggestions for addressing students' understanding.

To add clarity to these characteristics, drawing on examples of responses at each level provided in Watson and Callingham (2014), the look-fors in Table 3.5 were developed and used to identify each level.

Table 3.5
Levels of PCK look-fors

Hierarchical Level	Look-fors
Accomplished	Appropriate and inappropriate responses offered Intervention suggestion tied to content Series of questions offered to lead the student Suggest ways to challenge the student's belief
Competent	Appropriate and inappropriate responses offered Intervention suggestion tied to content Questions are more rhetorical Points student directly to a specific method/way of reasoning
Emerging	Several either appropriate or inappropriate responses offered Intervention suggestions are general and no tied to content Generic questions Generic comments
Aware	Single appropriate or inappropriate response offered Intervention suggestion doesn't address student misunderstanding Questions and comments not clearly tied to student response or content

After identifying the level for each task for each teacher, a *cross-case analysis* was carried out to identify categories of types of statistics PCK. These were then compared with categories of knowledge structures and categories of IIR in order to identify possible relationships between knowledge, IIR and statistics PCK.

Validity and Reliability

Both the GOALS-2 assessment (Sabbag & Zieffler, 2015) and the items used from the LOCUS assessment (Jacobbe et al., 2014) have been found to have content validity by their respective research teams. Moreover, the LOCUS assessment items were further validated to have content validity through the evidence model that was used to construct the assessment items (Haberstroh et al., 2015; Jacobbe et al., 2014). Regarding validity and reliability based on a psychometric analysis, Sabbag and Zieffler's analysis resulted in a claim that although the GOALS-2 assessment is still being refined, "the present level already has the potential to become a useful tool in studying and improving

students' statistical reasoning" (p. 110). The current version of the GOALS assessment is a revised version based on Sabbag and Zieffler's results. In regards to the LOCUS assessment, psychometric analyses are currently underway. However, for the purposes of this study, because I selected tasks from the assessment to use in an interview setting, the established content validity is sufficient to claim that the items measure what it is intended. In a similar way, there is a level of reliability in the questions that were asked in order to elicit responses to measure statistics PCK because of the field's acceptance of this method through multiple peer-reviewed publications within both mathematics education and statistics education (Beswick, Callingham, & Watson, 2012; Watson & Callingham, 2013, 2014; Watson et al., 2008).

As for data from task-based clinical interviews, reliability comes from the interview protocol (see Appendix G). As claimed by Goldin (1997), creating an interview protocol that anticipates "sufficiently many problem solving contingencies" leads to "the creation of reproducible task-based clinical interviews" (p. 53), thus providing reliability in the data collected during the interview through means of the carefully constructed interview instrument. Multiple contingencies were anticipated for every task through information obtained about students' common responses to LOCUS items (Jacobbe, 2016) as well as pilot study participants' responses to the items. Moreover, questions were asked in order to elicit verbalizations, evaluate them, and check for alternative hypotheses, while also aiding in determining participants' seriousness of responses, confirming their understanding of the task, and determining their strength of belief in their responses (Ginsburg, 1981). Through this process, the evidence provided by the data is strengthened to aid in making more valid claims.

Validity of data analysis is established first through *peer debriefing* (Creswell, 2013), in which I met with my advisor and other members of my dissertation committee on several occasions to discuss ongoing analysis. Moreover, I presented some of my data analysis to several research experts during a seminar to garner key feedback about portions of my analysis. These *peer debriefings* either confirmed analyses or led to refinements of the resulting claims I made. A second form of validity comes through triangulation of data sources (Patton, 2002). This was achieved by having multiple data sources to compare against in order to check consistency of evidence that was generated through all analysis methods. More specifically, when constructing knowledge maps, evidence was examined across all four LOCUS tasks, the GOALS assessment, and within questions about students. When analyzing teachers' IIR, not only were there four different LOCUS tasks to compare evidence across, but IIR was elicited when discussing student responses to each task as well. When analyzing teachers' statistics PCK, there were four different tasks in which to cross-check for consistency. The overall validity and credibility of the findings is strengthened by the presence of multiple data sources from which to examine consistency of evidence being used to support claims. Finally, although a potential limitation of the study, interrater reliability in the traditional sense was not established because of the inherent impracticality of identifying a second coder with sufficient expertise across all coding domains (statistics knowledge, IIR, and statistics PCK).

Summary

In this study, I used a task-based clinical interview design (Goldin, 1997) with cross-case analysis (e.g., Groth & Bergner, 2013) to study a stratified purposeful sample (Patton, 2002) of nine middle and secondary mathematics teachers. Data sources first

included teachers' written responses to the GOALS-2 assessment (Sabbag & Zieffler, 2015) as a measure of background knowledge in formal statistics as well as supplemental evidence of teachers' knowledge of center, spread, and shape (RQ1). The second data source was teachers' written and verbal responses to four released tasks from the LOCUS assessment (Jacobbe, 2016) during task-based clinical interviews. These tasks were used to examine 1) teachers' knowledge of center, spread, and shape, and 2) teachers' informal inferential reasoning. The final data source was teachers' written and verbal responses to the following two supplemental questions for each part of each LOCUS task, as used by Watson, Callingham, and Donne (2008), during task-based clinical interviews:

1. What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.
2. Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Responses to these supplemental questions were used first and foremost to aid in examining teachers' PCK for statistics (RQ3). However, responses were also used to gain insight into teachers' knowledge of center, spread, and shape of distributions (RQ1), as well as their informal inferential reasoning (RQ2).

To examine teachers' knowledge of center, spread, and shape, knowledge structure maps were constructed (Groth & Bergner, 2013) to indicate whether observed constructed knowledge elements were *desirable* or *undesirable* as well as connections teachers were observed to have made between elements. After a *within-case* analysis, a *cross-case* analysis was carried out to identify possible categories of types of knowledge

structures. To examine teachers' informal inferential reasoning (IIR), teachers' supports for claims were first identified as *acceptable* or *unacceptable* (Means & Voss, 1996). Then, for tasks that were designed to engage participants in IIR, responses were coded as to whether or not they included one of the three components of IIR (Makar & Rubin, 2009): (a) generalization beyond the data or claim about causality (*inference* component), (b) using data as evidence (*data component*), and (c) probabilistic language to indicate uncertainty in the inference (*uncertainty* component). After a *within-case* analysis was completed to identify each teacher's dominant form of reasoning, a *cross-case* analysis was carried out to identify possible categories of reasoning in both non-IIR and IIR contexts. These categories were then compared with knowledge structures categories to identify possible ways teachers' knowledge supported their IIR. Last, teachers' PCK for statistics was analyzed by identifying which of the four levels of PCK described by Callingham and Watson (Callingham & Watson, 2011; Watson & Callingham, 2014) best characterized each teacher—*aware*, *emerging*, *competent*, or *accomplished*. Teachers' PCK levels were then examined across knowledge structure types and reasoning categories in order to identify possible relationships among them.

The GOALS-2 assessment and LOCUS tasks were both identified as having content validity (Haberstroh et al., 2015; Jacobbe et al., 2014; Sabbag & Zieffler, 2015) and the supplemental questions used for identifying teachers' PCK have been established through multiple publications in the field (Callingham & Watson, 2011; Watson & Callingham, 2013, 2014; Watson et al., 2008), thus providing a validity to the questions and methods. Moreover, establishment of an interview protocol that anticipated many responses prior to the interviews aids in establishing a reproducibility of the interview,

providing some validity and reliability in the data collected from them (Goldin, 1997). Validity of data analysis is established through peer debriefing (Creswell, 2013) and triangulation of data sources and comparing evidence across sources (Patton, 2002).

CHAPTER 4: ANALYSIS OF THE DATA AND RESULTS

Considering the ever-increasing attention on statistics and the need for society to understand how inferences from data can be made and when they are possible, it is becoming more important society to be statistically literate. The direct implication of this goal is that teachers need to have rich experiences with *informal inferential reasoning* so that they can engage their students in the kinds of rich experiences that will prepare them for a data-driven world. This study examined nine middle and secondary teachers' knowledge structures, *informal inferential reasoning* (IIR), and pedagogical content knowledge (PCK) for statistics. The chapter is organized by each research question. First, I describe teachers' knowledge structures for center, spread, and shape (RQ1), placing them into largely distinct categories. Next, I present teachers' reasoning in non-IIR contexts and IIR contexts in order to express salient differences before identifying possible ways that teachers' knowledge structures may support their reasoning (RQ2). Last, I portray teachers' PCK in both non-IIR and IIR contexts and discern possible relationships among teachers' knowledge structures, IIR, and PCK for statistics (RQ3).

Knowledge Structures for Measures of Center, Spread, and Shape

Background Knowledge

The GOALS-2 assessment provided an initial measure of teachers' experience with inferential reasoning, revealing that the majority of the teachers in this study lacked significant background in formal statistics (see Table 4.1). Moreover, it showed that teachers of stand-alone statistics courses had more experience with formal statistics, with the exception of Michaela who had only taught a non-AP Statistics course one time, two years prior to the study. Therefore, it is no surprise that her knowledge of formal statistics

was more similar to those who taught statistics as a unit within the context of middle level or high school mathematics courses (e.g., Pre-Algebra, Algebra I).

Table 4.1
GOALS Assessment Results

Teacher	Course Statistics Taught In Most Recently	Years Teaching Statistics	GOALS Percent of Items Correct	GOALS Z-Score
Kathy	6th Grade Mathematics	4	40	-1.17
Ruby	Pre-Algebra & Algebra I	10	50	-0.75
Amalia	Algebra I	3	50	-0.75
Harrison	Pre-Algebra & Algebra I	5	55	-0.54
Ellie	Pre-Algebra	3	60	-0.33
Michaela	Non-AP Statistics	1 (3 as unit)	60	-0.33
Mike	AP Statistics	2 (7 as unit)	95	1.15
Rosalynn	AP Statistics	4 (7 as unit)	100	1.36
Tim	Non-AP Statistics	5 (10 as unit)	100	1.36

Knowledge Structures

Maps of teachers' knowledge of center, spread, and shape of distributions were categorized into 3 groups during cross-case analysis. These three categories were described in terms of the features of knowledge elements that were described in Chapter 3—that is, *desirable* and *undesirable* knowledge elements, and perceived connections between such knowledge elements. Categories are defined as *desirable-connected structures* (3 teachers), *undesirable-connected structures* (4 teachers), and *undesirable-disconnected structures* (2 teachers). Although these categories are largely distinct, one *undesirable* knowledge element was found to be in common across all knowledge maps. In particular, all teachers used the terms 'average' and 'mean' interchangeably, and, most importantly, when presented with the term 'average', they consistently replaced the term with 'mean.' This is considered an *undesirable* knowledge element because the term 'average' is a more general term that could be used to describe something as broad as

what is typical in a data set, or something more specific such as the geometric mean. Furthermore, teachers' interchangeable use of 'average' and 'mean' is *undesirable* because it could be the basis for fostering a misconception among students. For instance, two teachers (Harrison and Michaela) were more explicit when talking about average and mean, stating that the term 'average' can *only* imply the arithmetic mean. The following sections will describe the salient, distinct features of each of the three categories.

Desirable-connected knowledge structures. Knowledge structures that are in the category of *desirable-connected* had the following common characteristics: 1) the only *undesirable* knowledge element was the one found across all nine teachers' structures, 2) there were no connections to *undesirable* knowledge elements, and 3) there were many connections *within* and *between* knowledge element types (center, spread, and shape). This category includes the knowledge structures of Rosalynn, Mike, and Tim, all of whom also happened to score the highest on the GOALS-2 assessment and all of whom had taught stand-alone statistics courses at the secondary level for multiple years. To aid in describing this type of knowledge structure, Rosalynn's knowledge structure map is used as a representative case (Figure 4.1).

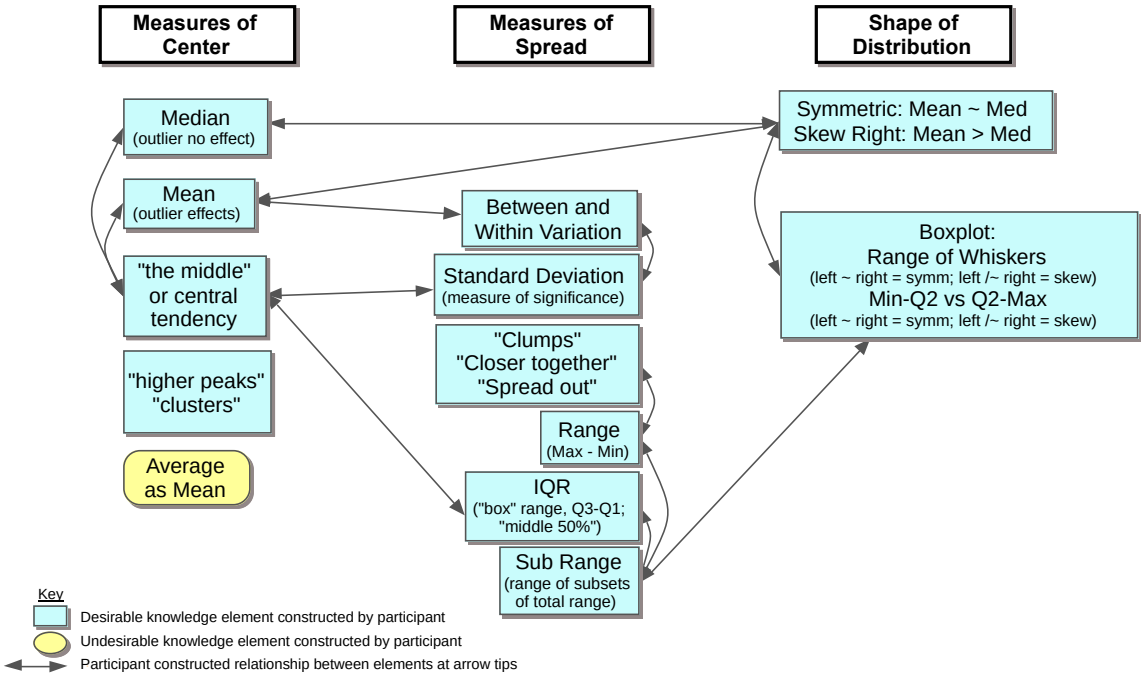


Figure 4.1. Rosalynn's knowledge structure map as an example of a *desirable-connected* structure

Connections within knowledge element types. As can be seen in Figure 4.1, there are multiple connections between knowledge elements *within* each knowledge element type. For instance, Rosalynn exhibited evidence of connecting the more general idea of center as "the middle" or using the term "central tendency" with the median as a measure of center. In the quote below, Rosalynn is describing a student response to part (b) of the New Year's Day Race task that she believes should be focused on the mean rather than the median:

> they say 'the middle of the graphs are really close together', where they have the idea that the mean or average is center, but the wrong application of it. So then you can talk about median being the actual frequency middle, versus um, arithmetic mean and the calculation that goes into that (Rosalynn, New Year's Day Race, Line 76).

In this quote, Rosalynn makes a connection to the more general idea of center as middle, and moves to connect that idea with the median as opposed to the mean. Therefore, these two knowledge elements are connected *within* the knowledge element type for measures of center.

Connections between knowledge element types. The other type of connections that are made in the category of *desirable-connected structures* are connections *between* knowledge element types. In Figure 4.1, both knowledge elements that were just discussed, "the middle" and "median", are connected to other knowledge element types. For example, "the middle" is connected to Rosalynn's conception of IQR, thus making a direct connection between the measure of center and measure of spread knowledge element types. This type of connection *between* knowledge element types of center and spread were observed across all three teachers' knowledge structures. Furthermore, "median" and "mean" are directly connected to the knowledge element describing the relation between mean, median, and symmetry as evidenced in this statement by Rosalynn: "with that somewhat symmetric shape your mean and your median are going to be about the same" (New Year's Day Race, Line 24). This type of connection *between* knowledge element types of center and shape was also observed across the knowledge structures of all three cases (Rosalynn, Mike, and Tim).

To further illustrate the integrated nature of these knowledge structures, the knowledge element of shape that relates to the respective locations of mean and median is also connected to the knowledge element that symmetry is observed through comparison of the spread of boxplot whiskers (see Figure 4.1). In turn, this way of observing symmetry is connected to the observed knowledge element of using sub-ranges in order

to understand variability. Although indirect, these kinds of paths provide evidence of the integrated use of center, spread, and shape. In this particular path, which was observed in the knowledge structures across all three cases, knowledge can be interpreted as center and spread being mediated through concepts of shape. In this specific instance, shape was identified through comparison of sub-ranges and this identification of shape was used in order to predict locations of, and compare, the mean and median measures of center—as observed in the previous quotes by Rosalynn.

This pattern of multiple connections among knowledge elements *within* knowledge element types, and among knowledge elements *between* knowledge element types was strong across the three cases of this knowledge structure category (See Appendix H for all knowledge structure maps). Therefore, their knowledge of center, spread, and shape is observed to be highly integrated. Moreover, *undesirable* knowledge elements were exceedingly rare, and when they did occur, they were not observed to be connected to other knowledge elements.

Undesirable-connected structures. Knowledge structures that were categorized as *undesirable-connected* had shared characteristics of multiple *undesirable* elements, and yet were also highly interconnected. Across all four cases of *undesirable-connected* structures, the total number of *undesirable* elements per structure was highly variable, ranging between 3 and 8. Moreover, connections were found between *undesirable* and *desirable* elements across all four cases. This category included the cases of Ruby, Ellie, Harrison, and Michaela.

A fairly substantial amount of variance is visible in their knowledge structure elements—more so than the other structure types. For instance, Ellie's and Harrison's

structures are the only ones with evidence of connections between two *undesirable* elements. Ruby's structure is the only one without a *desirable* knowledge element for shape and is also the only structure without a direct connection between shape and center. Furthermore, Ruby's and Ellie's structures were the only ones that evidenced at least one *undesirable* element for all knowledge element types (center, spread, and shape). However, the overall structure of their knowledge remains similar—highly connected, yet with multiple *undesirable* elements (see Figure 4.2).

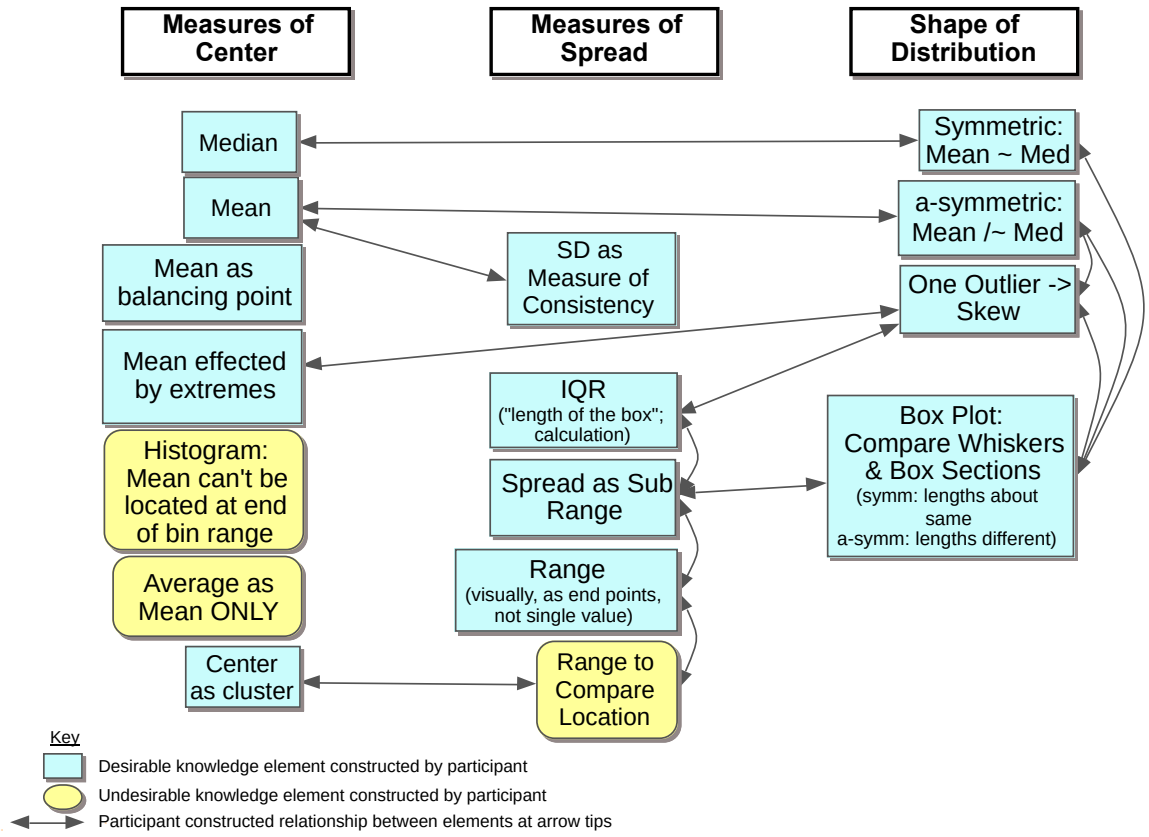


Figure 4.2. Harrison's knowledge structure map as an example of an undesirable-connected structure

Connections within knowledge element types. Harrison's knowledge map (Figure 4.2) is representative of the knowledge structures categorized as *undesirable-connected*, containing all of the common characteristics. In contrast to the *desirable-connected*

knowledge structures, although all the *undesirable-connected* structures contained connections *within* knowledge element types, at least one connection was between a *desirable* and an *undesirable* element. For instance, on the Jumping Distances task, one item prompted participants to compare the variability in jumping distance for a group with a target to jump toward and a group without a target to jump toward. Harrison begins by saying, "From an overall range, or min max, it looks like there's a difference of about 10 on the high end [gesturing to max of both plots], and like, what would that be on the low end, like 15 or so [gesturing to min of both plots], 15, 16 something like that" (Line 18). This is evidence of a *desirable* knowledge element that range is a measure of variability. Further evidence that he considers the range to be a measure of spread was provided in the New Year's Day Race task when he stated "it has a wider, the range is [gesturing to max and min on histogram of 5K runners' mile times], I guess that's the most obvious one that's bigger" (Line 9), after being prompted about consistent running times.

Harrison's focus on comparing the end points of the range on the Jumping Distances task, rather than the range as a distance, led him to then compare differences in location of all 5-number summary values and conclude "with the exception of that first quartile, everything else is shifted over in that 10 to 15 range, so yeah" (Line 19). This idea of *shift* is concerned with a broader sense of center rather than variability, thus connecting the *desirable* element of range as measure of spread with the *undesirable* element of range as a measure of center. Moreover, on the New Year's Day Race task, he indicated that students should focus on the "center mass" and "ignore the tails" (Line 122) when comparing consistency, providing further evidence to support that he believes

that range 1) is a measure of spread and 2) can be used to compare locations (centers). Because evidence indicates that he has constructed both of these contradictory views of range and considers both as measures of spread, they are, therefore, connected *within* the same knowledge element type to accurately reflect his knowledge structure.

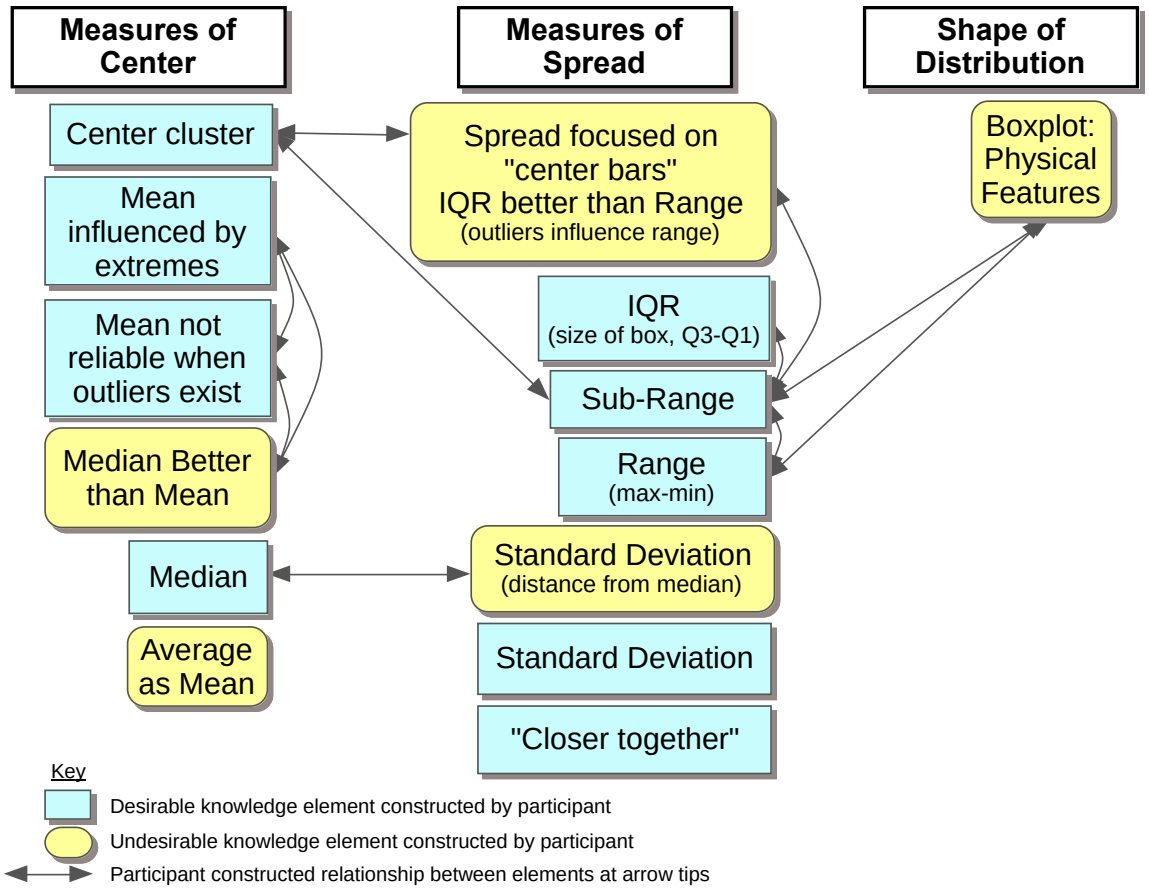


Figure 4.3. Ruby's knowledge structure map as an example of an *undesirable-connected* structure

Connections between knowledge element types. The second common feature across all cases of *undesirable-connected* knowledge structures is that all three knowledge element types are interconnected through either direct or indirect connections. An indirect connection is one that passes through one knowledge element type in order to get to another knowledge element type, but with no direct path between the two. These

may happen between *desirable* or *undesirable* knowledge elements. For instance, Ruby's knowledge structure (see Figure 4.3) indicates that she had not been observed to make a connection between her conceptions of shape and center. However, when she described the shape from a boxplot representation provided on the Jumping Distances task (referring only to physical characteristics, an *undesirable* element), she also described sub-ranges of the data, thus connecting shape to a *desirable* way of thinking more flexibly about spread.

Further extending the indirect connection, on both the Jumping Distances and New Year's Day Race tasks, she indicated that a sub-range that focused on the "center bars" (New Year's Day Race, Line 23 and 105), or the inter-quartile range (Jumping Distances, Line 12) is preferred. This connected the *desirable* element of sub-range to the *undesirable* element that spread *should* focus on the center portion—much like Harrison did. The last connection finally makes the indirect connection from shape to center when she speaks more generally to compare centers by talking about "the entire IQR being higher" (Jumping Distances, Line 57). This connected the *undesirable* knowledge that spread *should* focus on a central sub-range, to the *desirable* conception of center being focused on a central sub-range. Therefore, by indirect means, Ruby's knowledge of shape as physical features of a boxplot is connected to her knowledge of center as a central sub-range.

All of the *undesirable-connected* knowledge structures contain multiple indirect connections among center, spread, and shape knowledge element types. Although this provides further indication that knowledge structures were inter-connected, every knowledge structure contained at least one indirect path across knowledge element types

that included at least one *undesirable* knowledge element. In other words, many of the *undesirable* knowledge elements were integrated with the *desirable* elements and not disconnected from the overall structure.

Background characteristics. All four cases of *undesirable-connected* knowledge structures scored similarly on the GOALS-2 assessment (see Table 4.1). However, there was much wider variability in their backgrounds than the other two knowledge structure types. For instance, Ruby, Ellie, and Harrison had all taught statistics as a unit within either Pre-Algebra or Algebra I at the 8^{th} and 9^{th} grade level. However, Michaela had taught a stand-alone statistics course at the 12^{th} grade level. In terms of total teaching experience, Ellie had 3 years of experience while the others had between nine and 12 years, and Ruby had 10 years of experience teaching statistics as a unit while the others had between three and six years. Therefore, the only common characteristic in their background appears to be a similar experience with formal statistics content. The most salient difference is that the teachers observed to have *undesirable-connected* structures evidenced a much lower formal understanding of statistics, as seen in their substantially lower GOALS-2 scores. For instance, examining *z*-scores indicates that teachers with *undesirable-connected* structures scored between 1.48 and 2.11 standard deviations below the teachers with *desirable-connected* structures.

Undesirable-disconnected structures. Knowledge structures in the category of *undesirable-disconnected* are structures that were observed to contain multiple *undesirable* knowledge elements that are mostly disconnected from one another. Amalia and Kathy are the two cases whose knowledge structures are categorized as *undesirable-disconnected*. For illustration purposes, Kathy's structure is chosen as an example

because it represents the more connected of the two (see Figure 4.4). As can be seen in Kathy's knowledge structure map, there are multiple *undesirable* knowledge elements. However, on this point alone, there is not much difference from the *undesirable-connected* structures category in terms of frequency of *undesirable* elements. The most salient characteristic is that there are few observed connections between elements. This is true for both connections *within* knowledge element types and *between* knowledge element types. In fact, neither Kathy's nor Amalia's knowledge structures provide evidence of direct connections between center and shape, and Amalia's also does not provide evidence of an indirect connection between center and shape.

Connections between undesirable and desirable elements. Among the limited number of connections, both Amalia and Kathy were observed to make connections between *undesirable* and *desirable* knowledge elements. For instance, Kathy had constructed the *desirable* knowledge element of the inter-quartile range as being the range of the middle 50% of the data. She was also observed to have connected this element to an *undesirable* knowledge element, as seen below, when discussing the center of a dotplot on the Extended School Day task:

> And we talk about that you know, the average, when you're talking about data, is in that middle 50 percent" … "that's not the average, or that's not typical, that's not in the typical 50 percent or whatever. (Extended School Day, Line 158)

In this excerpt, Kathy twice connects the idea of average to the middle 50%, indicating that she has constructed the knowledge that the average should fall within the middle 50% of the data. This was identified as an *undesirable* knowledge element because it is not necessarily the case that the average, which earlier in the task she had

related to the mean, will be in the middle 50% of a data set—this only holds for unimodal symmetric distributions.

Connections within knowledge element types. Another characteristic found in both knowledge structures was that some connections were made *within* knowledge element types. For example, Kathy made several connections between knowledge elements relating to spread. On the Jumping Distances task, Kathy was comparing the two boxplots and stated "the target group, the middle 50 percent is so close together but then the range, I mean, you can tell that obviously the lower 25 percent has a wide range of ability" (Line 22). In this statement, she is simultaneously drawing on a flexible sense of range and comparing two different sub-ranges of data by using the term range. She is also using the more generic phrase "close together" to describe the range of the middle 50% of the data. This statement is one of multiple that indicates the constructed *desirable* knowledge elements of 1) "close together" as generic sense of spread, 2) range, and 3) sub-range—as well as connections between them. However, although these connections seem plausible as a foundation for considering what these subranges imply about the density and what that may imply about the shape, Kathy only referred to shape by describing range, IQR, and comparing sub-ranges as in the above quote.

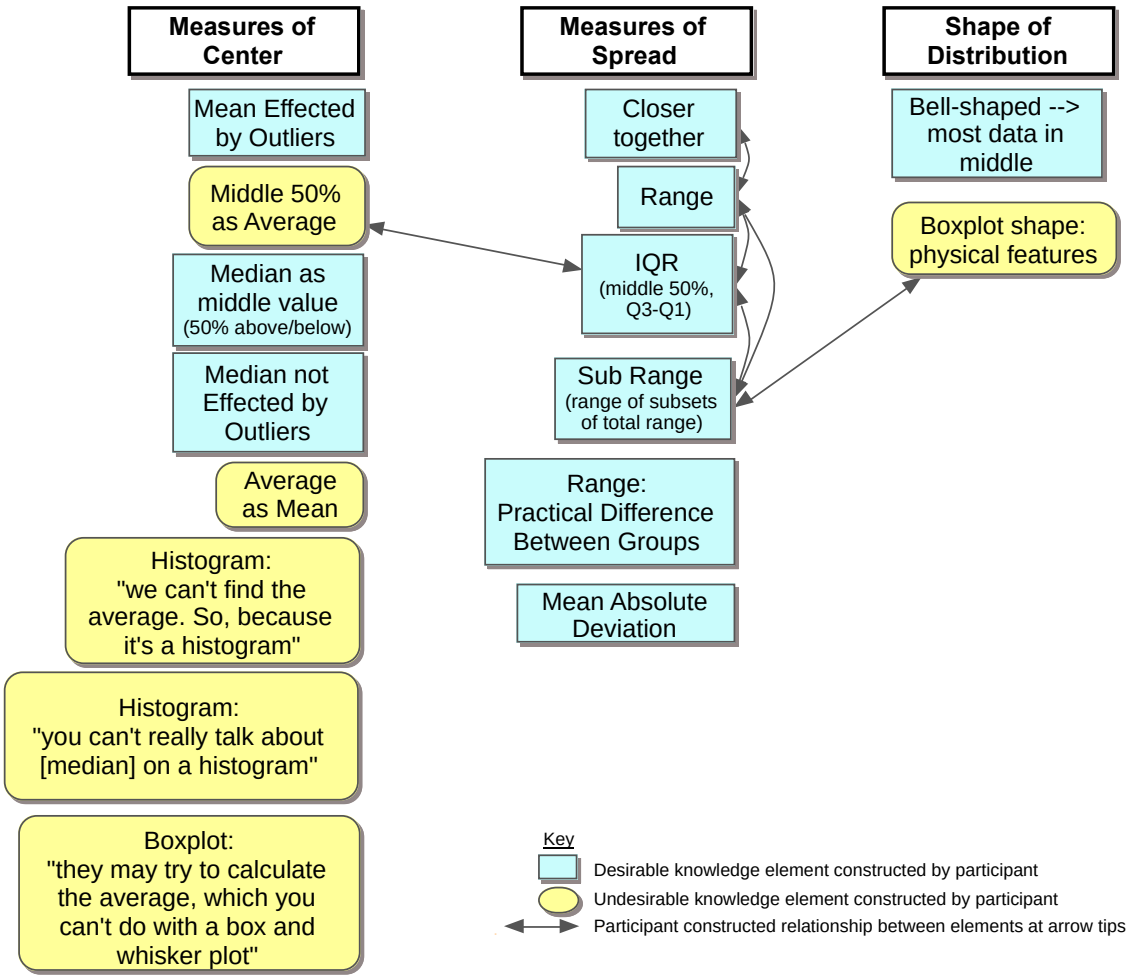


Figure 4.4. Kathy's knowledge structure map as an example of an *undesirable-disconnected* structure

Background characteristics. Amalia and Kathy have very similar teaching experience with both mathematics and statistics content, except that Kathy teaches statistics in the context of grade 6 mathematics while Amalia does so in Algebra I with grade nine students. They both have the same amount of post-secondary coursework in statistics and scored similarly on the GOALS-2 assessment. However, Kathy reported having attended several professional development workshops focused on statistics content, while Amalia reported attending none. It is also important to note that although Kathy and Amalia scored as low or lower than all other teachers in the study on GOALS-

2, their formal statistics knowledge can be assumed to be similar to the teachers with *undesirable-connected* knowledge structures. Comparing *z*-scores of their GOALS-2 scores indicates that teachers with *undesirable-disconnected* structures scored between 0 and 0.83 standard deviations below teachers with *undesirable-connected* structures. However, they also scored between 1.90 and 2.53 standard deviations below teachers with *desirable-connected* structures.

Summary. Knowledge structures fell into three broad categories—*desirable-connected*, *undesirable-connected*, and *undesirable-disconnected*. Not only are the knowledge structures distinctly different between these categories, the teachers with *desirable-connected* structures were observed to have substantially more formal statistics knowledge and experience teaching stand-alone statistics courses. It is at this point that the groups no longer differ by background characteristics (see Table 3.1). *Desirable-connected* knowledge structures (Figure 4.1) were observed to have one *undesirable* element, average as mean (observed in all nine teachers' structures), and were highly integrated both *within* and *between* knowledge element types (center, spread, shape). *Undesirable-connected* knowledge structures (Figure 4.2 and 4.3) were also highly integrated both *within* and *between* knowledge element types. However, there was more variability in that some knowledge structures only contained *indirect* connections *between* one pair of knowledge element types—all structures had *direct* connections *between* at least two of center, spread, and shape. Moreover, the highly-interconnected nature of *undesirable-connected* structures resulted in the integration of many of the *undesirable* elements. In all structures, at least one indirect path connecting center, spread, and shape included at least one *undesirable* knowledge element. *Undesirable-*

disconnected knowledge structures (Figure 4.4) had strikingly few connections between elements, and both cases in this category included *undesirable* elements in some of those connections in their structures. Direct paths between center and shape were absent, and the frequency of *undesirable* elements was fairly similar to the *undesirable-connected* structures.

Knowledge Structures Support for Informal Inferential Reasoning

Although each of the four LOCUS tasks incorporated IIR, they did not explicitly call for engagement in IIR among all sub-parts of each task. More specifically, they did not explicitly provide or call for an inferential statement and therefore did not call for inferential reasoning. In particular, the eight sub-parts across all four tasks were identified as either non-IIR or IIR as follows:

- Non-IIR sub-parts:
    - New Year's Day Race–Part (a): compare consistency of mile times
    - New Year's Day Race–Part (b): compare mile times on average
    - Extended School Day–Part (a): how surprising is the sample
    - Jumping Distances–Part (a): compare center, spread, shape
- IIR sub-parts:
    - New Year's Day Race–Part (c): respond to an inference
    - Tomatoes and Fertilizer (no sub-parts): respond to an inference
    - Extended School Day–Part (b): respond to an inference
    - Jumping Distances–Part (b): make an inference

It is also of note that due to the interview protocol that was designed to explore the strength of belief that each teacher had in their responses, most of the time teachers

offered multiple arguments for each sub-part. Therefore, it should not be assumed that only nine total responses exist for each sub-part, and their reasoning may have been coded as both *acceptable* and *unacceptable* for the same sub-part because each argument was coded separately.

Informal Reasoning on Non-IIR Tasks

In total, four of the eight sub-parts across all tasks did not explicitly encourage *informal inferential reasoning* (IIR). In other words, they did not 1) make inferential statements in the task stem, 2) prompt for criticisms of inferential statements, or 3) prompt for an inferential statement to be made. Moreover, teachers were not observed to make inferential statements to either generalize to a larger population or to make claims about causality on any of these tasks. Therefore, as expected, none of these task parts included inferential statements. This section will focus on teachers' claims about data and the reasoning used to support those claims.

Comparing center, spread, and shape of two distributions. Three of the four sub-parts included familiar activities of comparing center, spread, and shape of two distributions, and a histogram and a boxplot were supplied as representations of the data. These are familiar because, first, measures of center, spread, and shape, as well as boxplots and histograms, are explicitly included in the *Common Core State Standard for Mathematics* (CCSSM) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) beginning in 6th grade (see 6.SP.2–6.SP.5). Second, comparing two distributions is explicitly included in CCSSM as early as 7th grade (see 7.SP.3). These three sub-parts include parts (a) and (b) from the New Year's Day Race task, which included two histograms (see Figure 4.5) and prompted participants to decide which group's mile times were more "consistent" (part a) and

which group had a higher mile time "on average" (part b). The third sub-part is part (a) from the Jumping Distances task, which included two boxplots (See Figure 4.6) and prompted participants to compare the "center, variability, and shape" of the two groups' jumping distances. See Appendix F for exact wording and main task stems.

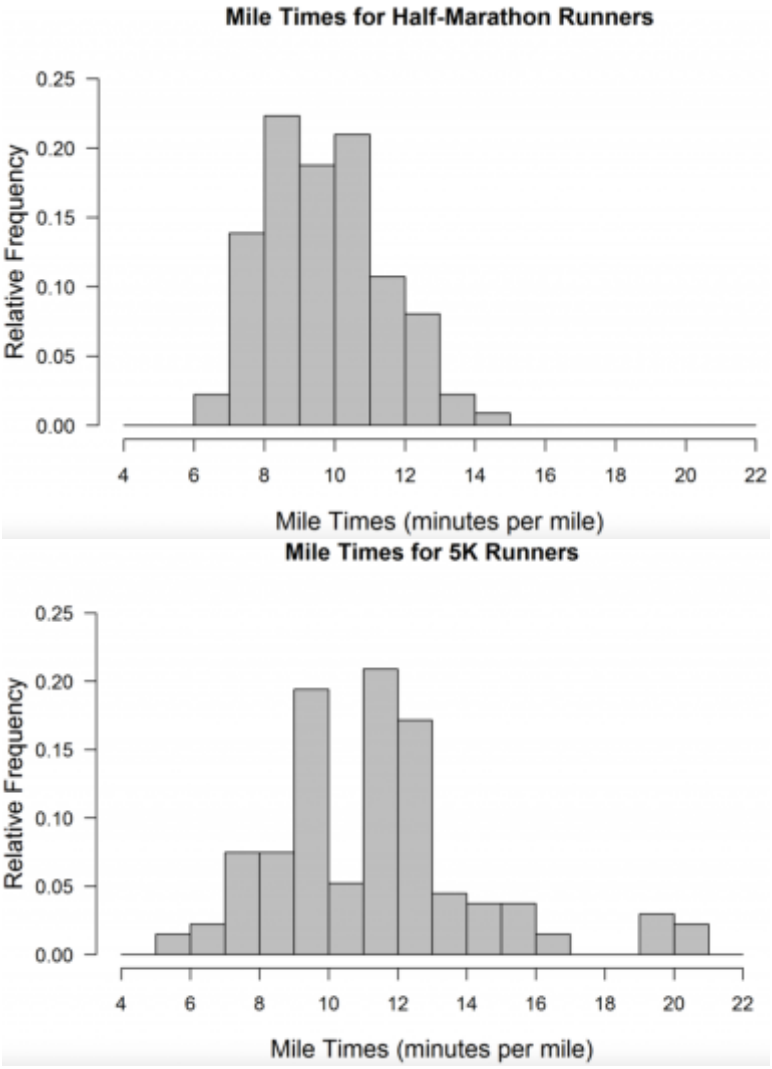


Figure 4.5. Histograms provided on the New Year's Day Race task. Released item from LOCUS assessment (Jacobbe, 2016).

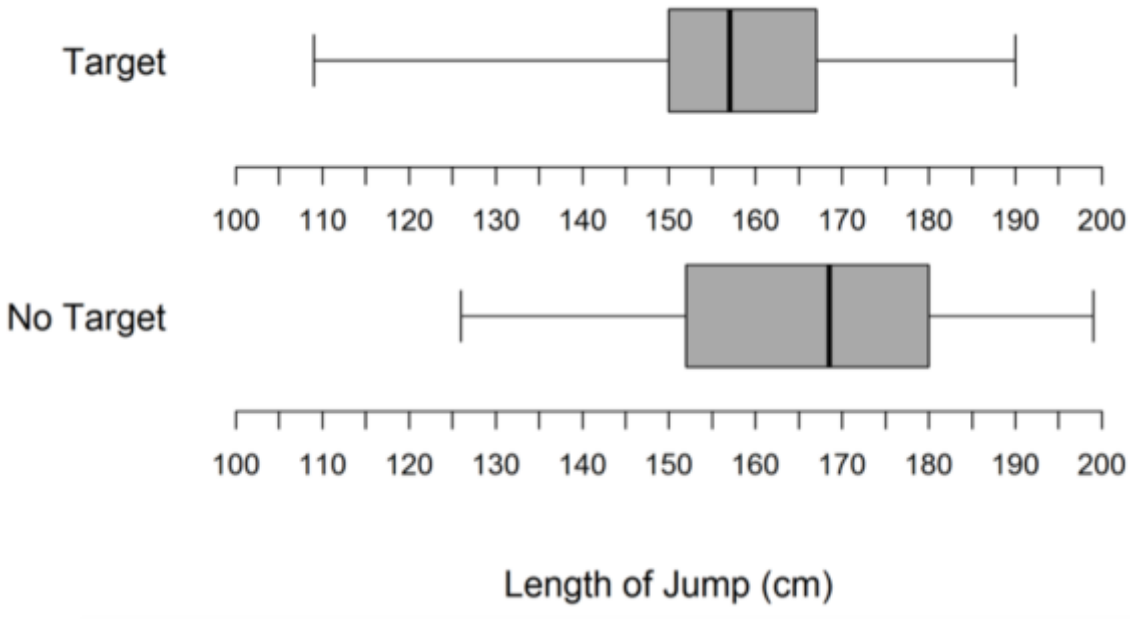


Figure 4.6. Boxplots provided on the Jumping Distances task. Released item from the LOCUS assessment (Jacobbe, 2016).

In general, teachers' reasoning on these three sub-parts was coded as *acceptable*. This means that their claims and reasons to support those claims were sound, drawing on relevant measures of center, spread, or shape as evidence. For instance, when asked about the consistency of mile times on the New Year's Day Race task (see Figure 4.5), a typical response was to claim that the half-marathon runners' mile times were more consistent because "the range is smaller from six to14, and then they're kind of clustered together. Rather than these [5K histogram] being spread out over a much bigger range" (Ellie, Line 16). Most teachers offered both a more generic sense of spread, such as Ellie's use of the phrase "clustered together", and a quantitative measurement such as the range. All of the teachers drew on the range.

Due to the lack of individual data values on the New Year's Day Race task, teachers were unable to calculate any measures of spread (even the range has to be approximated). However, some teachers indicated that a standard deviation could be

approximated or that they would like actual data values in order to carry out the calculation and be more precise with their comparison. Tim was the only teacher to attempt an approximation, drawing on his formal knowledge from calculus-based statistics:

> You could estimate the standard deviation on both [marks estimate of standard deviation distances from a supposed center mark on both plots] of them by finding the center and looking for the inflection point of the curve [draws bell shaped curve over plots and marks inflection point]. So I would say the standard deviation is about 2 on the half marathon and about 3 on the 5K [marks distances on each]. (Line 10).

Although reasoning with an approximate standard deviation was not anticipated, it is worth noting that only two teachers (Tim and Amalia) attempted to estimate a standard deviation despite seven of the teachers providing evidence that they viewed standard deviation as a measure of spread from center. Teachers were much more likely to draw on range as a measure of spread, rather than considering the spread from center. This is an example of some qualitative differences in the types of acceptable reasoning teachers offered.

Nearly all teachers offered at least one unacceptable reason, and at most two reasons, to support a claim. This is likely due to the interview protocol that explicitly attempted to exhaust all possible arguments teachers believed had validity. The majority of unacceptable reasoning came from New Year's Day Race part (b) and Jumping Distance part (a). On Jumping Distance part (a), three responses were coded as *unacceptable* ways of reasoning about the shape of a distribution by drawing on physical

characteristics of the boxplot—usually describing center or spread instead. For instance, Ellie reasoned from ideas of spread to describe a physical feature of the plot, that "this [distance from Q1-Q3 on no target plot] is bigger than that [target plot]" (Line 33)—informally describing the inter-quartile range. Similarly, Ruby reasoned with ideas of center and spread that "50 percent of the jumps without, I can say it this way, without a target, are greater than 75 percent of the jumps with a target" (Line 40), thus using sub-ranges to aid in describing general locations—a general sense of shifted centers. However, these teachers all explicitly stated that they were describing the shape.

On part (b) of the New Year's Day Race task, teachers were asked to respond to the following prompt:

> Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not. (Jacobbe, 2016)

Five teachers gave *unacceptable* reasons, one of which (Kathy) did not state an acceptable reason at all and all of the others offered at least one acceptable reason. Three of these responses reasoned with specific points, rather than considering what may be typical. For instance, Mike reasoned:

> Just looking right there [places pen vertically at half-marathon plot max across both plots], I mean, all of the marathon runners were done, or they, you know, would have had a max of 15, or 14ish, whereas up here [5K plot] there was a least a few people that went well above that [gesturing to far right bars of 5K plot]. (Line 61).

In this excerpt, notice that Mike is drawing on a comparison of extreme values in order to describe which group's mile times were greater "on average." Kathy and Ellie similarly reasoned that the 5K mile times were higher on average by comparing the maximum or minimum values. The other two teachers' responses (Amalia and Michaela) were *unacceptable* because their reasoning was based on *undesirable* knowledge elements that led them to conclude that the half-marathon runners' times were *higher* than the 5K runners' times. Michaela reasoned that skewed right distributions *decrease* the mean, thus resulting in the flawed argument that 5K runners' times are less than the half marathon runners' times. Amalia assumed that the bars on the far right of the 5K plot could be ignored because they were outliers and, she reasoned, someone "could explain that this [gesturing to 5K plot] could have a lower mile time, if that makes sense" (Line 48). Even if those bars were found to be outliers, they represent about 8% of the data and should not be disregarded. Furthermore, even if ignoring those bars was justifiable, the center of the distribution of the 5K runners' mile times is still higher, albeit closer, than the center for the half marathon runners' times.

When comparing center, spread, and shape of distributions in a non-IIR context, teachers were largely successful. However, nearly all teachers offered both *acceptable* and *unacceptable* forms of reasoning, indicating that their reasoning was weak. This is not surprising given the teachers' limited content knowledge of statistics.

Reasoning with a simulated sampling distribution. The fourth opportunity to reason in a non-IIR context presented to teachers was on part (a) of the Extended School Day task (see Figure 4.7). On part (a), participants were told that a school teacher wondered if 30% of the students at her school would favor an extended school day as was

found in a national study. A random sample of 50 of the 1200 students at her school showed that 24% favored an extended school day. The prompt then stated that the teacher wondered if this difference from 30% was surprising and if her school's proportion is less than 30%. A description of a simulation was provided (see Appendix F) "to see what values of the sample percentage would be expected if the school percentage was 30", and then teachers were presented the dotplot in Figure 4.7 and given the following prompt:

> If the school percentage were actually 30%, how surprising would it be to see a sample percentage of 24% or less? Justify your answer using the dotplot. (Jacobbe, 2016)

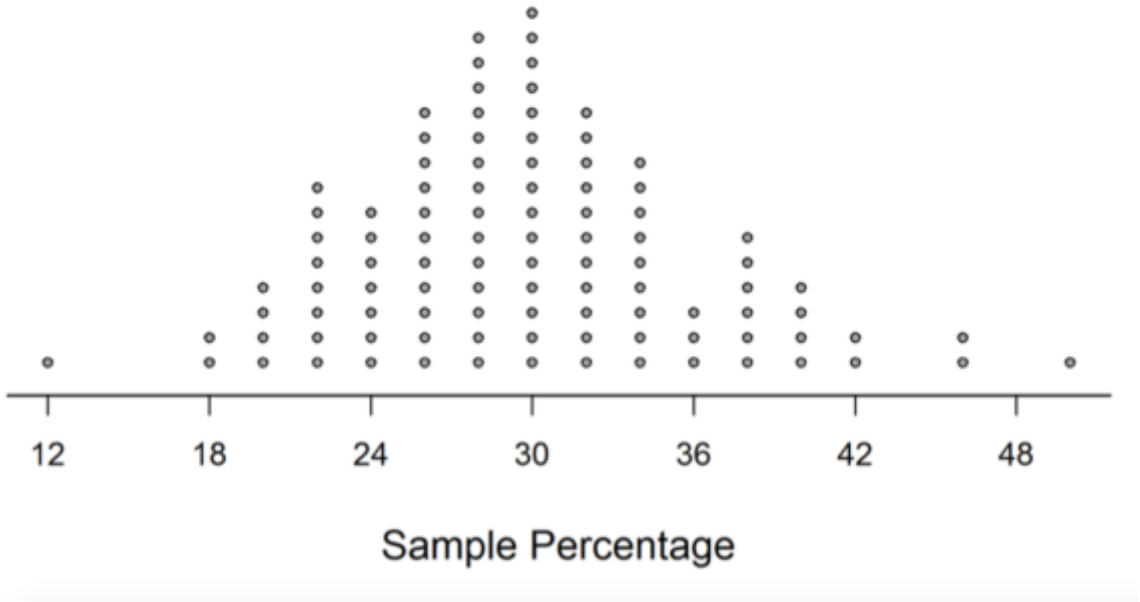


Figure 4.7. Dotplot provided on the Extended School Day task. Released item from the LOCUS assessment (Jacobbe, 2016).

All but Kathy provided at least one *acceptable* reason, and four teachers, namely Ellie, Michaela, Ruby, and Amalia, also offered at least one *unacceptable* way of reasoning. Most of the *acceptable* responses claimed that it would not be surprising, reasoning that "There's 100 of these guys, and there's [counting dots], so 22 out of 100

were 25, 24 percent or less, so that's 22 percent. So, 22 percent were that [draws arrow pointing to where she wrote "24% or less"]. So I don't think that's that surprising" (Ruby, Line 42). A select number attempted to estimate a standard deviation and use the 68-95-99.7 rule to identify that within two standard deviations accounts for 95% of the sample percentages and only things beyond that would be considered surprising. Therefore, Harrison and Tim claimed that it was not surprising "if I'm estimating where that one standard deviation below is [places mark at about 24]" (Harrison, Line 10). Other *acceptable* reasons were more general but reasoned from the dotplot that either 1) there were a lot of dots less than or equal to 24% or 2) that 24% is relatively close to an approximated mean—indicating a measure of spread to describe unusualness.

In contrast, of the eight *unacceptable* ways that five teachers reasoned, six were based on claiming that a sample percentage of 24% or less *is surprising* because "the majority of the time it's higher than 24 percent" (Kathy, Line 8) or because "it's not within your middle 50 percent of the population" (Michaela, Line 62). The remaining two were a result of misinterpreting the question.

In summary, although teachers were largely successful at constructing data-based arguments with sound reasoning, the majority also offered arguments that did not employ sound reasoning. In attempt to categorize types of reasoning in non-IIR contexts, three possibilities that could be useful are—*mostly acceptable*, *mostly unacceptable*, and *mixed* (see Table 4.2). Teachers in the *mostly acceptable* category were observed to engage in *acceptable* forms of reasoning in at least 75% of their responses. This group contained five teachers—Tim, Rosalynn, Mike, Michaela, and Harrison. Tim and Rosalynn were not observed to make any flawed arguments. On the other end of the spectrum, Kathy

was the only teacher observed to engage in *unacceptable* reasoning more often (71%) than *acceptable* forms of reasoning, and therefore she is the only teacher in the *mostly unacceptable* category. Ruby, Amalia and Ellie comprised the *mixed* group, who were observed to reason in *unacceptable* ways for between 50% and 33% of their responses. Moreover, teachers in the *mostly acceptable* category, as a whole, account for 22% of the total number of *unacceptable* forms of reasoning, while the *mixed* group accounts for 50% and Kathy accounts for the remaining 28%. The next section investigates this further by reflecting back on teachers' knowledge structures.

Table 4.2
Reasoning Categories in Non-IIR Contexts

Mostly Acceptable	Mixed	Mostly Unacceptable
Tim	Ruby	Kathy
Rosalynn	Amalia	
Mike	Ellie	
Michaela		
Harrison		

Knowledge structure support. Considering the types of knowledge structures that teachers were observed to have, one pattern is apparent (see Table 4.3). All teachers with *desirable-connected* knowledge structures (Tim, Rosalynn, and Mike) were also categorized as engaging in *mostly-acceptable* ways of reasoning in non-IIR contexts. Moreover, Kathy was observed to have an *undesirable-disconnected* knowledge structure and was also categorized as *mostly unacceptable*. However, Amalia, who also was observed to have an *undesirable-disconnected* knowledge structure was observed to have a *mixed* type of reasoning within non-IIR contexts—along with Ellie and Ruby, who have *undesirable-connected* knowledge structures. It is of note that even though Amalia's knowledge structure was categorized as *undesirable-disconnected*, she had constructed

multiple *desirable* knowledge elements of both center and spread, and three of her six instances of *acceptable* forms of reasoning occurred on the New Year's Day Race task—which was isolated to thinking about spread. The other two teachers observed to have *undesirable-connected* knowledge structures (Harrison and Michaela) were also categorized as engaging in *mostly acceptable* reasoning within non-IIR contexts. As mentioned previously, multiple tasks offered opportunities to reason with one knowledge element at a time, and therefore, making connections between knowledge elements was not always a necessary feature.

It is of note that the conceptual framework I am drawing upon does not assume that any particular knowledge is a necessary prerequisite for informal reasoning. However, a clear pattern has emerged that highly integrated knowledge structures of *desirable* knowledge elements is associated with substantially more consistent patterns of *acceptable* reasoning than knowledge structures that are largely disconnected with many *undesirable* knowledge elements.

Table 4.3
Types of Non-IIR Reasoning and Knowledge Structures

Reasoning Category	Teacher	Knowledge Structure Type
Mostly Unacceptable	Kathy	Undesirable-Disconnected
	Amalia	
Mixed	Ellie	Undesirable-Connected
	Ruby	
	Michaela	
	Harrison	
Mostly Acceptable	Mike	Desirable-Connected
	Rosalynn	
	Tim	

Types of Informal Inferential Reasoning

Three of the task sub-parts and the Tomatoes and Fertilizer task, which did not have multiple parts, were expected to involve *informal inferential reasoning* because they included an *inference* to respond to or called for an *inference* to be made. Moreover, teachers' responses were coded according to the three components of IIR (Makar & Rubin, 2009) and are referred to as follows:

- *Inference Component*: making or criticizing inferential statements that *generalize beyond* the data or imply *causality*
- *Data Component*: using or criticizing *evidence from data* to support an inference
- *Uncertainty Component*: using or criticizing *probabilistic language* that recognizes the *uncertainty* of the inference

Across the four different opportunities, part (b) of the Jumping Distances task encouraged engagement in all three components. The Tomatoes and Fertilizer task was intended to focus more on criticisms of the evidence being used (*data component*) to support a provided inferential statement (*inference component*), with a door open to criticize the lack of probabilistic language (*uncertainty component*) in the provided inferential statement. Part (b) of the Extended School Day task was intended to focus on the use of evidence from data (*data component*) to support, or not, a provided inferential statement (*inference component*). Moreover, part (a) encourages the calculation of a simulated *p*-value to be used in part (b)—thus providing the means for an assumed *uncertainty* in any inference if this value is used (as expected) in part (b). Therefore, the *uncertainty component* is only discussed for this task when there was an *unacceptable* form of reasoning with it. The last opportunity, part (c) of the New Year's Day Race task, was

limited to criticisms of the possibility of an inferential statement that generalizes beyond the data. Although these were the intentions of the tasks, *unacceptable* ways of reasoning within these three components was observed both with the intended components and with non-intended components of IIR. The next couple of sections will first examine the ways in which teachers reasoned in *acceptable* ways within and across components, and then the ways in which they reasoned in *unacceptable* ways within and across components. Afterward, possible deviations from the observed patterns in teachers' informal reasoning in non-IIR contexts from the previous section will be explored before reflecting on ways that teachers' knowledge structures may provide support for teachers' IIR.

Acceptable forms of reasoning with IIR components. In general, teachers rarely engaged with IIR in *acceptable* ways. Across all nine teachers and all four opportunities, there were 25 instances of *acceptable* forms of IIR—less than one per teacher per task, on average (recall that teachers offered multiple responses per sub-part). This is just over half of the number of instances of *acceptable* forms of reasoning in non-IIR contexts ($N = 47$), of which there was the same number of opportunities to engage with. All teachers except Amalia were observed to engage in an *acceptable* form of IIR at least once. To describe how teachers engaged with multiple components of IIR simultaneously, the eight *acceptable* responses on the Extended School Day task will be excluded because, by task design, only the *data component* was a possible point of variation among *acceptable* responses. Of the remaining 18 instances of *acceptable* IIR, one involved all three components (Rosalynn), 13 involved both the *inference component* and the *data component*, and the remaining four involved only one component. Moreover, the 13 instances involving at least 2 components account for all teachers

except Amalia, who was not observed to engage in IIR in an *acceptable* form on any of the four tasks. Note that probabilistic language (*uncertainty component*) was only provided in an *acceptable* form of reasoning with the one instance of an argument using all three components. The *inference component* and *data component* occurred with similar frequency overall. Before moving into further details, it should be disclosed that Tim was the *only* teacher who provided an *acceptable* form of IIR for both the New Year's Day Race part (c) task and the Tomatoes and Fertilizer task. Therefore, these will not be discussed here. There will be further details about reasoning on those tasks in the section regarding *unacceptable* IIR.

A majority of the *acceptable* informal inferential reasoning instances, 16 of the 25, occurred when responding to the Jumping Distances task, with representation from all teachers except Amalia. Moreover, all teachers except Amalia provided at least one response incorporating the *inference* and *data* components, 13 of the 16 total instances included two components, and one of the 16 included all three. As discussed in the previous section on reasoning in non-IIR contexts, the Jumping Distances task provided a more familiar context for teachers because of its close alignment with middle grades content. Part (b) (see Figure 4.6 for boxplots) of the task is also within the context of middle grades standards. The prompt says:

> Write a concluding statement to address whether the distances the male students jumped were affected by having a target. Justify your conclusion. (Jacobbe, 2016)

In order to be coded as *acceptable*, a response had to include reference to at least one of the following: (1) the center for those jumping with a target is lower than those without a target, (2) the range of the target group is higher or the IQR of the target group is lower,

or (3) a large amount of overlap between jumping distances. Among the 16 *acceptable* instances of IIR, three "conclusions" were made: 1) having a target resulted in shorter jumping distances, 2) having a target resulted in more consistent jumps, or 3) there was no observable effect of having a target, or that the difference between groups was not meaningful. As previously noted, there was some variability in which components were observed across responses.

Only Rosalynn offered an argument that incorporated all three components of IIR. She first stated, "Yes, I think there might be an effect of having a target. And I think in this particular case it might have caused them to jump shorter" (Line 13). In this claim, she is not simply stating a difference between the two groups, but is making a prediction about *causality*, even using the word "caused"—evidence of the *inference component*. This is in contrast to the claim, "When given a target, if that's the target [pointing to median of target boxplot], then they, then the middle 50 percent were a lot more consistent meeting that target, than not given a target" (Kathy, Line 34). This statement by Kathy does not clearly indicate an inference that the target may have *caused* the difference in consistency of the jumping distances, but rather points to a direct comparison of the spread of each group without making an inferential statement. Therefore, Kathy's response was not coded as involving the *inference component*.

Other arguments that involved the *inference component* were not as clear but still implied a *causation*, such as Harrison who said "people jump farther by having a target" (Line 30) and Ellie who said "I would say that having a target made the boys jump, their jumps weren't as long when they had a target as when they didn't" (Line 46). Causality is

being inferred in Harrison's statement through the use of the phrase "by having" and in Ellie's statement through the use of the word "made."

Rosalynn's inferential statement implying *causality* also includes evidence of the *uncertainty component* through her use of the probabilistic language inherent in the word "might." This use of language is not seen in Harrison's or Ellie's responses above. For instance, Harrison claimed that "people jump farther" and Ellie claimed that "having a target made the boys jump." Neither of these statements indicate that there is a level of *uncertainty* in the *causal* claim they made. Inclusion of the *uncertainty component* would require a softening of the *causal* claim with words like "might", as Rosalynn did, or, for example, stating "people [probably] jump farther" or "having a target [likely] made the boys jump." It is important to note that this is also different from use of similar language that does not indicate a degree of uncertainty in a *causal* claim but that recognizes the variability in the data. For instance, on part (b) of the New Year's Day Race task, Ruby makes a non-inferential statement that "I think these guys [5K runners' times] are mostly slower" (Line 55). The use of the word "mostly" recognizes that it is not possible to claim they are *all* slower because there is variability in run times. This kind of *uncertainty* is connected to a direct comparison of samples, rather than the type of *uncertainty* involved in an inferential statement.

Going back to Rosalynn's argument, she made use of *data as evidence* in order to support her *inference*. The entirety of her argument reads:

> Based solely on the data, I find it interesting that at least half of those with no target were able to jump farther than 3/4 of those with the target. So that big of a difference, we get 25 percent of them that are meeting or exceeding, beyond, that

there might be some sort of influencing factor of having that target fixed ahead of time. 'That target was far enough so I hit the target, I'm good', when they could have jumped farther. So, yes, I think there might be an effect of having a target. And I think in this particular case it might have caused them to jump shorter. (Rosalynn, Jumping Distance, Line 13)

Looking at the *data component* in isolation, Rosalynn clearly draws on data by comparing sub-ranges of the data to reason that those without a target appear to have jumped further than those with one: "at least half of those with no target were able to jump farther than 3/4 of those with the target" (Line 13). However, she continues by reasoning that such a difference implies that having a target "might have caused" this difference. This connection of the data to her inferential statement is clear from her use of the word "so" when she says "So, yes, I think…" (Line 13). Therefore, Rosalynn has incorporated all three components into her argument.

It is important to recognize that Rosalynn's *inference* was constructed from the evidence from the data that supported it. A different observation from the data could have led to a distinctly different *inference* regarding the effect of the target. In Rosalynn's case, evidence relating to the difference in a general sense of center supported the notion that a target *causes* shorter jumping distances. Although many teachers made a similar *inference*, without the *uncertainty component*, by drawing on evidence from differences in the medians or differences among all of the five number summary values, some teachers did not. For instance, Ruby stated that "the target did have an effect, because the lengths were more consistent. Or what? Have less variability. … they are more consistent when they have a target" (Line 110). Ruby provides evidence of a *causality inference*

("the target did have an effect" … "they are more consistent") that is a different kind of *inference* from those presented thus far because she is attending to the spread rather than to the center of the data. Thus, the *data component* appears to drive the *inference component*.

Interestingly, a third *inference* was made by considering these same types of data. For instance, when prompted, Kathy reasoned that there was no effect because "the range isn't that much bigger. It's only five centimeters different, so I guess if a kid was looking at that then they could assume that the range wasn't that much effected. So, they could claim that" (Line 45). As with Ruby's statement, the evidence rested on a comparison of the respective spreads in jumping distances. The difference is that Kathy's reasoning involved a comparison of the *size* of that difference in spread rather than whether a difference was observed. It should be noted that Kathy considered the range, while Ruby compared the inter-quartile ranges—which was, visually, a more distinct difference than the ranges (see Figure 4.6).

Part (b) of the Extended School Day task presented a different way of reasoning with *data as evidence*. A total of seven of the 25 *acceptable* forms of reasoning in an IIR context came from the Extended School Day task, perhaps due to less familiarity with the context. These seven responses came from four of the teachers: Rosalynn, Mike, Tim, and Michaela. See Appendix F for the task in its entirety. After presenting the dotplot in Figure 4.7 and having teachers consider how unusual a sample proportion of 24% might be in part (a), part (b) stated:

> Based on her sample data, should Stella conclude that the percentage of students at the school who favor an extended school day is less than 30%? Explain why or why not. (Jacobbe, 2016)

Two types of responses were offered. One type was to reason from the proportion of samples less than 24% found in part (a), as can be seen in Mike's response below (who also mixed in a bit of formal language from hypothesis testing):

> So, based on her sample should Stella conclude that the percentage of students at her school who favored an extended school day is less than 30 percent? I would currently say no, with a sample this size, it's not uncommon to see a percent as low as 24 percent. In fact, we already showed that there was about a 22 percent chance of seeing that, so. Based on that we would fail to reject the null hypothesis. (Mike, Extended School Day, Line 35)

In this excerpt, Mike is observed to draw on *data as evidence* ("it's not uncommon to see a percent as low as 24 percent", "a 22 percent chance of see that") in an *acceptable* form. It is important to note that the manner in which he reasoned from the evidence—recognizing that the dotplot's purpose was to rule out the possibility of the sample occurring by luck—provides the *acceptable* support for his claim.

The other 3 teachers offered a similar response that drew on the simulated *p*-value, without calling it that. Among the other three of the seven responses, Mike and Michaela suggested that a formal statistical test could be carried out since the actual data values are provided in the dotplot, and Rosalynn used formal ways of reasoning by considering the 22% simulated *p*-value to be "greater than any particular alpha level"—

using that as the evidence to support her claim that Stella should not conclude that "the true percentage is less than thirty percent" (Line 58).

In summary, when provided the most familiar context for engaging in IIR, all teachers except Amalia engaged in the *inference component* and *data component* simultaneously. Moreover, all teachers except Ruby and Kathy provided either at least two different forms of evidence as support for their *inference* or they provided two different *inferences* that involved both components in *acceptable* ways. Ruby only provided one response, and in Kathy's second response it was not clear that she had made an *inference*. Moreover, only Rosalynn engaged in all three components, and she was the only teacher to use *probabilistic language* to indicate a level of *uncertainty* in her *inference*.

Unacceptable forms of reasoning with IIR components. The majority of responses to the four tasks that encouraged IIR were coded as *unacceptable* forms of reasoning. A total of 59 instances were coded as *unacceptable*, in comparison to the 25 *acceptable* instances of IIR—more than twice as many. Furthermore, the frequency of *unacceptable* instances of IIR per person ranged from three to 10, as opposed to a range of zero to five for *acceptable* instances. Every teacher except Tim was observed to reason in *unacceptable* ways more frequently than *acceptable* ways. Even so, Tim was observed to reason in four *unacceptable* ways and five *acceptable* ways. In relative terms, the percentage of total responses to IIR tasks that were coded as *unacceptable* ranged from 38% (Rosalynn) to 100% (Amalia), with Rosalynn and Tim being the only teachers to be observed reasoning in *unacceptable* ways in IIR contexts for less than 50% of their responses. Four of the teachers (Amalia, Kathy, Ellie, Ruby) offered *unacceptable*

reasoning on more than 75% of their responses. Disregarding the Extended School Day task for reasons described previously, among the remaining 47 instances of IIR coded as *unacceptable*, one included no component, 44 included one component, and 3 included two components—2 of which were provided by Michaela.

Before describing patterns in reasoning, it is useful to also consider that among the four opportunities to engage in *informal inferential reasoning* (IIR), all teachers were observed to have not provided nor recognized an appropriate argument on at least one of the four tasks intended for IIR. Moreover, all teachers except for Tim and Ruby were observed either making contradictory statements when constructing their own arguments, making contradictory statements between their own arguments and what they identified as appropriate and inappropriate student arguments, or between student arguments.

Because part (b) of *the* Jumping Distances task was the most accessible in terms of engaging in IIR, it is perhaps a representation of teachers' *informal inferential reasoning* in a best-case scenario. To compare with the overall description, five teachers offered seven of the 59 instances of IIR coded as *unacceptable* on part (b) of the Jumping Distances task—the lowest frequency of all four IIR tasks. These five teachers were Ruby, Amalia, Michaela, Mike, and Tim—Tim and Mike each offering two *unacceptable* arguments and the rest offering one. All teachers except for Amalia also provided either one or two *acceptable* arguments for the task, yet Amalia recognized a provided *acceptable* student response as being appropriate. Moreover, all teachers at least recognized an *acceptable* argument for this task (the only task for which this was the case).

Two teachers, Mike and Michaela, reasoned in ways coded as *unacceptable* while also incorporating both the *inference component* and the *data component.* Mike and Michaela both reasoned in this way on part (b) of the Jumping Distances task. Both instances were categorized as *unacceptable* for similar reasons. Both made inferences that having a target *caused* shorter jumping distances, but with inadequate or inappropriate support. For instance, Mike's supporting evidence from data (*data component*) was that "the graphs just are, just plain shifted. The no target group is shifted to the right compared to the target group" (Line 73). Although this statement indicates drawing on a general form of center, it was coded as *unacceptable* because the evidence is very weak—not even specifying, for instance, what is meant by "to the right." Somewhat similarly, Michaela's supporting evidence from data (*data component*) was "because overall the variation given by having a target was spread out so there was less consistency jumping when a target was given" (Line 48). Although this would be *acceptable* supporting information for a claim about the consistency of jumping distances, the inference (*inference component*) that was made was about the effect on jump distances and not the consistency of them. Therefore, the evidence from data (*data component*) did not support the claim.

For comparison, Michaela's second instance involved the use of both the *data component* and the *uncertainty component* on the Tomatoes and Fertilizer task. On this task, the main stem (see Appendix F) stated that a farmer had randomly assigned tomato plants to receive either a new fertilizer or an old fertilizer—all under the same growing conditions. The task then revealed that tomatoes in the new fertilizer group had a mean weight of 0.4 ounces heavier. The task then says:

> Based on the results, the farmer is convinced that the new fertilizer produces heavier tomatoes on average. Briefly explain to the farmer why simply comparing the two means is not enough to provide convincing evidence that the new fertilizer produces heavier tomatoes. (Jacobbe, 2016)

In order to be assigned an *acceptable* code for reasoning on this task, responses had to address the main issue—namely, that the farmer had not addressed the possibility of sampling variability explaining the difference in the means. This could be demonstrated through recognition of the possibility that the random assignment, by chance, led to some of the more fruitful plants being placed in the treatment group. Many *unacceptable* arguments were given (outlined later in this section) that did not address the primary issue of sampling variability. For example, in her *unacceptable* argument, Michaela argued that the difference in the means was not enough evidence because

> You would have to look at the possibility of error, the confidence interval, talking about what percentage of confident you are that they'd be, you couldn't just outright say that it's going to be, but you need to actually calculate, you need to talk about how confident you are but you can't outright say it works every time. Because I do feel like you have to calculate in error, chance. (Michaela, Tomatoes and Fertilizer, Line 25)

This response is *unacceptable* because although inclusion of the confidence interval could allow for deciding whether the observed difference in the means was likely due to sampling variability, Michaela's response is more to focused on the farmer's lack of probabilistic language (*uncertainty component*) to indicate uncertainty ("you can't outright say it works every time"). To verify this, I asked her, "what extra information

would such a thing like a confidence interval allow you to say?" (Researcher, Line 26). Notice that her response below is not describing a confidence interval or indicating a concern with ruling out chance, but rather she is concerned with 1) comparing variabilities and 2) being less deterministic in the inference:

> Well the thing is, I feel like you would need the standard deviation for the weight. Because if you're talking about the averages, again you could have an extremely fat, random tomato that boosted an average and therefore it doesn't work. So, you really need to look at more than just the average, which is where the confidence, if I'm remembering correctly, I think that's where the confidence interval can help you talk about okay here's my standard deviation, here's where my average, or my middle was, and so here's where I'm actually confident that this would happen every time, where these tomatoes would be heavier than the other ones. And so for that purpose, I think you can't just say that because the two means are, one's bigger, again you could have a super fat tomato and a super unhealthy one as an outlier in the other group, but the majority of the tomatoes could be the same weight. So, you couldn't just actually say that the two means were evidence that one beats the other one. (Michaela, Tomatoes and Fertilizer, Line 27).

As seen above, although Michaela mentions the confidence interval, further inspection revealed that her reasoning that drew on the confidence interval was more generally related to comparing variability. However, regardless of this, her response still involved the *data component* because it criticized the evidence provided by the farmer and suggested that another type of evidence "the confidence interval" or the "standard deviation" be provided. Therefore, this response involved both the *data component* and

the *uncertainty component*, but was *unacceptable* because the reasoning involved would still leave the farmer with inadequate supportive evidence.

All other instances of *unacceptable* IIR involved a single component—either the *inference component* or the *data component*. On the Tomatoes and Fertilizer task, 25 responses were coded as *unacceptable* and all teachers were represented. Reasoning involving the *inference component* were coded as *unacceptable* ($N = 7$) because they criticized the farmer's causal inference (*inference component*) by suggesting changes to the study design that would still not lead to sufficient evidence—moreover, the study design as described in the task does allow for a causal inference. These reasons included 1) increasing sample size, 2) replicating the experiment, and 3) including a control group. *Unacceptable* reasons involving the *data component* ($N = 16$) criticized the farmer for only using the mean, but in a similar way, suggested different evidence that would still not be sufficient—that is, evidence that would still not rule out the possibility of the observed difference being just due to chance. These reasons *all* included one or both of 1) suggesting other measures of center or spread, or 2) pointing out that the mean is sensitive to outliers and isn't the best measure of center. Five teachers did not generate or recognize an *acceptable* reason—Kathy, Michaela, Harrison, Ruby, and Mike. When Michaela, Harrison, and Ruby were presented with the *acceptable* reason in the form of a possible student response, all three of them identified it as an *inappropriate* justification.

For broadly similar reasons, 15 of the 16 responses to part (c) of the New Year's Day Race task were coded as *unacceptable*, and all teachers but Tim were represented. In order to be assigned a code of *acceptable*, responses had to explicitly point out that runners having the option to choose which race to run is a study design flaw that prohibits

an inference. This could also be achieved through a reference to the lack of random assignment. The nine responses that criticized the intended *inference component* focused on either 1) irrelevant features of the task context or 2) changes to the study design that avoided explicitly recognizing the study design flaw—that runners were able to choose their own race; there was no random assignment. For instance, Rosalynn claimed that it would not be reasonable "to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K" (Jacobbe, 2016) because "those that ran the half marathon and not the 5K would most likely be in better shape" (Line 40). Although this could be a motivating factor for not allowing runners to choose their race, it only draws on contextual features—ignoring the study design flaw. Another common response was "the only way you could do this is if you literally had the same runners do both races and then just compared their times and see if they got faster" (Kathy, Line 56). Although this change would allow for such an inference to be made, it avoids pointing out the study's design flaw. Therefore, both were considered *unacceptable*.

Unacceptable responses that incorporated the *data component* all ignored the study design flaw and claimed it *was* a reasonable conclusion—drawing on differences in center to support it. For instance, Michaela drew on a general location of the median:

> If you just look at, I mean eyeballing where perhaps the median might be, the median time, like the half way marker for this one, the half marathon runners, looks like it would be lower than the half way time for this, the 5K runners. So, if that was the case, then you could say that they're, the half marathon runners,

could be lower than this one, than the runners in the 5K. (Michaela, New Year's Day Race, Line 118).

In this quote from Michaela, the median is used as a support to the *inference* provided in the task stem, thus not attending to the study design flaw that invalidates such inferences. As already stated previously, Tim was the only teacher who provided an *acceptable* response on this task.

Because part (b) of the Extended School Day task involved more advanced background in statistics—although still in the context of high school statistics standards—it resulted in mostly separating teachers into those who have taught a stand-alone statistics course and those who have not. Namely, Rosalynn, Tim, and Michaela did not provide any *unacceptable* forms of IIR—recall that Rosalynn, Tim, Mike, and Michaela were the only teachers who provided *acceptable* reasoning on the task. Because of the fairly drastic separation by content taught and the observed struggle teachers had in responding to the task during interviews, results of this task do not provide much indication of teachers' *informal inferential reasoning* in *unacceptable* ways. However, it is noteworthy that the *unacceptable* forms of reasoning can broadly be described as:

- Misinterpreting the dotplot: "I think if anything that she would conclude that it's just what the national newspaper thought. Right at 30. Although again, I would want to calculate the actual mean and median to actually verify that" (Kathy, Line 62).
- Not recognizing the role of the dotplot: "the dotplot represents beads and not actual students" (Ruby, Line 111)

- A focus on criticizing the study design: "she doesn't have the data from the entire school so she can't guarantee her data is accurate" (Ellie, Line 76)

In summary, attempting to categorize teachers' reasoning based on common patterns is complex because many of the teachers exhibited similar ways of reasoning—teachers typically gave both *acceptable* and *unacceptable* arguments, for instance. However, taking a similar approach to categorizing reasoning types as in the non-IIR section aids in observing a couple of common threads. Teachers whose reasoning was categorized as *mostly-acceptable* were those observed to exhibit *unacceptable* ways of reasoning in IIR contexts for less than 50% of their responses. Moreover, when the task prompted for a response to a provided *inference*, those instances of reasoning were balanced between criticizing and suggesting within both *data* and *inference* components. Teachers whose reasoning might be categorized as *more-unacceptable* were those observed to exhibit *unacceptable* ways of reasoning in IIR contexts for between 64% and 70% of their responses. This group did not have any other salient distinguishing features. Teachers' whose reasoning might be categorized as *mostly-unacceptable* were observed to exhibit *unacceptable* ways of reasoning in IIR contexts for between 75% and 100% of their responses. Moreover, the majority of the time, this group of teachers focused their *unacceptable* forms of reasoning on criticizing study designs (e.g., sample size, treatment assignment) rather than on suggesting possible ways to overcome the issues.

Comparing IIR, non-IIR, and possible supports from knowledge structures. Overlaying the two types of reasoning with the knowledge structure types indicates some possible broad characterizations of relationships between the two types of reasoning and between reasoning and knowledge (see Table 4.4).

Table 4.4

Comparing Reasoning Types and Knowledge Structures

Non-IIR Reasoning	Teacher	Knowledge Structure
Mostly Unacceptable	Kathy	Undesirable-Disconnected
Mixed	Amalia	
	Ellie	Undesirable-Connected
	Ruby	
	Michaela	
	Harrison	
Mostly Acceptable	Mike	Desirable-Connected
	Rosalynn	
	Tim	

IIR Reasoning
- Mostly Unacceptable IIR
- More Unacceptable IIR
- Mostly Acceptable IIR

For instance, focusing only on the reasoning types, at the extremes, there is more meaningful overlap so that Kathy (the only teacher in the *mostly-unacceptable* reasoning category for non-IIR contexts) was observed to have *mostly-unacceptable* reasoning in both non-IIR and IIR contexts. The same is true at the other extreme among *mostly-acceptable* reasoning types. The middle group is where things become slightly less distinct and defined. For instance, in non-IIR contexts, Mike, Harrison, and Michaela exhibited *mostly-acceptable* forms of reasoning, while in IIR-contexts, they exhibited *more-unacceptable* forms of reasoning. On the other end of the spectrum, Ruby, Ellie, and Amalia all switched from a *mixed* type of non-IIR to a *mostly unacceptable* type of IIR. All in all, the general categories hold, with the *mostly-acceptable* reasoning type shrinking from non-IIR to IIR contexts and the *mostly-unacceptable* reasoning type growing from non-IIR to IIR contexts.

It is noteworthy that although teaching experience in statistics was fairly similar among Tim, Rosalynn, and Mike, whose knowledge structures were all described as *desirable-connected*, Mike had the least amount of experience and he also shifted reasoning categories. Although a difference in background experience is not a common feature among Mike, Harrison, and Michaela, relative to their shift in reasoning type,

they each were observed to provide less *acceptable* reasoning types in IIR contexts (33%–50% less) and substantially more *unacceptable* reasoning types in IIR contexts (250%–600% more). Moreover, roughly half of each of their *unacceptable* reasoning types were observed on responses to the Tomatoes and Fertilizer task.

Examining the types of knowledge structures that may provide support for IIR, once again the opposite ends of the spectrum are solidly within seemingly aligned reasoning and knowledge element types. From Table 4.4 it appears that *desirable-connected* knowledge structures are associated with *mostly-acceptable* forms of IIR—namely, Rosalynn and Tim. Moreover, it appears that *undesirable-disconnected* knowledge structures are associated with *mostly unacceptable* forms of IIR—namely, Kathy and Amalia. Among teachers with *undesirable-connected* knowledge structures, both *more-unacceptable* and *mostly unacceptable* types of IIR appear. It is of note that the change from non-IIR types to IIR types resulted in a clearer association between knowledge structures and reasoning types. Moreover, teachers who were categorized as having a *mixed* type of reasoning in non-IIR contexts were all found to be categorized as engaging in IIR in *mostly unacceptable* ways.

Pedagogical Content Knowledge for Statistics

Overview

Teachers' pedagogical content knowledge (PCK) for statistics was categorized according to the four hierarchical levels described by Callingham and Watson (2011): *Aware*, *Emerging*, *Competent*, and *Accomplished* (see Table 3.4 and Table 3.5). Although Callingham and Watson (2011) found that these hierarchical levels are determined, in part, by the number and type of student responses that teachers generated, this was not the case for the teachers in this study. All teachers were found to generate multiple

appropriate and inappropriate responses both in general and across non-IIR and IIR intended sub-parts. Moreover, among nine teachers' responses to four tasks, only one instance across 36 was identified as *Emergent*—Kathy's responses on the Extended School Day task. Ellie, Kathy, Mike, and Michaela were observed to have substantial variance in PCK levels across tasks. All four of them were identified at both *Competent* and *Aware* levels of PCK across tasks. This was also true among task parts within non-IIR and IIR contexts (which will be detailed later). Therefore, rather than using Callingham and Watson's description of *Emergent*, since it was exceedingly rare to observe a teacher on any given task at this level, I will call the level *Competent/Aware* to reflect the wide range of PCK within the group.

Accomplished and Competent. Tim was the only teacher observed at the *Accomplished* level, and Rosalynn was the only teacher observed at the *Competent* level. Moreover, there was no variance *across* tasks—both Tim and Rosalynn were identified at these respective levels for all tasks. The most distinguishing characteristics of the *Accomplished* and *Competent* levels are suggesting a student intervention for an inappropriate student response that makes use of questions to lead the student or challenge his or her belief, or offers rhetorical questions or directly points the student toward a method or way of reasoning, respectively. For instance, on part (a) of the New Year's Day Race task (see Figure 4.5), Tim identified an inappropriate student response as stating that the half-marathon runners' times are lower because the half-marathon graph is "further left" (Tim, Line 104)—a response identified as a common misunderstanding on the LOCUS assessment (Jacobbe, 2016). Tim suggested the following way of interacting with such a student:

> So I would ask that student, how do you recognize that's further left? What are you looking at? And they might do something with their hand, asking well if you take this thing [placing fingers at min/max of 5K graph], you have to scoot left to do it [moves hand straight up and then left to align with half marathon graph], so they might be redirected to start paying attention to the balance point of the graph, or a measure of center, instead of being further left. (Tim, New Year's Day Race, Line 105).

In Tim's suggestion, he is intending to lead the student, according to him ("they might be redirected to…"), toward a "balance point" perspective of center with a question such as "how do you recognize that's further left?" These types of suggestions were typical for Tim, showing evidence of the *Accomplished* level across all tasks.

Rosalynn was the only teacher with a dominant PCK level of *Competent* across all tasks and contexts. At this level, her suggestions for students were noticeably different from Tim's. For instance, on part (a) of the New Year's Day Race task, Rosalynn offers the following suggested way of interacting with a student who claims that the 5K runners' times are more consistent because the bars are all close to the same height—a common student misunderstanding on the LOCUS assessment (Jacobbe, 2016):

> I would likely address with that student that the idea of consistency is not bad. That they're looking an aspect of the graph that is the same. But, if we're looking to compare the times, we're looking for how well can we get the times closer together rather than the heights of the bars. So, acknowledging an idea of consistent as being correct, but the application of that to be reversed. (Rosalynn, New Year's Day Race, Line 57).

In this suggestion, Rosalynn is more direct in pointing the student toward a specific way of reasoning than Tim's suggested intervention. For instance, by stating "we're looking for how well can we get the times closer together rather than the heights of the bars", she is directly pointing the student to his or her error, rather than using questions as a way to the lead student.

Competent/Aware. Four teachers were identified at the *Competent/Aware* level of PCK—Ellie, Kathy, Mike, and Michaela. A distinguishing characteristic of this level is that they were observed to sometimes provide more direct responses to students that addressed the misunderstanding in the student response. At other times, although they made suggestions for interventions, they did so without addressing the misunderstanding or offered a suggestion to address a misunderstanding that did not exist (i.e., not recognizing a student response as being appropriate). For instance, on part (a) of the Jumping Distances task, Ellie offered the following suggestion for a student who might be "thinking of center as the actual middle of the min and max" (Ellie, Line 107):

> I would ask them what each point, or what each line, what's the box and whisker plot mean? Like what are the different values represented in a box and whisker plot? And hopefully they would say something here [gesturing to median line on the target boxplot] about some sort of center. [pause] I might ask what center means in statistics versus in the real world." (Ellie, Jumping Distances, Lines 124–126)

In this suggestion, Ellie is using rhetorical questions in order to direct the student away from his or her way of thinking about center as a midpoint of the range, and toward the

measure of center that is explicitly found on the box. This type of suggestion was common on this task, and it was identified at the *competent* level.

On the other hand, on the Tomatoes and Fertilizer task, Ellie offered the following suggestion for a student who might claim, correctly, that the farmer cannot only use the difference in means as evidence because the difference may be due to chance—that by luck the farmer had better plants in the treatment group:

> I think that it would be kind of an easy way to get out of it. But, I think that it would be, if they said that, I might ask, well then how can he, if he were to do it again, how could he ensure that he doesn't do that, or how can he make it so that's not the case? Just to get them to think a little more. (Ellie, Tomatoes and Fertilizer, Line 105)

Although it may be the case that Ellie is using a rhetorical question in order to point the student toward disregarding this as a viable response, she is actually directing him away from an appropriate way of thinking. Therefore, this suggestion exhibits an *aware* level of PCK for statistics. These ways of seeming to exist at multiple levels of PCK for statistics were common within this group and therefore more difficult to characterize holistically.

Aware. In contrast to the teachers at the *competent/aware* level, the three teachers at the *aware* level (Ruby, Amalia, and Harrison) exhibited a high level of consistency in their PCK across tasks. However, on the Tomatoes and Fertilizer task, Amalia was identified at the *competent* level, and on the Jumping Distances task, Harrison was identified at the *accomplished* level. Nevertheless, their dominant operating level was solidly *aware* when looking across all tasks. As mentioned previously, a distinguishing

characteristic of the *aware* level is to not address the student misunderstanding, or to offer questions and comments that do not appear to be tied to the content. For instance, on part (a) of the New Year's Day Race task, Ruby suggested the following intervention for a student who reasoned that the 5K runners' times were more consistent because the bars were closer to the same height:

> I would ask them, well what does it mean to be consistent? What do you think that that means? And try to see if they can come up with what that word means in not math world. And then say, you know, just kind of talk them through it. That these are all the different times [gesturing to each bar of the 5K graph and along the x-axis], are these times [half marathon plot] closer together, or are these times [5K plot] closer together? And what makes you, you know, me, I would try and rephrase it like that. Trying to define consistent again and see if they can see this one [half marathon plot] being more consistent. I'd probably try to point out right away these guys on the end [far right bars on 5K plot], um, I'd probably point out this bar that's much smaller [bar between 10 and 11 mark on 5K plot] and see if I could just kind of lead them in this direction." (Ruby, New Year's Day Race, Line 154)

Although at first Ruby's response may seem to address the student's misunderstanding—considering the consistency along the vertical axis rather than the horizontal axis—ultimately, her response fails to address it. At the end of this suggestion, Ruby attempts to direct the student's attention to the bars on the far-right end of the 5K histogram as well as the bar "that's much smaller" between the 10 and 11 mark. However, drawing the student's attention to individual bars does not address the student's misunderstanding of

focusing on the *heights* of the bars—the student could very well continue to consider the heights of the individual bars Ruby pointed out. Therefore, her response is viewed as not addressing the misunderstanding of the student and therefore exhibiting an *aware* level of PCK for statistics.

Comparing PCK within IIR and Non-IIR Contexts

Because it is not assumed that teachers would have the same level of PCK for statistics within *informal inferential reasoning* (IIR) contexts as they would within non-IIR contexts, levels of PCK were disaggregated by each of these two contexts (see Table 4.5). In comparison to teachers' overall PCK levels for statistics, generally teachers demonstrated higher PCK levels within non-IIR contexts. This upward movement only occurred, however, with three teachers. Amalia and Harrison, whose overall PCK level was identified as *aware*, increased to the *competent/aware* level among non-IIR contexts, while Ellie was observed to increase from *competent/aware* to *competent*. All other teachers' levels of reasoning remained constant. This resulted in a narrowing of the *aware* level that left only Ruby remaining, and the *competent/aware* and *competent* levels both widened somewhat.

Table 4.5
PCK Levels Across Contexts

Teacher	Overall PCK Level	Non-IIR PCK	IIR PCK
Tim	Accomplished	Accomplished	Competent
Rosalynn	Competent	Competent	Competent/Aware
Mike	Competent/Aware	Competent/Aware	Competent/Aware
Kathy	Competent/Aware	Competent/Aware	Aware
Ellie	Competent/Aware	Competent	Aware
Michaela	Competent/Aware	Competent/Aware	Aware
Ruby	Aware	Aware	Aware
Amalia	Aware	Competent/Aware	Aware
Harrison	Aware	Competent/Aware	Aware

In comparing teachers' overall PCK levels to their levels within IIR contexts, teachers' levels were generally lower among IIR contexts. This downward movement from an overall perspective to IIR contexts was observed among five teachers, and movement occurred between all levels. Tim was observed to move from an overall level of *accomplished* to a *competent* level within IIR contexts—thus leaving no one at the *accomplished* level. Furthermore, Rosalynn was observed moving from *competent* to *competent/aware* (leaving Mike as the only teacher at the *competent* level) and Ellie, Kathy, and Michaela moved from *competent/aware* to *aware*.

Comparing across reasoning contexts, teachers' PCK levels were either observed to be constant across IIR and non-IIR contexts, or they were observed to be lower in IIR contexts. No teacher was observed to have a higher PCK level in an IIR context than a Non-IIR context—even when further disaggregating levels among the most familiar task contexts (New Year's Day Race and Jumping Distances). More specifically, two teachers (Ruby, and Mike) were observed to have the same level of PCK in both reasoning contexts, six teachers were observed at one level of PCK lower within IIR contexts (Amalia, Kathy, Harrison, Michaela, Rosalynn, and Tim), and Ellie was observed to be two levels lower within IIR contexts. The broad downward movement is perhaps most obvious when noting that six teachers were observed at the *aware* level within IIR contexts as opposed to one teacher at the *aware* level within non-IIR contexts. Further, Tim was observed to move from *accomplished* to *competent*, thus widening the lowest level and removing the highest level of PCK within IIR contexts.

There were also important differences in patterns observed in the consistency of PCK levels across tasks within each reasoning context. Within non-IIR contexts, teachers

at the *aware*, *competent*, and *accomplished* levels exhibited a high level of consistency across tasks. For instance, Rosalynn was observed at the *competent* level on two tasks and the *accomplished* level on one task, a difference of one level resulting in a dominant PCK level of *competent*. However, teachers at the *competent/aware* level were observed to have less consistency. For instance, Amalia, Michaela, and Mike were all observed at both *accomplished* and *aware* levels on different tasks, thus spanning the entire spectrum.

In contrast, among IIR contexts, a difference of two levels was observed between at least two tasks for all teachers in at the *competent* and *competent/aware* levels—less variance than was observed within non-IIR contexts for the *competent/aware* level. Moreover, among the six teachers at the *aware* level, two teachers were observed at the *aware* level for three tasks and the *accomplished* level for one task, thus spanning the entire spectrum. Ruby was observed to *only* be at the *aware* level, and the other three teachers were observed to span from *aware* to *competent*.

Possible Connections Between Knowledge, Reasoning, and PCK

Due to substantial differences in teachers' PCK between non-IIR and IIR contexts, analysis was disaggregated by these two contexts. From Table 4.6 it is evident that a more discernable pattern exists among PCK within IIR contexts and knowledge structure types than for PCK within non-IIR contexts. For instance, all knowledge structure types categorized as having *undesirable* knowledge elements for center, spread, and shape are associated with PCK at the *aware* level—which includes all teachers who were identified at the *aware* level. Moreover, all six of these teachers were categorized as engaging in IIR in *mostly unacceptable* or *more unacceptable* forms. Therefore, teachers in this study whose PCK for IIR contexts was at the *aware* level also engaged in *unacceptable* forms of IIR more times than not, and their knowledge structures contained

multiple *undesirable* knowledge elements for center, spread, and shape that were at least somewhat integrated into their structures. However, teachers whose knowledge structures were categorized as *desirable-connected* were identified at both the *competent/aware* PCK level for IIR contexts and as engaging in *more unacceptable* and *mostly unacceptable* forms of IIR. Tim is the only teacher who appears to fit solidly across these categorizations. However, a somewhat less constrained pattern emerged. Teachers in this study whose PCK for IIR contexts was at least the *competent/aware* level also were observed to have *desirable-connected* knowledge structures, yet engage in IIR with mixed results.

Table 4.6

Comparing PCK with Reasoning Types and Knowledge Structures

Non-IIR PCK Level	Non-IIR Reasoning Type	Teacher	Knowledge Structure Type	IIR PCK Level
Competent/Aware	Mostly Unacceptable	Kathy	Undesirable	Aware
Competent/Aware	Mixed	Amalia	Disconnected	Aware
Competent		Ellie	Undesirable	Aware
Aware		Ruby	Connected	Aware
Competent/Aware	Mostly Acceptable	Michaela		Aware
Competent/Aware		Harrison		Aware
Competent/Aware		Mike	Desirable	Competent/Aware
Competent		Rosalynn	Connected	Competent/Aware
Accomplished		Tim		Competent

IIR Reasoning Type
- Mostly Unacceptable IIR
- More Unacceptable IIR
- Mostly Acceptable IIR

Possible connections to non-IIR forms of reasoning are much less apparent. However, it seems to be the case for the teachers in this study, that PCK for statistics within non-IIR contexts is largely unrelated to teachers' reasoning or their knowledge structures. For instance, teachers whose knowledge structures were identified as *undesirable-connected* were observed to have PCK across three different levels ranging from *aware* to *competent*. Moreover, among teachers with *desirable-connected* structures, PCK ranges from *competent/aware* to *accomplished*. The two teachers with

undesirable-disconnected knowledge structures were observed to be at the *competent/aware* level of PCK.

Two teachers in particular have made pattern observation more difficult—Ruby and Ellie. Ruby was observed to be at the lowest level of PCK, *aware*, in all contexts, indicating a high amount of consistency. However, she was identified as having a *mixed* form of reasoning in non-IIR contexts and an *undesirable-connected* knowledge structure, putting her into a category with Ellie. Within an IIR context, Ellie and Ruby fit into a narrowly definable category. However, within the context of non-IIR reasoning, Ellie is observed to be two levels above Ruby in PCK for statistics. Moreover, Ellie's PCK within non-IIR contexts was observed to be highly consistent, ranging from *accomplished* to *competent*. In fact, on the task with the most familiar context (New Year's Day Race), the gap widens further with Ellie moving solidly into the *accomplished* level of PCK and Ruby remaining at the *aware* level. To contextualize the matter further, Ellie reported having taught a statistics unit within the context of a Pre-Algebra course for the prior three years to the study. By way of contrast, Ruby reported teaching a unit of statistics in the same context for the prior 10 years to the study. Therefore, a cohesive pattern between knowledge, PCK and reasoning within non-IIR contexts goes largely unobserved.

Summary

The findings from this study reveal that teachers' knowledge structures for center, spread, and shape of distributions can be distinctly categorized as *desirable-connected*, *undesirable-connected*, and *undesirable-disconnected*. The majority of teachers were observed to have *undesirable-connected* structures and the teachers with the most experience teaching formal statistics courses were observed to have *desirable-connected*

structures. This indicates that the teachers in this study who were teaching statistics as a unit within the context of a mathematics course had constructed multiple *undesirable* knowledge elements related to center, spread, and shape of distributions, many of which were highly integrated into their knowledge structures. Results of their *informal inferential reasoning* indicated that teachers with *desirable-connected* structures tended to also engage in mostly acceptable ways in both IIR and non-IIR contexts. Moreover, teachers observed to have *undesirable-disconnected* knowledge structures were observed to reason in non-IIR and IIR contexts in largely unacceptable ways.

There were mixed results of teachers' reasoning among those with *undesirable-connected* structures. In a similar manner, patterns amongst knowledge, IIR and PCK provide evidence that teachers with *undesirable-disconnected* knowledge structures who engage in IIR in largely *unacceptable* ways also have PCK for statistics within IIR contexts at the lowest level. Moreover, teachers with *desirable-connected* structures who were observed to engage in IIR in largely *acceptable* ways have higher levels of PCK for statistics within IIR contexts. Some mixed results are found among teachers with *undesirable-disconnected* knowledge structures, but all of them were observed at the lowest level of PCK for IIR contexts, and in general they engaged in IIR in more *unacceptable* than *acceptable* ways. These patterns did not hold for non-IIR contexts. Although all teachers were observed at higher levels of PCK in non-IIR contexts, there was wide variability within these levels with regard to both their reasoning and their knowledge structures.

CHAPTER 5: DISCUSSION

Over the past decade there has been increased global attention to statistics at the K-12 level, as observed in standards reform documents (Australian Curriculum Assessment and Reporting Authority, 2014; England Department of Education, 2014; Franklin et al., 2007; NGA & CCSSO, 2010). Moreover, these documents advocate for attention to informal inference, especially at the middle and secondary levels. Such calls have led to a global crisis for teachers because preparation programs continue to lag behind the demands. For instance, recommendations for teacher preparation in statistics changed significantly from the *Mathematical Education of Teachers* (MET) (Conference Board of the Mathematical Sciences, 2001) to the MET II (Conference Board of the Mathematical Sciences, 2012) due to the increased demands of CCSSM that included recommendations from the GAISE report (Franklin et al., 2007). These changes include a focus on informal inference, which previously had not been a focus. Echoing these calls, the *Statistical Education of Teachers* document (Franklin et al., 2015) continues to demand reform of teacher education programs regarding the experiences preservice and inservice teachers receive. Moreover, Banilower and colleagues (2013) found that although most middle level and secondary mathematics teachers in the US had taken one or two courses in statistics, most also lacked confidence in teaching the content required of them. In addition to the lack of coursework to develop pedagogical knowledge for statistics, this lack of confidence is also likely a direct result of the widespread documentation that introductory statistics courses at the post-secondary level are largely ineffective at improving teachers' statistical reasoning because of their likely focus on formal procedures rather than the reasoning process (Castro Sotos et al., 2009, 2007; delMas et al., 2007).

Currently, there is a dearth of research on how mathematics teachers engage in *informal inferential reasoning*, and how it may be related to their statistics knowledge and pedagogical content knowledge for statistics (e.g., Langrall et al., 2017). A deeper understanding of these relationships is needed in order to develop the necessary preparation courses to ensure that prospective and practicing teachers have the necessary skills to develop their students' statistical reasoning abilities in these new ways. Therefore, this study sought to contribute to this void by exploring the following research questions:

1. What knowledge structures do middle level and secondary mathematics teachers have, regarding center, spread, and shape of distributions? (RQ1)
2. How do teachers' knowledge structures support informal inferential reasoning? (RQ2)
3. What is the relationship between teachers' informal inferential reasoning and pedagogical content knowledge? (RQ3)

In the sections that follow, I provide a summary of the study and findings, a discussion of the findings, limitations of the study, implications for teacher education, and recommendations for future research.

Summary of the Study and Findings

Method

The participants included nine middle and secondary mathematics teachers who were recruited to achieve a stratified purposeful sampling method (Patton, 2002) with strata to represent teachers who had taught (a) an Advanced Placement (AP) (N = 2) Statistics, (b) a non-AP statistics course (N = 2), (c) at least one unit of statistics in the context of a high school mathematics course (N = 2), and (d) at least one unit of statistics

in the context of a middle level mathematics course (N =3). The *Goals and Outcomes Associated with Learning Statistics-2* (GOALS-2) assessment (Sabbag & Zieffler, 2015) was administered to gauge participating teachers' formal knowledge and to aid as secondary information for RQ1. To understand teachers' knowledge (RQ1) and their *informal inferential reasoning* (IIR) (RQ2), four released items from the *Levels of Conceptual Understanding in Statistics* (LOCUS) assessment (Jacobbe, 2016; Whitaker et al., 2015) were administered during task-based clinical interviews (Goldin, 1997). All four tasks were open-ended and ill-structured to encourage informal reasoning, and four of the eight total parts specifically included some form of IIR (Huey & Jackson, 2015; Zieffler et al., 2008). To examine teachers' pedagogical content knowledge (PCK) for statistics (RQ3), for all parts of the four LOCUS tasks, teachers were asked to (a) generate a list of possible student responses and to categorize them as appropriate or inappropriate, and (b) describe how they might intervene with a student who offered one of the generated inappropriate student responses (Callingham & Watson, 2011; Watson et al., 2008). When necessary, student responses were offered to teachers for consideration. Responses to these items sometimes offered supplemental data to aid in answering RQ1 and RQ2 as well.

Transcripts of video-recorded interviews were analyzed to identify what knowledge elements teachers were drawing on, coded as *undesirable* or *desirable*, and what connections teachers may have constructed between knowledge elements (RQ1). Knowledge structure maps were constructed (Groth & Bergner, 2013) to represent if knowledge elements were *undesirable* or *desirable* and arrows were drawn between elements for which there was evidence the teacher had related the elements in some way.

Maps of teachers' knowledge structures were categorized based on common characteristics found during cross-case analysis.

Teachers' informal reasoning was coded by first identifying if a claim was supported by *acceptable* evidence (Means & Voss, 1996). For the four parts that focused on IIR (RQ2), it was then identified whether the argument incorporated (a) an inference beyond the data or regarding causality, (b) data as evidence, and (c) probabilistic language to indicate a level of uncertainty in the inference (Makar & Rubin, 2009; Rossman, 2008). During cross-case analysis (Creswell, 2013), teachers' reasoning, within non-IIR and IIR contexts, was then categorized into broad types and these types were then compared with knowledge structure types to characterize possible ways teachers' knowledge supported their IIR.

In order to answer RQ3, the frequency of inappropriate and appropriate responses per task part was counted, and then suggestions for intervention were coded based on (a) whether the suggestion addressed the student misunderstanding, (b) whether the suggestion drew on the student's response, and (c) whether the suggestion was content-specific (Callingham & Watson, 2011; Watson & Callingham, 2014). Moreover, teachers' responses to each task part were categorized as one the following hierarchical levels: *aware*, *emerging*, *competent*, or *accomplished* (Callingham & Watson, 2011; Watson & Callingham, 2014; Watson et al., 2008). Dominant forms of PCK for each teacher were identified as either *aware*, *competent/aware*[3], *competent*, or *accomplished* overall and within both non-IIR and IIR contexts. Teachers' PCK levels were then

[3]The competent/aware category was created during analysis because the "average" of competent and aware, emergent, was not an observed PCK level—rather, teachers were observed at both competent and aware levels of PCK.

compared to the types of reasoning and knowledge structures previously identified in order to propose possible relationships to PCK levels.

Results of the Study

Knowledge structures. Responses to GOALS-2 and the LOCUS tasks indicated that teachers' knowledge structures were of three types: *desirable-connected* ($N = 3$), *undesirable-connected* ($N = 4$), and *undesirable-disconnected* ($N = 2$). *Desirable-connected* structures included almost no *undesirable* knowledge elements for center, spread, and shape of distributions and knowledge was highly interconnected, with connections observed both *within* and *between* knowledge types (center, spread, shape). This knowledge structure category was highly consistent across teachers—salient features were observed in all cases.

On the other end of the spectrum, *undesirable-disconnected* knowledge structures contained multiple *undesirable* knowledge elements that were observed to be largely disconnected, with no connections observed between center and shape knowledge types. These characteristics were consistent across *undesirable-disconnected* structures. Knowledge structures described as *undesirable-connected* also contained multiple *undesirable* knowledge elements, but knowledge elements were highly connected—thus integrating *undesirable* knowledge elements into the overall structure. Knowledge structures in this category were more varied than the other two types. For instance, not all structures included direct connections *between* all knowledge types, but all included direct connections *between* at least two of the three types and all included indirect connections *between* all knowledge types. Moreover, the majority of knowledge elements were connected to at least one other knowledge element—reflecting the highly integrated nature of teachers' knowledge.

Knowledge as support for IIR. Within non-IIR contexts teachers were observed to exhibit more *acceptable* forms of reasoning than within IIR contexts—the supports they provided for claims were adequate and based on *desirable* knowledge elements. Specifically, five teachers were observed to reason in *acceptable* ways in non-IIR contexts for at least 75% of their responses and two teachers were observed to *only* offer *acceptable* forms of reasoning (*mostly acceptable*), one teacher was observed to engage in *unacceptable* forms of reasoning for more than 70% of responses (*mostly unacceptable*), and the remaining were a combination of *acceptable* and *unacceptable* (*mixed*).

In contrast to non-IIR contexts, no teachers were observed to engage in IIR in *acceptable* ways for at least 75% of their responses. Teachers whose reasoning was described as *mostly acceptable* ($N = 2$) reasoned in *unacceptable* ways for less than 50% of their responses, reasoning described as *mostly unacceptable* included teachers ($N = 4$) who reasoned in *unacceptable* ways for at least 75% of their responses, and the remaining ($N = 3$) were observed to reason in *more unacceptable* ways for between 64% and 70% of their responses. Moreover, teachers in the *mostly acceptable* category of IIR were observed to balance their responses to *inferences* between critiques and suggestions while those in the *mostly unacceptable* category typically offered critiques of the study design. Considering the three components of IIR, two teachers each utilized all three components one time and all teachers except Amalia incorporated both the *inference* and *data* components when engaging in IIR in *acceptable* forms. The *uncertainty* component was only observed in the cases where all three components were incorporated, indicating that teachers did not consider the deterministic nature of their inferential statements.

Many teachers were observed to be in different categories of reasoning within non-IIR contexts than they were within IIR contexts. Two teachers in the *mostly acceptable* non-IIR reasoning type remained in the *mostly-acceptable* type of reasoning within IIR contexts. The same is true of the one teacher who was in the *mostly unacceptable* category for non-IIR contexts—she was also in the *mostly unacceptable* category for IIR contexts. The remaining six teachers were in different categories. Three teachers from the *mostly acceptable* category of non-IIR moved to the *more unacceptable* reasoning type for IIR and three teachers from the *mixed* type of non-IIR moved to *mostly unacceptable* for IIR. Thus, teachers struggled to reason in *acceptable* forms in IIR contexts much more than in non-IIR contexts.

Comparing reasoning types to knowledge types, the greatest observable pattern is that teachers with mostly *desirable* elements that were highly integrated also reasoned in *mostly acceptable* ways while teachers with many *undesirable* elements that were largely *disconnected* tended to reason in *mostly unacceptable* ways. A second salient pattern is that those who reasoned in more *acceptable* forms across both contexts tended to have more connected knowledge structures. Moreover, the relationship between reasoning types and knowledge structures appeared most evident at opposite ends of the spectrum. In particular, the teachers who were in the *mostly acceptable* category within both reasoning contexts were observed to also have *desirable-connected* knowledge structures. On the other end of the spectrum, the one teacher who was in the *mostly unacceptable* reasoning category within both reasoning contexts was also observed to have an *undesirable-disconnected* knowledge structure. The relationship is not as clear between these two ends of the spectrum. However, none of the teachers who moved from the

mixed category of reasoning within non-IIR contexts to the *mostly unacceptable* category for IIR contexts were observed to have *desirable-connected* knowledge structures. Furthermore, none of the teachers who moved from *mostly acceptable* reasoning within non-IIR contexts to *more unacceptable* within IIR contexts were observed to have *undesirable-disconnected* knowledge structures.

Statistics PCK and relations to knowledge and IIR. Overall, teachers rarely exhibited the highest levels of PCK. Only Tim was observed at the *accomplished* level of PCK and only Rosalynn was observed at the *competent* level. Five teachers were at the two lower levels of PCK—*competent/aware* ($N = 4$) and *aware* ($N = 1$). Substantial changes occurred when disaggregating to non-IIR and IIR contexts. For instance, within non-IIR contexts only Ruby was observed at the *aware* level, while within IIR contexts there were six teachers at the *aware* level. The general trend is that most teachers were one level lower in PCK among IIR contexts than among non-IIR contexts. Ellie was two levels lower within non-IIR contexts than IIR contexts. However, two teachers remained at the same level of PCK.

Comparing PCK levels with knowledge and reasoning categories revealed a general pattern. Teachers who reasoned in *mostly unacceptable* or *mixed* forms within IIR contexts and in *mostly unacceptable* forms within IIR contexts were also observed to have either *undesirable-disconnected* or *undesirable-connected* knowledge structures and were at the *aware* level of PCK for IIR contexts. However, within non-IIR contexts, this same group of teachers was observed across the *aware*, *competent/aware*, and *competent* levels of PCK—indicating that knowledge and reasoning types appear to be largely unrelated to teachers' PCK within non-IIR contexts.

Discussion of Findings

The purpose of this study was to describe ways teachers' knowledge of center, spread, and shape may provide support for teachers' informal inferential reasoning and to describe their PCK within IIR contexts. In the following sections, I discuss findings in relation to results of prior empirical studies and theoretical discussions from the published literature in statistics education.

Knowledge Structures for Center, Spread, and Shape

A major finding of this study is that middle and secondary teachers' knowledge structures for center, spread, and shape of distributions fell into three categories—*desirable-connected*, *undesirable-connected*, and *undesirable-disconnected*. Looking across these categories, seven of the nine teachers were found to have highly interconnected knowledge structures that all made connections among the knowledge types of center and spread, and some made connections across all three knowledge types. Moreover, despite many teachers having *undesirable* elements integrated into their knowledge structures, most were observed to have at least one connection between *desirable* elements of center and spread. Although there was evidence that teachers' knowledge structures were highly interconnected, teachers struggled to integrate their knowledge of center, spread, and shape.

In a similar way, Groth and Bergner (2006) observed few elementary pre-service teachers at the highest level of reasoning (extended abstract) with measures of center—which required an understanding of when different types of center were more useful than others. Although many teachers in the current study were observed to have many connections between conceptions of center, several were found to exhibit *undesirable* generalizations from those connections. For instance, on both the New Year's Day Race

task and Tomatoes and Fertlizer task, Ruby insisted that the mean was not appropriate because it was sensitive outliers and on several occasions admitted her preference for the median as a result. Thus, she appears to have constructed a belief that the median is always a better measure of center because it is not influenced by outliers—despite the fact that outliers should not necessarily be ignored. On the other hand, Amalia, Ellie, and Michaela showed evidence that they believed the mean was always a better measure of center because it took the magnitude of all values into consideration. For instance, Amalia goes so far as to call the mean the "true center" (Jumping Distances, Lines 18, 57). Therefore, although they may appear to be at the highest level of Groth and Bergner's (2006) framework, based on the knowledge structures, many teachers ultimately did not fully comprehend when *each* measure of center was preferable.

With respect to variability, Silva and Coutihno (2008) found that most middle level and secondary teachers in their study were at a verbal reasoning stage, indicating that they could only describe variability as a measure of how similar values were to a measure of center. The present study appears to mirror this finding. Teachers who recognized that standard deviation was a measure of spread from the mean largely did not provide evidence that they understood it as intervals around the mean. The majority of the teachers in this study relied on the range to describe spread and although every teacher connected ideas of standard deviation to a measure of center (some incorrectly), most are likely at a verbal or transitional reasoning level because they did not attempt to estimate standard deviation. The exceptions are Tim, Harrison, and Rosalynn who provided some indication of thinking about intervals—a procedural level of reasoning.

Also related to variability, Noll and Shaughnessy (2012) introduced a hierarchical framework for reasoning with variability that characterized the highest level of reasoning as involving an integration of multiple knowledge types (center, spread, shape). Although most teachers in this study were observed to make some connection between center and spread, teachers largely did not draw on any two knowledge types simultaneously in their responses to tasks. Therefore, as found in Noll and Shaughnessy's study with middle level and secondary students, most teachers were at the lower levels of the framework.

Regarding shape, Doerr and Jacobs (2011) found that secondary mathematics teachers tended to perceive bell-shaped distributions as having less variability than skewed distributions. This study supports this finding, but it was not a widespread trend—only Ellie and Amalia were observed to offer some evidence of this perspective. However, teachers in this study were not explicitly asked to compare such situations, so it could be more widespread than it appears.

Teachers in this study were largely observed to recognize that connections existed between knowledge element types (although not as much between center and shape), yet they did not utilize these connections and draw on multiple knowledge types simultaneously, as prior research has found for students and pre-service teachers (Doerr & Jacob, 2011; Groth & Bergner, 2006; Noll & Shaughnessy, 2012).

Knowledge Structures that Support IIR

Although the conceptual framework for this study does not assume that there is any particular prior knowledge necessary for engaging in IIR (Means & Voss, 1996; Perkins, 1985; Zieffler et al., 2008) and that knowledge and informal reasoning work together to allow engagement in IIR, the teachers who engaged in IIR in *mostly acceptable* forms were those with *desirable-connected* knowledge structures. This

finding aligns with Makar and colleagues' (2011) claim that statistical knowledge is an important support for IIR.

When narrowing to part (b) of the *Jumping Distances* task, eight of the nine teachers were able to provide at least one *inference* that resulted from *acceptable* reasoning within an IIR context. Moreover, all eight of these teachers incorporated two of the three components in at least one of their arguments. Given that all teachers were observed to have at least some *desirable* knowledge elements for center and spread, and the majority of *acceptable* forms of IIR involved only reasoning about center, it remains aligned to the theory that the knowledge supports their reasoning (Makar et al., 2011). Furthermore, it is worth noting that no one on this task reasoned from a general indication of center, but rather used the median—a more formal measure of center explicitly visible on the provided boxplots.

Another important finding was that reasoning types among non-IIR contexts were much more homogeneous. Seven of the nine teachers offered more *acceptable* forms of reasoning within non-IIR contexts than they did *unacceptable* forms of reasoning. Moreover, three of the four non-IIR tasks did not require integrating multiple knowledge types in order to make and justify claims from the data— providing further support to Makar and colleagues' (2011) claims. Teachers did not need support from an integrated knowledge structure in order to respond to these items, resulting in a larger number of responses that were considered acceptable.

Noll and Shaughnessy's (2012) framework indicated that when multiple knowledge types are integrated, a higher level of reasoning is observed—and that most middle level and secondary students in their study were not observed at this level.

Likewise, the teachers in this study were observed to have similarly limited reasoning levels. Teachers in this study, even when observed to have connected ideas of multiple knowledge elements, mostly did not integrate multiple concepts into their reasoning. As found by Huey (2011) with preservice mathematics teachers and Watson (2003) with elementary, middle, and secondary students, teachers in this study tended to focus only on measures of center unless explicitly prompted otherwise.

Relations Between Knowledge, IIR, and PCK

Prior literature has suggested that teachers of stand-alone statistics courses may not have taken more courses focused on statistics content than teachers who do not teach stand-alone courses (Engledowl, 2016). Although the teachers in this study also reported completing similar coursework in statistics, they also generally reported feeling confident in teaching statistics. However, Engledowl (2016) found that teachers of stand-alone statistics courses reported feeling much more confident teaching statistics than mathematics teachers who do not teach stand-alone statistics courses. In this study, a surprising result is that despite similar backgrounds with statistics and confidence in teaching the content, the teachers of stand-alone statistics courses in this study were found to (a) have different knowledge structures that included nearly no *undesirable* knowledge elements, (b) reason in *mostly acceptable* ways in IIR contexts, and (c) have a higher level of PCK for statistics within IIR contexts. Although it was beyond the scope of this study to investigate why these patterns may be different, prior research has found that expert AP Statistics teachers are highly motivated to seek out ways to deepen their knowledge that extends outside of college coursework and professional development (Peters, 2013; Whitaker, 2016).

Although there have been no studies conducted to assess teachers' PCK for IIR contexts, a few studies are potentially informative. Over the course of several studies focused on various statistics sub-domains of content, Watson and Callingham (Callingham & Watson, 2011; Watson & Callingham, 2013, 2014; Watson et al., 2008) found that most of the elementary teachers in their studies were at the *competent* or *emerging* level of PCK for statistics. This study reflects these results *only for non-IIR contexts*, with most teachers being at the *competent/aware* or *competent* level. Within IIR contexts, eight of the nine teachers were at the *aware* level ($N = 6$) or *competent/aware* level ($N = 2$). On the other hand, Leavy (2010) found that preservice teachers struggled to press students to use data as evidence when teaching lessons designed to engage elementary students in IIR. In contrast, teachers in this study, even those at the *aware* level of PCK for IIR contexts, tended to focus first on the use of data as evidence to support reasoning—even though it may have been focused on evidence that did not support a statement in *acceptable* ways.

Comparing knowledge structures and PCK for statistics, Groth and Bergner (2013) found that even when pre-service elementary teachers had correct knowledge structures for both content and pedagogy, it did not always result in appropriate interactions with students. In a similar way, this study found that teachers' PCK for statistics did not always follow predictable patterns. Ruby and Ellie, for example, were observed to have similar knowledge constructions and to reason in similar ways for both non-IIR and IIR contexts. Moreover, they were both found to be at the *aware* level of PCK for IIR contexts. However, within non-IIR contexts, Ellie was observed at the *competent* level of PCK (a difference of two levels), yet Ruby remained at the *aware*

level. On the other end of the spectrum, Mike was observed to have a *desirable-connected* knowledge structure and to reason in *mostly acceptable* ways within non-IIR contexts. However, he was found to be at the *competent-aware* level of PCK, implying that his highly-integrated knowledge structure of *desirable* elements and his ability to reason in *acceptable* ways from that knowledge was not sufficient for him to consistently recognize errors in reasoning and suggest interventions that addressed the errors.

Limitations of the Study

Assessment Tasks

The statistical problem solving process involves formulating questions, collecting data, analyzing data, and interpreting results (Franklin et al., 2007). However, the LOCUS tasks used in this study incorporated only the *analyze data* and *interpret results* portions of this process in order to more clearly focus on the point at which an inferential statement is constructed or justified. Although this decision allowed for a more direct access to teachers' reasoning with inference, and maintained the recommended task features for engagement in IIR (Huey & Jackson, 2015), by not constructing the statistical question and engaging in collecting the data, the tasks were perhaps less authentic to participants. Moreover, some of the tasks were only intended to observe one or two of the three components of IIR (Makar & Rubin, 2009). This decision was made to ensure that all components were more likely to be observed. However, the interpretation of results of this study are limited to the restricted environment imposed by the tasks.

Pedagogical Content Knowledge

Currently, no measures exist for evaluating teachers' PCK for statistics at the middle and secondary level, and the method this study employed to analyze PCK was based on the limited knowledge that exists on the topic. Therefore, the levels of PCK

reported in this study should be considered an initial attempt at characterizing teachers' PCK in the context of middle and secondary statistics content within the context of *informal inferential reasoning*. Drawing on the literature, I inferred teachers' PCK based on their suggestions of how they would interact with hypothetical students, devoid from the realities of the classroom. In this regard, this study presents a best-case scenario because teachers were allowed an opportunity to consider an interaction with a student that was removed from the complexities of a classroom environment. For example, teachers were unconcerned about how much time they would have to spend engaging with a single student. It is, therefore, unknown whether the teachers in this study would actually interact with students in the ways that they suggested.

Implications for Teacher Education

Tasks Encouraging Integrating Knowledge Elements Within and Between Types

Some of the teachers in this study had one-sided perspectives, such as that the mean was always better or that the median was always better. In such cases, teachers' conceptions were largely based on dichotomous views of outliers as either *always* or *never* ignorable. Therefore, teachers need to be exposed to situations where outliers can be disregarded and situations where outliers should *not* be disregarded to avoid an over-generalization that may result in misconceptions about measures of center and about the occurrence of outliers. More broadly, teachers need opportunities to develop a fuller understanding of how different measures of center are related to one another and when each is more useful than the others.

Teachers in this study largely focused on range and sub-ranges instead of considering spread as a measure from center. The consequence is that teachers were not considering the center *in relation to* the spread. Not being able to reason with center and

spread simultaneously can make it difficult for teachers to respond to students who reason in such ways, and can result in inadvertently persuading students away from more complex ways of reasoning. It is important that teachers be provided opportunities to informally explore the connections between spread and center to allow a more robust conception to develop.

In this study, many teachers only drew on ideas of center to support *inferences*, even when the tasks did not restrict responses in such a way. However, by only incorporating single concepts, teachers may not integrate multiple knowledge types, even when afforded an opportunity. Consequently, tasks designed to engage teachers in IIR should also encourage attention to center, spread, and shape, and not isolate one knowledge element type. In doing so, such tasks have the potential to not only strengthen teachers' understanding of the connections between knowledge element types, but to also improve their statistical reasoning by providing multiple pieces of evidence that more coherently support conclusions drawn from data.

Opportunities to Weigh the Evidence of Inferential Statements

In this study, all teachers provided both *appropriate* and *inappropriate* ways of reasoning within IIR contexts. Moreover, when teachers were confronted with different ways of reasoning, they had an opportunity to weigh the evidence to support the claims they had made. Not only did this provide teachers another opportunity to evaluate the quality of their own reasoning, but it also allowed teachers to explore multiple claims and multiple supports in order to construct a stronger argument. This notion of weighing the evidence was noted by Pfannkuch (2006a) as an essential part of IIR, and Makar and colleagues (2011) also claimed that engaging in this process improves confidence in the *inference*. In light of this evidence, professional development that focuses on statistics

content should provide teachers with opportunities to 1) construct multiple *inferential* statements and 2) respond to multiple *inferential* statements and explore their validity.

Providing teachers with these two situations will also allow them to engage in the statistical analog of the mathematical practice from CCSSM (NGA & CCSSO, 2010) that states students should "construct viable arguments and critique the reasoning of others" (p. 6). According to the SET report (Franklin et al., 2015), the analog to this practice for statistics is that "statistically proficient students also are able to compare the plausibility of alternative conclusions and distinguish correct statistical reasoning from that which is flawed" (p. 14). Therefore, being presented with multiple *inferences* that motivate teachers to explore the ideas and weigh the evidence will encourage engagement in this practice that teachers' students are expected to engage in as well—giving teachers an experience to draw on when planning statistics lessons.

Another reason to have teachers consider and construct multiple *inferences* and then weigh the evidence of each one is that teachers can engage in the IIR component of using *data as evidence* that yield different *inferences*. For instance, on the Jumping Distances task, a focus on the difference in center leads to an inference that the target caused shorter jump distances, while a focus on the difference in IQR leads to an inference that the target caused more consistent jumps. On the other hand, the same evidence can lead to different inferences when there is a weighing of the evidence.

Although several teachers were observed making different inferences by drawing on different *data as evidence*, it was rare that a teacher spontaneously engaged in this idea of weighing the evidence. This rarity is likely due to the lack of experience teachers have with statistical investigations. As described by Makar and colleagues (2011),

weighing the evidence, which they described as a search for "deeper statistical evidence" (p. 170), allows for an improvement in the confidence of the inference. Accordingly, teacher preparation and development programs should incorporate the use of statistical investigations that encourage this type of weighing the evidence.

Explicit Attention to Using Probabilistic Language

This study observed a profound lack of attention to the component of IIR described by Makar and Rubin (2009) as using probabilistic language to indicate a level of uncertainty in an inferential statement. Unlike mathematics, where deterministic statements are the norm, inferences in statistics are predictions that extend beyond the data and rather imply a "*statistical tendency*, and/or *level of confidence* or *uncertainty* in a prediction" (Makar & Rubin, 2009, p. 87, emphasis in original). At the post-secondary level, this probabilistic language should not be proceduralized to include this information. For instance, the very introduction of an alpha level allows for the proceduralization of the *uncertainty* component by being able to ignore how *uncertain* an inference may be and instead simply refer to a *p*-value. This proceduralization is problematic, as evidenced by many decades of research that have found that students can compute *p*-values and make correct decisions about rejecting, or not, the null hypothesis, and yet they have widespread misconceptions of what these values mean (Castro Sotos et al., 2009, 2007; delMas et al., 2007; Falk & Greenbaum, 1995).

This study provides evidence that the *uncertainty* of the teachers' *inferential* statements was absent, despite evidence of a formal understanding of *p*-values and confidence intervals. For the purposes of a formal setting, perhaps this is not a cause for concern. However, when considering the development of statistical literacy that hopes to instill an ability to 1) interpret these kinds of statements and 2) apply the principles in

situations that are less formal, it *does* give cause for concern. In this study, it seems likely that these formal tests were viewed only as procedures because 1) all teachers had completed at least one tertiary course in statistics, and therefore would have experienced hypothesis testing, and 2) it was exceedingly rare to observe teachers give attention to the *uncertainty* component when making *inferences*. Therefore, teacher development programs that aim to improve teachers' statistical knowledge for teaching should provide an explicit attention on incorporating the *uncertainty* component into *inferential* statements.

Recommendations for Future Research

A major finding of this study was the wide range of knowledge structures of middle and high school mathematics teachers. Although multiple points of evidence were provided for connections made between knowledge elements, more research is needed to ascertain the *strength* of such relationships. Such investigations might aid in explaining differences in teachers' knowledge structures that may appear to be similar but whose reasoning indicates salient differences in how that knowledge is applied in IIR and non-IIR contexts. Moreover, teachers' knowledge structures need to be examined as they develop—similar to Groth and Bergner (2013). This type of study design would have the potential to explain the formation of *undesirable* knowledge elements, and subsequently lead to the development of interventions to foster the development of *desirable* constructions of knowledge elements and connections between them. Ultimately, such research could inform a more purposeful design of professional development programs recommended by *The Statistical Education of Teachers* report (Franklin et al., 2015).

In this study, teachers' informal inferential reasoning was analyzed in a cross-sectional way—at one point in time. Although some work has been done to understand

how students' (e.g., Garfield et al., 2012) and preservice teachers' (Huey, 2011) IIR develops over time, more research is needed to understand how informal reasoning necessary for IIR may develop alongside construction of highly connected *desirable* knowledge structures to support teachers' IIR.

A final recommendation stems from the context of this study, namely that teachers' PCK was assessed in the restricted context of IIR tasks. Moreover, PCK was assessed based on teachers' suggestions for intervening with hypothetical students. It follows that additional research is needed to examine teachers' PCK in classroom contexts, rather than in the context of a task-based clinical interview. Such research would allow for a more robust characterization of the challenges teachers face while beginning to engage their students in IIR and what supports enable them to overcome such challenges.

Reflections

This study addressed a lack of research on middle and secondary mathematics teachers' statistics knowledge, IIR, and PCK for statistics, and possible ways they are related. Findings indicated that teachers with the strongest and weakest knowledge structures were associated with stronger and weaker IIR and stronger and weaker PCK, respectively. However, in general, all teachers struggled to reason in acceptable ways in IIR contexts and to identify student misunderstandings within IIR contexts and offer interventions to address them. Moreover, there were no observable differences in course types among those teaching statistics as a unit within a mathematics course. Given these findings, more work is needed to identify catalysts for learning that lead to development of highly connected knowledge structures alongside IIR that involves all three components, while also increasing teachers' PCK level.

Given that the world has quickly become dependent on data, statistical literacy has necessarily ascended to become a prominent goal of K-16 education. It is worthy of recognition that although much of the statistics instruction that occurs in K-12 settings takes place within mathematics courses—as well it should (see Franklin et al., 2015)—instruction for statistics is likely occurring in other contexts. For instance, scientific models are based on statistical inference, meaning that biology, chemistry, and even physics teachers must be able to reason with and, at least, support the teaching of inference occurring in mathematics classrooms. To further extend the boundaries, the *Common Core State Standards for English & Language Arts* (NGA & CCSSO, 2010) requires middle grades students to "distinguish among facts, reasoned judgment based on research findings, and speculation in a text", and 9th and 10th grade students to "assess the extent to which the reasoning and evidence in a text support the author's claim or a recommendation for solving a scientific or technical problem" (see RST.6-8.8 and RST.9-10.8). The bottom line is that although mathematics teachers are most in need of being supported to teach students to reason with inference, *all* teachers need to be statistically literate and experienced with reasoning inferentially in order to support the development of a statistically literate society.

REFERENCES

Amerom, B. A. Van. (2003). Focusing on informal strategies when linking arithmetic to early algebra. *Educational Studies in Mathematics, 54*(1), 63–75.

Australian Curriculum Assessment and Reporting Authority. (2014). *Autstralian Curriculum*. Retrieved from http://www.australiancurriculum.edu.au/mathematics/curriculum/f-10?layout=1

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83. Retrieved from http://iase-web.org/documents/SERJ/SERJ3%282%29_Bakker.pdf

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Springer Science + Business, Inc.

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistical Education Research Journal, 7*(2), 130–145. Retrieved from http://www.stat.auckland.ac.nz/serj

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). *Report of the 2012 national survey of science and mathematics education*. Chapel Hill, NC: Horizon Research, Inc.

Batanero, C., Burrill, G., & Reading, C. (Eds.). (2011). *Teaching statistics in school mathematics–Challenges for teaching and teacher education. A Joint ICMI/IASE Study*. The 18th ICMI Study. New York, NY: Springer Science + Business Media.

Ben-Zvi, D., & Garfield, J. (2005a). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 3–15). Springer

Science + Business Media, Inc. Retrieved from http://link.springer.com/chapter/10.1007/1-4020-2278-6_1

Ben-Zvi, D., & Garfield, J. (Eds.). (2005b). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Springer Science + Business Media, Inc.

Beswick, K., Callingham, R., & Watson, J. (2012). The nature and development of middle school mathematics teachers' knowledge. *Journal of Mathematics Teacher Education, 15*, 131–157. Retrieved from http://link.springer.com/article/10.1007/s10857-011-9177-9

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic.

Booker, G. (1996). Constructing mathematical conventions formed by the abstraction and generalization of earlier ideas: The development of initial fraction ideas. In L. P. Steffe, P. Nesher, P. Cobb, G. A. Goldin, & B. Greer (Eds.), *Theories of Mathematical Learning* (pp. 381–395). New York, NY: Routledge.

Burrill, G., & Biehler, R. (2011). Fundamental Statistical Ideas in the School Curriculum and in Training Teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics–Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 57–69). The 18th ICMI Study. New York, NY: Springer Science + Business Media.

Callingham, R., & Watson, J. (2011). Measuring levels of statistical pedagogical content knowledge. In C. Batanero, G. Burill, & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education. A Joint ICME/IASE Study* (pp. 283–293). The 18th ICMI Study. New York, NY: Springer

Science + Business Media.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Education Research Journal*, *26*(4), 499–531.

Castro Sotos, A. E., Vanhoof, S., Noortgate, W. Van den, & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education*, *17*(2).

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98–113. https://doi.org/10.1016/j.edurev.2007.04.001

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

Conference Board of the Mathematical Sciences. (2012). *The mathematical education of teachers II*. Providence, RI and Washington, DC: American Mathematical Society and Mathematical Association of America. Retrieved from http://www.cbmsweb.org/MET2/index.htm

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). SAGE. Retrieved from https://books.google.com/books?id=Ykruxor10cYC

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*,

6(2), 28–58.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.219.7563&rep=rep1&type=pdf

Dierdorp, A., Bakker, A., van Maanen, J. A., & Eijkelhof, H. M. C. (2012). Supporting students to develop concepts underlying sampling and to shuttle between contextual and statistical spheres. In D. Ben-Zvi, J.-C. Oriol, & L. Cordani (Eds.), *Topic Study Group 12: Teaching and learning of statistics*. Topic Study Group conducted at the 12th International Congress on Mathematics Education, Seoul, Korea.

Doerr, H. M., & Jacob, B. (2011). Investigating secondary teachers' statistical understandings. In M. Pytlak, T. Rowland, & E. Swoboda (Eds.), *Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education* (pp. 776–786). University of Rzeszów, Poland. Retrieved from https://www.researchgate.net/profile/Maria_Meletiou-Mavrotheris/publication/283347725_Stochastic_thinking_Proceedings_of_Working_Group_5_at_CERME_7/links/5635ada008aebc003fff737c.pdf#page=80

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding of hypothesis testing concepts. *Statistics Education Research Journal*, *14*(1), 60–89. Retrieved from http://iase-web.org/documents/SERJ/SERJ14(1)_Dolor.pdf

England Department of Education. (2014). *National curriculum in England: mathematics programmes of study*. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-

mathematics-programmes-of-study/national-curriculum-in-england-mathematics-programmes-of-study

Engledowl, C. (2016). Characteristics of secondary statistics teachers. In M. B. Wood, E. E. Turner, M. Civil, & J. A. Eli (Eds.), *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (p. 1008). Tucson, AZ: The University of Arizona. https://doi.org/10.1016/B978-0-12-802121-7.01602-2

Falk, R., & Greenbaum, C. W. (1995). Significance Tests Die Hard The Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology*, *5*(1), 75–98. https://doi.org/10.1177/0959354395051004

Farmaki, V., & Paschos, T. (2007). The interaction between intuitive and formal mathematical thinking: A case study. *International Journal of Mathematical Education in Science & Technology*, *38*(3), 353–365.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education*, *28*(1), 96–105.

Franklin, C. (2013). Common Core State Standards and the future of teacher preparation in statistics. *The Mathematics Educator*, *22*(2), 3–10. Retrieved from http://tme.journals.libs.uga.edu/index.php/tme/article/viewFile/252/239

Franklin, C., Bargagliotti, A. E., Case, C. A., Kader, G., Scheaffer, R., & Spangler, D. A. (2015). *The Statistical Education of Teachers*. Retrieved from http://www.amstat.org/education/SET/SET.pdf

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*. *American Statistical Association*. Alexandria, VA. Retrieved from http://www.amstat.org/Education/gaise/GAISEPreK-12_Full.pdf

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association. Retrieved from http://www.amstat.org/education/gaise

Garfield, J., DelMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM - The International Journal on Mathematics Education*, *44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Gfeller, M. K., Niess, M. L., & Lederman, N. G. (1999). Preservice teachers' use of multiple representations in solving arithmetic mean problems. *School Science and Mathematics*, *99*(5), 250–257. https://doi.org/10.1111/j.1949-8594.1999.tb17483.x

Gil, E., & Ben-Zvi, D. (2011). Explanations and context in the emergence of students' informal inferential reasoning. *Mathematical Thinking and Learning*, *13*(1–2), 87–108.

Ginsburg, H. (1981). The clinical interview in psychological research on mathematical thinking: Aims, rationales, techniques. *For the Learning of Mathematics*. Retrieved from http://www.jstor.org/stable/40247721

Godino, J. D., Ortiz, J. J., Roa, R., & Wilhelmi, M. R. (2011). Models for Statistical Pedagogical Knowledge. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics–Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 271–282). The 18th ICMI Study. New York, NY:

Springer Science + Business Media.

Goldin, G. A. (1997). Chapter 4: Observing mathematical problem solving through task-based interviews. *Journal for Research in Mathematics Education. Monograph*, *9*(Qualitative Research Methods in Mathematics Education), 40-62-177.

Gravemeijer, K., & Doorman, M. (1999). Context problems in realistic mathematics education: A calculus course as an example. *Educational Studies in Mathematics*, *39*(1), 111–129. Retrieved from http://link.springer.com/article/10.1023/A:1003749919816

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, *38*(5), 427–437. Retrieved from http://www.jstor.org/stable/30034960

Groth, R. E. (2013). Characterizing key developmental understandings and pedagogically powerful ideas within a statistical knowledge for teaching framework. *Mathematical Thinking and Learning*, *15*(2), 121–145. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10986065.2013.770718

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning*, *8*(1), 37–63. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15327833mtl0801_3

Groth, R. E., & Bergner, J. A. (2013). Mapping the structure of knowledge for teaching nominal categorical data analysis. *Educational Studies in Mathematics*, *83*(2), 247–265. https://doi.org/10.1007/s10649-012-9452-4

Haberstroh, J., Jacobbe, T., DelMas, R., Hartlaub, B., Miller, D., Scheaffer, R., …

Watson, J. (2015). LOCUS Evidence Model.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7*(1), 1–20.

Herscovics, N. (1996). The construction of conceptual schemes in mathematics. In L. P. Steffe, P. Nesher, P. Cobb, G. A. Goldin, & B. Greer (Eds.), *Theories of Mathematical Learning* (pp. 351–379). New York, NY: Routledge.

Hill, H. C., Ball, D. L., & Schilling, S. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372–400. Retrieved from http://www.jstor.org/stable/40539304

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511. https://doi.org/10.1080/07370000802177235

Hill, H. C., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406. Retrieved from http://aer.sagepub.com/content/42/2/371.short

Hill, H., Schilling, S., & Ball, D. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, *105*(1), 11–30. Retrieved from http://www.jstor.org/stable/10.1086/428763

Huey, M. E. (2011). *Characterizing middle and secondary preservice teachers' change in inferential reasoning*. University of Missouri, Columbia, MO. Retrieved from http://search.proquest.com/docview/1262397961/abstract/AB7A075587244E0FPQ/

1?accountid=14576

Huey, M. E., & Jackson, C. D. (2015). A framework for analyzing informal inferential reasoning tasks in middle school textbooks. In T. G. Bartell, K. N. Bieda, R. T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 431–438). East Lansing, MI: Michigan State University.

Inspiration. (2014). Inspiration (version 9.2.2). Inspiration Software, Inc. Retrieved from http://www.inspiration.com/

Izsák, A. (2008). Mathematical knowledge for teaching fraction multiplication. *Cognition and Instruction*, *26*(1), 95–143.

Jacobbe, T. (2016). Levels of Conceptual Understanding in Statistics (LOCUS). Retrieved April 18, 2016, from https://locus.statisticseducation.org/

Jacobbe, T., & Carvalho, C. (2011). Teachers' understanding of averages. In C. Batanero, G. Burill, & C. Reading (Eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 199–209). The 18th ICMI Study. New York, NY: Springer Science + Business Media.

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the validity of the LOCUS assessments through an evidenced-centered design approach. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, Arizona, USA* (pp. 1–6). Voorburg, The Netherlands: International Statistical Institute. Retrieved from

http://icots.info/9/proceedings/pdfs/ICOTS9_7C2_JACOBBE.pdf

Jones, K. (1995). Acquiring abstract geometrical concepts: The interaction between the formal and the intuitive. In M. Selinger & T. Smart (Eds.), *Proceedings of the 3rd British Congress on Mathematics Education (BCME3)* (pp. 239–246). BCME.

Langrall, C. W., Makar, K., Nilsson, P., & Shaughnessy, J. M. (2017). The teaching and learning of probability and statistics: An integrated perspective. In J. Cai (Ed.), *Compendium for Research in Mathematics Education* (pp. 490–525). Reston, VA: National Council of Teachers of Mathematics.

Lave, J., Murtaugh, M., & de la Rocha, O. (1984). The dialectic of arithmetic in grocery shopping. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 67–94). Harvard University Press.

Lavigne, N. C., & Lajoie, S. P. (2007). Statistical reasoning of middle school children engaged in survey inquiry. *Contemporary Educational Psychology*, *32*(4), 630–666.

Leavy, A. M. (2010). The challenge of preparing preservice teachers to teach informal inferential reasoning. *Statistics Education Research Journal*, *9*(1), 46–67. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ9(1)_Leavy.pdf

Leavy, A., & O'Loughlin, N. (2006). Preservice teachers understanding of the mean: Moving beyond the arithmetic average. *Journal of Mathematics Teacher Education*, *9*(1), 53–90.

Lester Jr., F. K. (Ed.). (2007). *Second Handbook of Research on Mathematics Teaching and Learning*. Reston, VA: National Council of Teachers of Mathematics.

Makar, K. (2014). Young children's explorations of average through informal inferential reasoning. *Educational Studies in Mathematics*, *86*(1), 61–78.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 152–173.

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 1–4.

Makar, K., & Confrey, J. (2003). Chunks, clumps, and spread out: Secondary preservice teachers' informal notions of variation and distribution. *Reasoning about Variability: A Collection of Current Research Studies*. Retrieved from http://people.cst.cmich.edu/lee1c/SRTL3/SRLT_3_papers/Makar-Confrey SRTL3 FINAL.pdf

Makar, K., & Confrey, J. (2005). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353–373). Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/1-4020-2278-6_15.pdf

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, *14*(2), 139–178. https://doi.org/10.1207/s1532690xci1402_1

National Assessment Governing Board, & U.S. Department of Education. (2012). *Mathematics Framework for the 2013 National Assessment of Educational Progress*. Washington, DC.

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

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington D.C.

National Governors Association Center for Best Practices & Council of Chief State SchoolOfficers. (2010). *Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects*. Washington D.C.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. US Department of Education.

Noll, J., & Shaughnessy, J. M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education*, *43*(5), 509–556. https://doi.org/10.5951/jresematheduc.43.5.0509

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, *7*(2), 83–106. Retrieved from http://www.stat.auckland.ac.nz/serj

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods* (3rd ed.). Thousand Oaks, California: Sage Publications Inc.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*(5), 562–571. https://doi.org/10.1037/0022-0663.77.5.562

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and the roots of intelligence* (pp. 83–106). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. Retrieved from http://psycnet.apa.org/psycinfo/1991-97717-005

Peters, S. A. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal*, *10*(1), 52–88. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ10(1)_Peters.pdf

Peters, S. A. (2013). Developing understanding of statistical variation: Secondary statistics teachers' perceptions and recollections of learning factors. *Journal of Mathematics Teacher Education*, *17*(6), 539–582. https://doi.org/10.1007/s10857-013-9242-7

Pfannkuch, M. (2006a). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27–45. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2).pdf#page=30

Pfannkuch, M. (2006b). Informal inferential reasoning. In *Conference proceedings of the 7th International Conference on Teaching Statistics* (pp. 1–6). Retrieved from http://iase-web.org/documents/papers/icots7/6A2_PFAN.pdf

Pfannkuch, M. (2011). The Role of Context in Developing Informal Statistical Inferential Reasoning: A Classroom Study. *Mathematical Thinking and Learning*, *13*(1–2), 27–46. https://doi.org/10.1080/10986065.2011.538302

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research*, *5*(2), 4–9. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2).pdf#page=7

Piaget, J. (1997). Development and learning. In M. Guavain & M. Cole (Eds.), *Readings on the development of children* (pp. 19–28). New York: W.H. Freeman and Company. Retrieved from (Reprinted from Piaget Rediscovered, pp. 7-20, by R. E. Ripple & V. N. Rockcastle (Eds.), 1964)

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 3–4. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Pratt_Ainley.pdf

Reading, C., & Canada, D. (2011). Teachers' Knowledge of Distribution. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics–Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 223–234). The 18th ICMI Study. New York, NY: Springer Science + Business Media.

Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal*, *7*(1), 40–59. Retrieved from http://www.stat.auckland.ac.nz/serj

Ron, G., Dreyfus, T., & Hershkowitz, R. (2010). Partially correct constructs illuminate students' inconsistent answers. *Educational Studies in Mathematics*, *75*(1), 65–87. https://doi.org/10.1007/s10649-010-9241-x

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 5–19. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2).pdf#page=8

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 Instrument. *Statistics Education Research Journal*, *14*(2), 93–116. Retrieved from http://iase-web.org/documents/SERJ/SERJ14(2)_Sabbag.pdf

Sánchez, E., da Silva, C. B., & Coutinho, C. (2011). Teachers' understanding of variation. In C. Batanero, G. Burill, & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education. A Joint ICME/IASE Study* (pp. 211–221). The 18th ICMI Study. New York, NY: Springer Science + Business Media.

Scheaffer, R. L. (1998). Statistics education - Bridging the gaps among school, college and the workplace. In *Conference proceedings of the 5th International Conference on Teaching Statistics* (pp. 20–27). Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/2/Keynote3.pdf

Shaughnessy, M. (2007). Research on statistics learning and reasoning. In F. K. Lester Jr. (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 957–1009). Reston, VA: NCTM.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.2307/1175860

Silva, C. B. da, & Coutinho, C. (2008). Reasoning about variation of a univariate distribution: A study with secondary mathematics teachers. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings from the ICMI Study 18 and 2008 IASE Roundtable Conference* (pp. 1–6). Monterrey, Mexico: ICME/IASE. Retrieved from https://iase-web.org/Conference_Proceedings.php?p=Joint_ICMI-IASE_Study_2008

Siraj Datoo. (2014, January 10). How tracking customers in-store will soon be the norm. *The Guardian*. Retrieved from

http://www.theguardian.com/technology/datablog/2014/jan/10/how-tracking-customers-in-store-will-soon-be-the-norm

Smith, J. P. I., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163. Retrieved from http://www.jstor.org.proxy.mul.missouri.edu/stable/1466679

Speer, N. M., King, K. D., & Howell, H. (2015). Definitions of mathematical knowledge for teaching: Using these constructs in research on secondary and college mathematics teachers. *Journal of Mathematics Teacher Education*, *18*(2), 105–122. https://doi.org/10.1007/s10857-014-9277-4

Steffe, L. (2001). A new hypothesis concerning children's fractional knowledge. *The Journal of Mathematical Behavior*, *20*(3), 267–307. Retrieved from http://www.sciencedirect.com/science/article/pii/S0732312302000755

Stohl Lee, H., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: Students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal*, *9*(1), 68–96. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ9(1)_Lee.pdf

Voss, J. F., Perkins, D. N., & Segal, J. W. (Eds.). (1991). *Informal reasoning and education*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46. Retrieved from http://iase-web.org/documents/SERJ/SERJ2(2)_Watson_Callingham.pdf

Watson, J., & Callingham, R. (2013). Likelihood and sample size: The understandings of

students and their teachers. *The Journal of Mathematical Behavior*, *32*(3), 660–672. https://doi.org/10.1016/j.jmathb.2013.08.003

Watson, J., & Callingham, R. (2014). Two-way tables: Issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning*, *16*(4), 254–284. https://doi.org/10.1080/10986065.2014.953019

Watson, J., Callingham, R., & Donne, J. (2008). Establishing PCK for teaching statistics. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICME/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education*. (Proceedings of the ICMI Study 18 and the 2008 IASE Round Table Conference, Monterrey, Mexico, July, 2008). Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/rt08

Whitaker, D. (2016). The development of a professional statistics teaching identity. In M. B. Wood, E. E. Turner, M. Civil, & J. A. Eli (Eds.), *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Pyschology of Mathematics Education.* (pp. 992–999). Tucson, AZ: The University of Arizona. Retrieved from http://www.pmena.org/pmenaproceedings/PMENA 38 2016 Proceedings.pdf

Whitaker, D., Foti, S., & Jacobbe, T. (2015). The levels of conceptual understanding in statistics (LOCUS) project: Results of the pilot study. *Numeracy*, *8*(2), 1–16. Retrieved from http://scholarcommons.usf.edu/numeracy/vol8/iss2/art3/

Whitaker, D., & Jacobbe, T. (2014). Lessons from the LOCUS assessments (part 1): Comparing groups. *Statistics Teacher Network*, *83*, 13–15. Retrieved from http://www.amstat.org/education/stn/pdfs/stn83.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. Retrieved from http://iase-web.org/documents/SERJ/SERJ7(2)_Zieffler.pdf

APPENDIX A: APPROVAL LETTER FROM HUMAN SUBJECTS IRB AT THE UNIVERSITY OF MISSOURI

Institutional Review Board
University of Missouri-Columbia

190 Galena Hall; Dc074.00
Columbia, MO 65212
573-882-3181
irb@missouri.edu

July 1, 2016

Principal Investigator: Christopher Michael Engledowl
Department: Learning Teaching & Curriculum

Your Exempt Application to project entitled Mapping Teachers' Cognitive Structures that Support Informal Inferential Reasoning was reviewed and approved by the MU Institutional Review Board according to the terms and conditions described below:

IRB Project Number	2006053
IRB Review Number	217065
Initial Application Approval Date	July 01, 2016
IRB Expiration Date	July 01, 2017
Level of Review	Exempt
Project Status	Active - Open to Enrollment
Exempt Categories	45 CFR 46.101b(2)
Risk Level	Minimal Risk
Internal Funding	Personal funds

The principal investigator (PI) is responsible for all aspects and conduct of this study. The PI must comply with the following conditions of the approval:

1. No subjects may be involved in any study procedure prior to the IRB approval date or after the expiration date.
2. All unanticipated problems, adverse events, and deviations must be reported to the IRB within 5 days.
3. All changes must be IRB approved prior to implementation unless they are intended to reduce immediate risk.
4. All recruitment materials and methods must be approved by the IRB prior to being used.
5. The Annual Exempt Form must be submitted to the IRB for review and approval at least 30 days prior to the project expiration date. If the study is complete, the Completion/Withdrawal Form may be submitted in lieu of the Annual Exempt Form
6. Maintain all research records for a period of seven years from the project completion date.
7. Utilize all approved research documents located within the attached files section of eCompliance. These documents are highlighted green.

If you are offering subject payments and would like more information about research participant payments, please click here to view the MU Business Policy and Procedure:
http://bppm.missouri.edu/chapter2/2_250.html

If you have any questions, please contact the IRB at 573-882-3181 or irb@missouri.edu.

Thank you,
MU Institutional Review Board

APPENDIX B: INITIAL EMAIL TO MATHEMATICS DEPARTMENT HEADS

Teacher Name,

I am a mathematics education doctoral candidate at the University of Missouri working under the direction of James Tarr. The reason I am contacting you today is that I thought you might be able to help direct me in recruiting teachers as participants in my dissertation (or perhaps you are interested in this yourself).

For my dissertation, I am recruiting both middle level and high school teachers with experience in teaching statistics content that includes data analysis—this includes teachers of all possible courses that might include some statistics content. For instance, it could include calculating summary statistics, examining dotplots, histograms, and boxplots, comparing such plots, and making inferences about what such representations of samples might mean about the population. Knowledge of formal statistical tests (t-tests, chi-square, confidence intervals, etc.) are not important for my study. I am hoping that you can help direct me to teachers whom you think would 1) meet my criteria (I am thinking that I have cast a wide net of possibilities), and 2) be interested in participating in my study.

For my study, I am interested in understanding how teachers' knowledge of statistical concepts influences how they engage in statistical tasks that employ what is commonly referred to as "informal inferential reasoning." This type of reasoning is a special type of reasoning with statistics that is *informal* because statistical tests (i.e., p-values, t-tests, confidence intervals) are not used when making claims about data. A secondary interest is also related to what knowledge teachers have about how students might approach specific tasks of this type and how teachers might respond to specific types of student responses to these tasks. In order to understand these things, I have collected and developed some tasks to engage teachers in thinking about these topics for use during one-on-one interview sessions.

Specifically, I would like to have a range of teachers included in my study. So, a few middle school mathematics teachers, some high school mathematics teachers who do not teach a statistics course (AP or non-AP), and high school teachers of both AP and non-AP Statistics. My target is a total of 10 teachers and I plan to carry out the interviews in July or August. Do you have any thoughts on who might interested in participating? I greatly appreciate any help you can give me!

Thanks again for your time and any help or advice in recruiting some teachers. I look forward to hearing back from you.

Chris Engledowl
Mathematics Education Doctoral Candidate
University of Missouri
119 Townsend
ce8c7@mail.missouri.edu

APPENDIX C: INITIAL EMAIL TO MATHEMATICS TEACHERS

Teacher Name,

I am a mathematics education doctoral candidate at the University of Missouri working under the direction Dr. James Tarr. I am searching for teachers who might be interested in participating in my dissertation. I am studying how teachers reason about statistics and your background in teaching ____ (mathematics course) gives you something unique to add to my study. I have included some basic information below about my study and what it would require of you.

What am I studying? Generally speaking, I am interested in understanding the types of statistical knowledge teachers have and how it supports their engagement in a specific type of statistical reasoning termed *informal inferential reasoning*. This term means using informal knowledge (i.e., *not* statistical tests, confidence intervals, etc.) to make inferences from samples about some larger population. Further, I am interested in how teachers' knowledge and engagement in tasks that require this type of reasoning is related to their knowledge of students. Understanding these relationships is important for developing professional development programs that can more efficiently aid teachers in engaging their own students in this type of reasoning.

Who can participate? Anyone who has taught statistics content that involved some type of data analysis. For instance, calculating summary statistics and making decisions about the data based on those statistics. Moreover, I am specifically wanting to include teachers who have taught ____ (course) in my study so I hope you will consider participating.

Monetary Compensation: I do not *yet* have funds acquired but I am actively seeking funds and my advisor is also aiding in my endeavor.

Expected Time Commitments and Responsibilities:
1. Background survey and assessment taken online: no more than 60 minutes
2. Completion of four *informal inferential reasoning* tasks with follow-up questions regarding anticipating students responses and how you might address specific student responses during two one-on-one interviews: approximately 90 minutes per interview (total of 3 hours)
3. Possible follow-up questions at a later point for verification purposes: approximately 15 minutes

Total expected time commitment: 4 hours

Thank-you for considering speaking further with me about participating in my study. If you have any questions, I am more than happy to provide more detail. I look forward to hearing from you.

Chris Engledowl
Mathematics Education Doctoral Candidate
University of Missouri
119 Townsend
ce8c7@mail.missouri.edu

APPENDIX D: PARTICIPANT CONSENT FORM

Campus Institutional Review Board
Informed Consent

Researcher's Name(s): Christopher Engledowl (Principal Investigator), James E. Tarr (Advisor/Co-Investigator)

Researcher's Contact Information:
119 Townsend Hall
University of Missouri - Columbia
ce8c7@mail.missouri.edu

Project Title: Mapping Teachers' Cognitive Structures that Support Informal Inferential Reasoning

YOU ARE BEING ASKED TO VOLUNTEER TO PARTICIPATE IN A RESEARCH STUDY

You are being asked to participate in a research study. This research is being conducted to help understand how teachers' knowledge of statistics is related to the informal inferential reasoning (IIR) and how that is related to one aspect of *pedagogical content knowledge*, namely knowledge of students. This research is important because IIR is an important precursor to formal inferential reasoning (i.e., using hypothesis testing, *p*-values, or confidence intervals to make inferences about data). Moreover, a better understanding of the relationships this study seeks to examine will aid in further research into teacher knowledge that will help in developing both professional development and teacher preparation courses. When you receive an invitation to participate in the research, you have the right to be informed about the study procedures so that you can decide whether you want to consent to participation. This form may contain words that you do not know. Please ask the researcher to explain any words or information that you do not understand.

You have the right to know what you will be asked to do so that you can decide whether or not to be in the study. Your participation is voluntary. You do not have to be in the study if you do not want to. You may refuse to be in the study and nothing will happen. If you want to withdraw from the study, you may stop at any time without penalty or loss of benefits to which you are otherwise entitled.

WHY AM I DOING THIS STUDY?

The purpose of this research is to examine (1) the types of statistical knowledge middle level and high school mathematics teacher have, (2) how that knowledge supports their engagement in informal inferential reasoning, and (3) how their knowledge and engagement in this type of reasoning is related to their knowledge of students. Through a better understanding of these relationships, professional development programs can be constructed that will more efficiently aid teachers in engaging their own students in this type of reasoning. Furthermore, it will aid in developing teacher preparation courses and will inform further research into these relationships.

Why IIR? IIR is a special type of reasoning that involves making inferences about data through the use of various representations of the data (i.e., summary statistics, plots). It is informal because it does not make use of hypothesis tests, *p*-values, confidence intervals, etc., which are considered to be used for formal inferential reasoning. Several important documents (such as the *Common Core State Standards for Mathematics* and the *Guidelines for Assessment and Instruction in Statistics Education* report by the American Statistical Association) outlining what students should know regarding statistics content by the end of high school have included informal inferential reasoning as an important thread throughout middle school and high school.

HOW LONG WILL I BE IN THE STUDY?

Data will be collected for about one month and will require approximately **4 hours** of your time. There are three phases to this study.

- Phase I: Background survey and background content assessment to be taken online. Estimated time: **30 to 60 minutes**
- Phase II: Two one-on-one interviews in which informal inferential reasoning tasks will completed and discussed. Estimated time: 60 to 90 minutes, total of **2 to 3 hours**
- Phase III: A final email contact to clarify any responses made during Phase II interviews. Estimated time: **15 minutes**.

WHAT AM I BEING ASKED TO DO?

You will be asked to complete:

- a **background survey** regarding your teaching experience
- a **background assessment** that includes 20 multiple choice items to assess your statistical content knowledge,
- **4 informal inferential reasoning tasks** to completed during two separate interview sessions, and follow-up questions related to those informal inferential reasoning tasks that inquire about your knowledge of students.
- **A final email contact** that may ask you to clarify statements you made about tasks completed during interview sessions.

Interview sessions will be video and audio recorded. The audio recording only serves as a backup in case the video camera malfunctions. A video camera will record *only what is being written* so that faces will not appear. The purpose of the video recording is to 1) capture the dialogue between the researcher and participant for transcription and analysis purposes, and 2) to capture any gestures made to what was written in order to clarify statements such as "over here", or "when I wrote that" during analysis.

HOW MANY PEOPLE WILL BE IN THE STUDY?

Between **8 and 12** people are expected to be involved in the study.

WHAT ARE THE BENEFITS OF BEING IN THE STUDY?

Your participation will benefit (1) other researchers of teachers' statistical knowledge, (2) people who design professional development programs that include statistics content, (3) instructors of preservice teachers, and (4) teachers of middle level and high school students who teach statistics content. Personal benefits to you may include new experiences of engaging in non-routine statistics tasks as well as prompts during interviews regarding the teaching of such tasks. Such experiences may aid in improving your own understanding of, and teaching of, statistics.

WHAT ARE THE RISKS OF BEING IN THE STUDY?

Your participation in this study is not expected to cause you any risks greater than those encountered in everyday life. However, you may experience feelings of frustration associated with solving non-routine statistics problems. Moreover, since interviews will be video and audio recorded, there is some risk of being identified. The section below will describe the methods for maintaining confidentiality. There are

no anticipated circumstances under which your participation will be terminated without regard to your consent. Further, there are no negative consequences should you choose to withdraw from the research.

CONFIDENTIALITY

Your identity and participation will remain confidential. Only myself (Chris Engledowl) and my advisor (James E. Tarr) will have access to data collected about you during this study and we are the only ones who will know your identity.

All records collected through video, audio, and written formats will be stored in password protected files and will only be accessed either locally or through a secure network. Once data analysis begins, all participants will be given pseudonyms and any published content will make use of pseudonyms in order to maintain confidentiality. Moreover, any reference to participant background information will be appropriately disguised so as not to reveal any identifying information, such as referring to "a high school mathematics teacher in the Midwest" as opposed to "a high school mathematics teacher in _____ school/district." Raw data will only be accessible to members of the research team.
Per IRB requirements, data will be stored for 7 years and then be destroyed.

WHAT WILL I RECEIVE FOR BEING IN THE STUDY?

Currently, there is no monetary compensation associated with participation in this study. There are no costs to you for participating in this study.

WILL THE RESEARCHER TELL ME IF SOMETHING CHANGES IN THE STUDY?

Informed Consent is an ongoing process that requires communication between the researcher and participants. The participant should comprehend what they are being asked to do so that they can make an informed decision about whether they will participate in the research study. You will be informed of any new information discovered during the course of this study that might influence your health, welfare, or willingness to be in this study.

WHERE CAN I LEARN MORE ABOUT PARTICIPATING IN RESEARCH?

The Campus Institutional Review Board offers educational opportunities to research participants, prospective participants, or their communities to enhance their understanding of research involving human participants, the IRB process, the responsibilities of the investigator and the IRB. You may access the Campus IRB website to learn more about the human subject research process at http://www.research.missouri.edu/cirb/index.htm

WHO DO I CONTACT IF I HAVE QUESTIONS, CONCERNS, OR COMPLAINTS?

Please contact Christopher Engledowl (ce8c7@mail.missouri.edu) if you have questions about the research. Additionally, you may ask questions, express concerns or complaints to the research team.

Investigator Contact Information

Christopher Engledowl (Principal Investigator)
119 Townsend Hall
University of Missouri

ce8c7@mail.missouri.edu

James E. Tarr (Advisor/Co-Principal Investigator)
303 Townsend Hall
University of Missouri
573.882.4034
TarrJ@missouri.edu

WHO DO I CONTACT IF I HAVE QUESTIONS ABOUT MY RIGHTS, CONCERNS, COMPLAINTS OR COMMENTS ABOUT THE RESEARCH?

You may contact the Campus Institutional Review Board if you have questions about your rights, concerns, complaints or comments as a research participant. You can contact the Campus Institutional Review Board directly by telephone or email to voice or solicit any concerns, questions, input or complaints about the research study.

Campus Institutional Review Board
483 McReynolds Hall
Columbia, MO 65211
573-882-9585
E-Mail: umcresearchcirb@missouri.edu
Website: http://www.research.missouri.edu/cirb/index.htm

WILL I GET A COPY OF THIS FORM TO TAKE WITH ME?

A copy of this Informed Consent form will be given to you before you participate in the research.

SIGNATURES

I have read this consent form and my questions have been answered. My signature represents my desire to participate in the study. I know that I can withdraw from the study at any time without any problems.

_____ _____

Your Signature Date

FULL DISCLOSURE

Video and audio recordings will not capture faces, but voices and gestures will be captured. Video and audio clips may be used in research presentations or for professional development purposes. In these cases, only short clips will be used, pseudonyms will be used to refer to individuals, and this type of data will not be used for professional development in participants' own school districts in order to minimize the likelihood of identification by voice or some bodily feature that may be visible on the video. If you *do not* give consent for video and audio of your interview sessions to be used for the purposes of research presentations and professional development, please check the box below:

☐ I *do not* give consent for video and audio of myself to be used for research presentations or for professional development.

APPENDIX E: BACKGROUND SURVEY

Background Survey

1. Name (First and Last):

2. What grade level did you teach in the last academic year?

3. (a) What courses did you teach at this grade level?

(b) How long have you been teaching each of these courses?

4. How long have you been teaching at this grade level?

5. How long have you been teaching mathematics?

6. How many times have you taught a course specifically designed for statistics content?

7. How many times have you taught a course that involved at least one unit of statistics?

8. (a) When is the last time you taught a course that involved statistics content, and specifically involved data analysis—data analysis referring to calculating summary statistics from a data set, creating plots of the data, and making decisions and inferences about the data from this information.

(b) What was the course(s) where this content was taught?

(c) What grade level was the course(s)?

10. Please describe the nature of the statistics content that you taught. For example: What statistics content was included? How was data analyzed? What kinds of decisions and inferences were made from that analysis? How much, or long, did you spend with data analysis and making decisions and inferences about the data (e.g., days, weeks)?

11. On a scale of 1 to 5, with 5 being the highest, how confident were you when teaching this content?

12. What college level courses, if any, have you taken that involved a significant amount of statistics content, and at what level are they (e.g., undergraduate/graduate)?

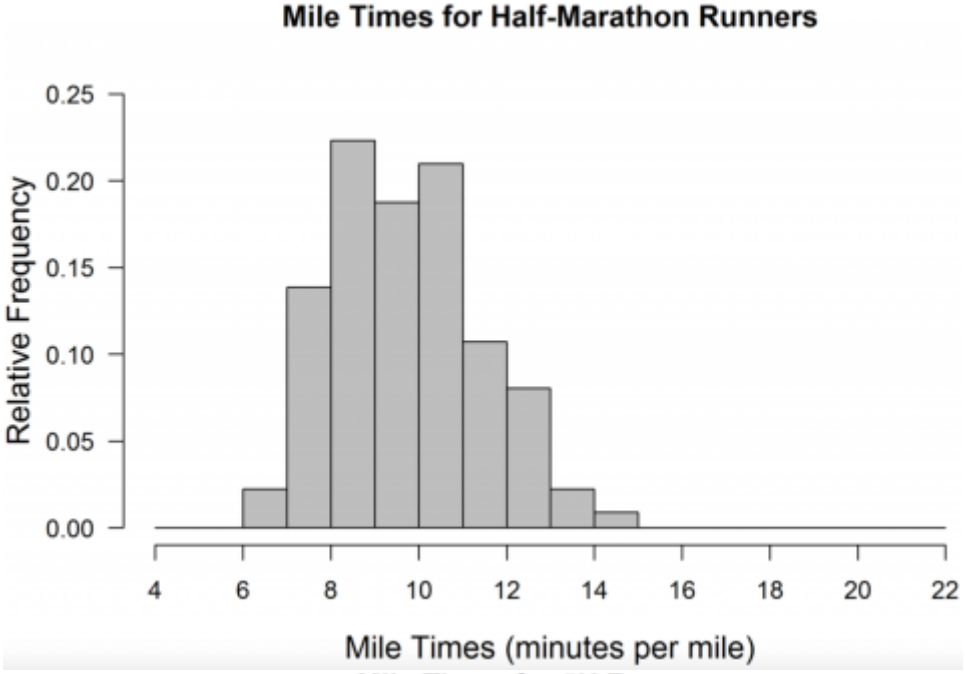
13. What professional development, if any, have you attended specifically for statistics content?

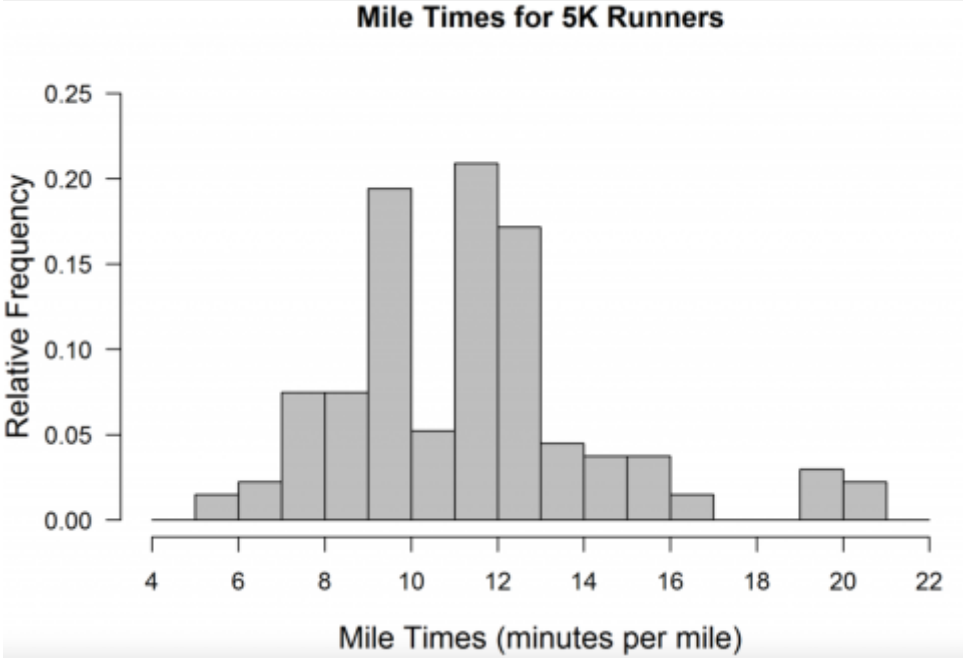
14. What college degree(s) do you have?

APPENDIX F: LOCUS TASKS

New Year's Day Race

The city of Gainesville hosted two races last year on New Year's Day. Individual runners chose to run either a 5K (3.1 miles) or a half-marathon (13.1 miles). One hundred thirty-four people ran in the 5K, and 224 people ran the half-marathon. The mile time, which is the average amount of time it takes a runner to run a mile, was calculated for each runner by dividing the time it took the runner to finish the race by the length of the race. The histograms show the distributions of mile times (in minutes per mile) for the runners in the two races.





(a) Jaron predicted that the mile times of runners in the 5K race would be more consistent than the mile times of runners in the half-marathon. Do these data support Jaron's statement? Explain why or why not.

(b) Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not.

(c) Recall that individual runners chose to run only one of the two races. Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K? Explain why or why not.

Please refer to part (a), which asks about Jaron's statement, when responding to parts (d) and (e).

(d) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(e) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Please refer to part (b), which asks about Sierra's statement, when responding to parts (f) and (g).

(f) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(g) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Please refer to part (c) when responding to parts (h) and (i).

(h) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(i) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Tomatoes and Fertilizer

A farmer conducted an experiment to find out whether a new type of fertilizer would increase the size of tomatoes grown on his farm. The farmer randomly assigned 10 tomato plants to receive the new fertilizer and 10 tomato plants to receive the old fertilizer. All other growing conditions were the same for the 20 plants. At the end of the experiment, the mean weight of tomatoes grown with the new fertilizer was 0.4 ounce heavier than the mean weight of the tomatoes grown with the old fertilizer.

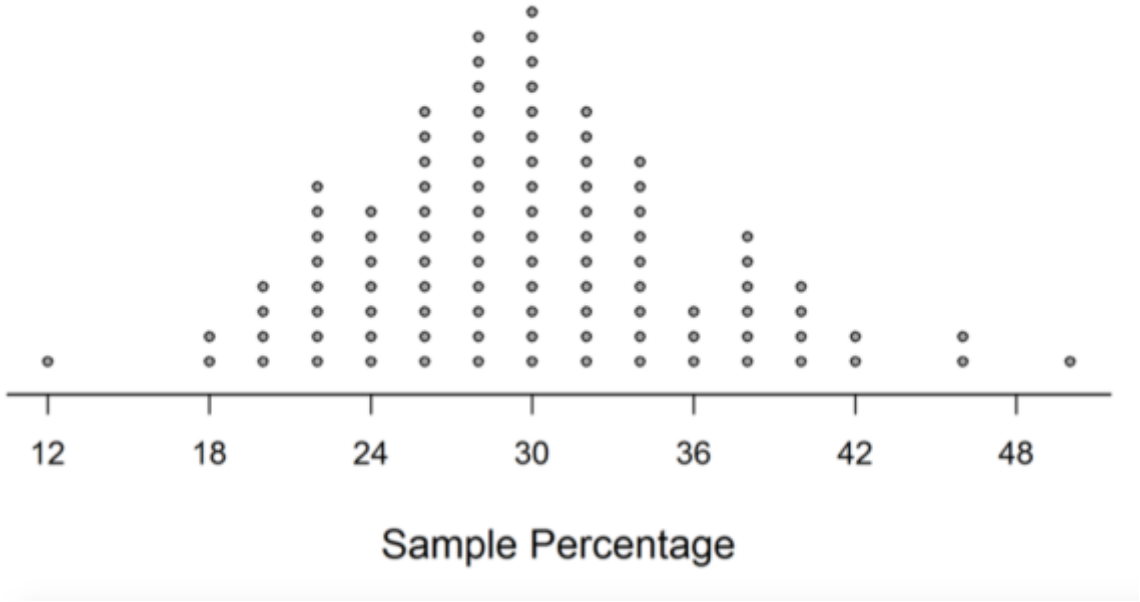
a) Based on the results, the farmer is convinced that the new fertilizer produces heavier tomatoes on average. Briefly explain to the farmer why simply comparing the two means is not enough to provide convincing evidence that the new fertilizer produces heavier tomatoes.
b) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.
c) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Extended School Day

Stella saw the following headline in a national newspaper: "30 Percent of High School Students Favor Extended School Day." She wondered if the percentage of students at her school who favor an extended school day was less than 30 percent. To investigate, she selected a random sample of 50 students from the 1,200 students at her school and asked each student in the sample if he or she favors an extended school day.

Only 12 of the students in the sample favored an extended school day. Because the sample percentage is (12/50)100 = 24%, Stella thinks that fewer than 30 percent of the students at her school favor an extended school day. She wonders if it would be surprising to see a sample percentage of 24 or less if the school percentage is really 30.

To see what values of the sample percentage would be expected if the school percentage was 30, she decides to use 1,200 beads to represent the population of 1,200 students. She will use a red bead to represent a student who favors an extended school day and a white bead to represent a student who does not. Stella put all the beads in a box. After mixing the beads, she selected 50 of them and computed the percentage of red beads. She put the 50 beads back in the box and repeated this process 99 more times. Then, she made the following dotplot of the 100 sample percentages:



(a) If the school percentage were actually 30%, how surprising would it be to see a sample percentage of 24% or less? Justify your answer using the dotplot.

(b) Based on her sample data, should Stella conclude that the percentage of students at the school who favor an extended school day is less than 30%? Explain why or why not.

Please refer to part (a) when responding to parts (c) and (d) below.

(c) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(d) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

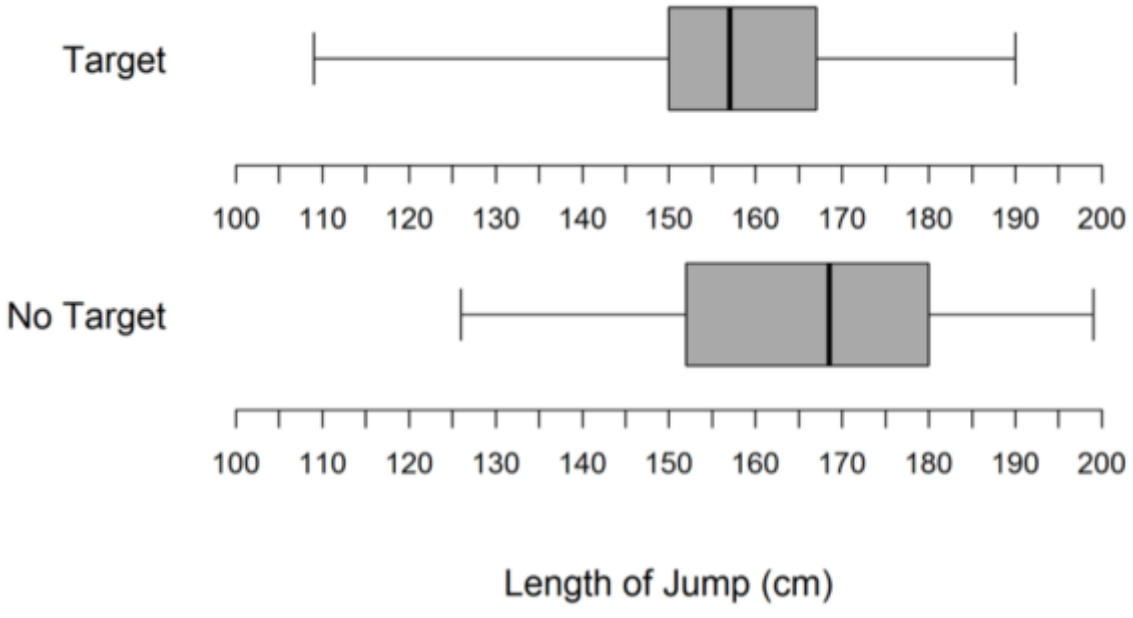
Please refer to part (b) when responding to parts (e) and (f) below.

(e) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(f) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Jumping Distances

Students wanted to investigate whether the distance a male student can jump is affected by having a target to jump toward. The students decide to perform an experiment comparing two groups. One group will have male students jumping toward a fixed target, and the other group will have male students jumping without a fixed target. There are 28 male students available for the experiment.
After randomly assigning each male student to one of the two groups, data were collected on the length (in centimeters) of the jump for each male student. The data for 28 male students are summarized in the boxplots below.



(a) Based on the boxplots, how do the lengths of the jumps compare for the two groups. Make sure to compare center, variability, and shape.

(b) Write a concluding statement to address whether the distances the male students jumped were affected by having a target? Justify your conclusion.

Please refer to part (a), regarding center, spread, and shape, when responding to parts (c) and (d).

(c) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(d) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

Please refer to part (b), regarding a concluding statement, when responding to parts (e) and (f).

(e) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

(f) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

APPENDIX G: INTERVIEW PROTOCOL

New Year's Day Race Interview Protocol

Initial question before moving into specific parts:

- *Before we get started, is there anything about the problem or the context that you don't understand or need clarification about?*

Task part (a) Jaron predicted that the mile times of runners in the 5K race would be more consistent than the mile times of runners in the half-marathon. Do these data support Jaron's statement? Explain why or why not.

Desired response draws on variability in responses. In particular, more consistent implies that there is less variance from center. Misconceptions reported by Whitaker and Jacobbe (2014) are that students focused on differences in heights of bars, such as with responses like "there are more spikes in the graph."

Phase 1 of interview questions will be "non-directive follow-up questions" (Goldin, 1997, p. 45) to elicit verbalizations and require reflection (Ginsburg, 1981):
Examples:
Can you show you me what you mean on the graphs?
Can you explain to me how you came to that conclusion?
What makes you think that? Can you say more about that?
Moreover, if someone confuses parts (a) and (b), allow them to continue and then use the questions below to help determine if there was a reading error or if there is a misconception.

Contingency 1: Correct response drawing on the range (or other measure of spread: e.g., standard deviation) as a measure of variability and clearly drawing a comparison between the two groups using language such as more than, less than, etc.
Questions:

- *What do you mean by range?* (require reflection, non-directive follow-up)
- *How did you determine what the range is?* (reflection, challenge)
- *Why do you think the range is appropriate to use to argue against Jaron's claims?* (require reflection, non-directive follow-up, challenge response)
- *Is there any possibility that the data **do** support Jaron's claims?* (challenge response)

Contingency 2: Seems a correct response but comparative language is implicit -- one group is mentioned and the other is implied by language such as more than, less than, etc. Along with similar questions as in Contingency 1:

- *What do you mean by "more than" (or whatever language is used)?* (require reflection, non-directive follow-up)
- *What about _____ (the other group not mentioned)?* (challenge response)

Contingency 3: Incorrect response that correctly claims that the data **don't** support Jaron's statement, but that incorrectly focuses on variability in the heights of bars.

Whitaker and Jacobbe (2014) note this as a common incorrect response, claiming that many students "concluded that variability in bar heights implies inconsistent mile times", such as one student who said, "No, because there are more spikes in the graph for the 5K, and less in the graph for the half marathon" (p. 14).

- *What do you mean by "spikes" (or whatever term is used)?*
- *Why do you think that the "spikes" (or other term) are appropriate to use to argue against Jaron's claims?*
- *Is there any possibility that the data **do** support Jaron's statement?*

All other possible contingencies will result in similar questions such as:

- *What do you mean by ____ (graphical feature/statistic)?*
- *Why do you think that ____ (graphical feature/statistic) is appropriate to use?*
- *Is there any possibility that the data **do/don't** support Jaron's statement?*

Task part (b) Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not.

Contingency 1: Correct response claiming no support and using a measure of center to support the argument. Whitaker and Jacobbe (2014) claim "many students" made such claims.

- *What do you mean by ____ (whichever measure of center is used)?*
- *How did you determine what the center is?*
- *Why do you think that ____ (measure of center) is appropriate to use to argue against Sierra's statement?*
- *Is there any possibility that the data **do** support Sierra's statement?*

Contingency 2: Response claims no support but uses a measure of spread as the argument. Whitaker and Jacobbe claimed the following student response was representative of a common mistake: "Yes, because the times stayed consistent during the 8–11 mile times" (2014, p. 14). Whitaker and Jacobbe claimed that this kind of response "attends to only part of the data in one group, does not make an appropriate comparison with the 5K runners group, and seems to be based on reasoning using the mode rather than a measure of center for quantitative data such as the mean or median" (p. 14).

- *What do you mean by "stayed consistent during the 8-11 mile time"?*
- *How did you come to focus on the 8-11 mile time period?*
- *Why do you think the focus on consistency during that time period is appropriate to argue against Sierra's statement?*
- *Is there any possibility that the data **do** support Sierra's statement?*

All other possibilities will receive similar treatment with questions.

Task part (c) Recall that individual runners chose to run only one of the two races. Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K? Explain why or why not.

The desired response will claim 1) that it is not reasonable, and 2) draw on the way runners were assigned to groups to support their claim -- that because there was not *random assignment*, a causal inference cannot be made. This type of inference is one of two included in Rossman's (2008) description of types of statistical inferences -- the other is inference from a sample to a population.

Contingency 1: Correct response providing a valid reason (valid meaning addressing random assignment) why no such conclusion can be drawn. Whitaker and Jacobbe claimed that "students were imaginative and many potentially valid reasons were given" (2014, p. 15). Therefore, regardless of what reason is provided, some questions to be asked are:

- *What makes you think _____ (reason) could occur?*
- *Why would _____ (reason) cause no conclusion to be drawn?*
- *If any terms are used (e.g., random selection, random assignment) ask what is meant by them.*
- *Is there any possibility that these data **do** allow you to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K?*

Contingency 2: Response correctly claims that no such conclusion can be made, but reasoning is flawed. Whitaker and Jacobbe describe one such example response, "No, because to do that you need to have an individual run both races to compare differences in time", and further claim that 1) the type of design described by the student is not necessary, and 2) it ignores the role of random assignment (2014, p. 15).

- *Can you explain more about what you mean?* (non-directive, reflective)
- *What makes you think having an individual run both races would allow such a conclusion to be drawn?*
- *Why would you need to "compare differences in time"?*
- *Is there any other reason why such a conclusion cannot be drawn?*
- *Is there any possibility that these data **do** allow you to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K?*

Contingency 3: Response incorrectly claims that such a conclusion **can** be made using flawed reasoning. Whitaker and Jacobbe (2014) described one such example involving assumptions about the level of endurance individuals have in each of the two categories. A student claimed "Yes, if a person chose the half marathon over the 5K you can assume they are in better shape and are better runners than someone who would have chose to run less in the 5K" (p. 14).

- *Can you explain more about what you mean?*
- *Why would being a better runner allow you to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K?*
- *Is there any other reason why such a conclusion could be drawn?*

- *Is there any possibility that these data **don't** allow you to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K?*

Contingency 4: Response correctly claims that such a conclusion **cannot** be made using the same flawed reasoning as in Contingency 3: referring to people choosing which race based on their physical fitness level or endurance without mentioning random assignment.

- *Can you explain more about what you mean?*
- *What makes you think that by allowing runners to choose which race they run in, such a conclusion cannot be drawn?*
- *Is there any way the study could be altered to allow such a conclusion?*
- *Is there any other reason why such a conclusion could be drawn?*
- *Is there any possibility that these data **do** allow you to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K?*

Task parts (d), (f), and (h) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

- All responses will receive the following questions in order to bring clarity:
  - *What is it that makes those responses appropriate or inappropriate?*
- Contingency 1: A non-response, such as "I don't know."
  - *Do you think students might respond in a similar way that you did?*
  - *What are some things you think students might place their attention that you would not want them to focus on?*
- Contingency 2: Responding only with one type
  - *Can you think of any other types of responses students might have?*

Task parts (e), (g) and (i) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

- All responses will receive the following questions in order to bring clarity:
  - *Why did you choose that as the inappropriate response to talk about?*
  - *Why would you _____ (whatever they say they will do)?*
- Contingency: The inappropriate response is vague or does not address a common mistake students might make (e.g., students will struggle).
  - *What would you do if a student gave ____ (pick a response that was a common mistake) response?*

Tomatoes and Fertilizer Interview Protocol

Initial question before moving into specific parts:

- *Before we get started, is there anything about the problem or the context that you don't understand or need clarification about?*

Task part (a): Based on the results, the farmer is convinced that the new fertilizer produces heavier tomatoes on average. Briefly explain to the farmer why simply comparing the two means is not enough to provide convincing evidence that the new fertilizer produces heavier tomatoes.

Contingency 1: Correct response that refers to sampling variability and indicates that the difference could be due to chance. Some examples are "The results could have happened by random chance which means that the really good plants could have all ended up in the new group by chance and the type of fertilizer had nothing to do with the difference in mean" and "This is not enough evidence because although the new fertilizer has a heavier weight on average, it could just be from the random assignment of the groups and that one or two of the plants were better and both could have ended up in the new fertilizer group" (LOCUS website).

- *What do you mean by "random chance" (or random assignment)?*
- *Why is the idea of "random chance" (or random assignment) important in your explanation to the farmer?*
- *Is there another possible reason why comparing the means is not enough to justify that the new fertilizer is better?*

Contingency 2: No reference to sampling variability. Could reference such things as "sample size, the possibility of confounding variables, or the potential effect of an outlier" (LOCUS website). Some example student responses are "This is only testing 20 of all tomatoes. This is not enough to conclude about the population", "the farmer needs to consider confounding variables like weather", "the mean is not resistant to outliers, in other words, a single larger-than-average tomato may increase the mean leading us to draw incorrect conclusions" (LOCUS website).

- Sample Size:
  - *How many would you say is enough?*
  - *Is there another possible reason why comparing the means is not enough to justify that the new fertilizer is better?*
  - *Suppose it **was** enough. Could you then conclude that the difference in means is enough to say that the new fertilizer is better?*
  - *What would you say was the reason the farmer randomly assigned the plants to the fertilizer groups?*
- Confounding Variables:
  - *What do you mean by "the farmer needs to consider confounding variables like weather"?*
  - *Do you think the weather could impact the plants in different ways based on the type of fertilizer?*
  - *Is there another possible reason why comparing the means is not enough to justify that the new fertilizer is better?*

    - *Suppose there were **no confounding variables**. Then, could you conclude that the difference in means is enough to conclude the new fertilizer is better?*
    - *What would you say was the reason the farmer randomly assigned the plants to the fertilizer groups?*
- Outlier effects:
    - *Are you saying the farmer should use a different measure for center? If so, which one?*
    - *[assuming another measure is proposed] If the farmer were to use ____ instead of the mean, would you then say that comparing the two ____ is enough to claim that the new fertilizer is better?*
    - *Is there another possible reason why comparing the means is not enough to justify that the new fertilizer is better?*
    - *[this question might already be answered by the second bullet above] Suppose there were **no outliers**. Then, could you conclude that the difference in means is enough to conclude the new fertilizer is better?*
    - *What would you say was the reason the farmer randomly assigned the plants to the fertilizer groups?*

Task part (b): What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

- All responses will receive the following questions in order to bring clarity:
    - *What is it that makes those responses appropriate or inappropriate?*
- Contingency 1: A non-response, such as "I don't know."
    - *Do you think students might respond in a similar way that you did?*
    - *What are some things you think students might place their attention that you would not want them to focus on?*
- Contingency 2: Responding only with one type
    - *Can you think of any other types of responses students might have?*

Task part (c): Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

- All responses will receive the following questions in order to bring clarity:
    - *Why did you choose that as the inappropriate response to talk about?*
    - *Why would you _____ (whatever they say they will do)?*
- Contingency: The inappropriate response is vague or does not address a common mistake students might make (e.g., students will struggle).
    - *What would you do if a student gave ____ (pick a response that was a common mistake) response?*

Extended School Day Interview Protocol

Initial question before moving into specific parts:

- *Before we get started, is there anything about the problem or the context that you don't understand or need clarification about?*

Task part (a) If the school percentage were actually 30%, how surprising would it be to see a sample percentage of 24% or less? Justify your answer using the dotplot.

**Note*: Part (a) could be further addressed in part (b), and part (b) could be further addressed in part (a). Look in both places before questioning further.

Contingency 1: Correct response (could be addressed in part b as well) indicating that it would *not* be surprising (given the sample size and population proportion) and uses the simulated sampling distribution to justify the response. Some examples are "22/100=.22. It would not be surprising, as, from the dotplot, the probability of seeing a sample percentage of 24% or less is .22, which is fairly likely.", and "22 of the 100 times resulted in a sample percentage of 24% or less. That means 22% resulted in this and therefore it wouldn't be all that surprising to see a sample percentage of 24% or less because it is almost a 1 in 4 chance" (LOCUS website).

- *What do you mean by "sample percentage"?*
- *Why did you divide 22 by 100?*
- *How did you get 0.22?*
- *What do you mean by "the probability of seeing a sample percentage of 24% or less is 0.22?*
- *What do you mean by "22 of the 100 times resulted in a sample percentage of 24% or less?*
- *Can you show me on the plot?*
- *What do you mean by "fairly likely"?*
- *Why are the percentages/probabilities you are referring to important to look at to answer this question?*
- *Is there any possibility that the sample percentage of 24% or less* ***is surprising****?*

Contingency 2: Correct response indicating it would *not* be surprising *but* explanation does not refer to the simulated sampling distribution as requested in the question stem ("using the dotplot"). An example is "not surprising at all it could happen by chance alone" (LOCUS website).

- *What do you mean by "it could happen by chance alone"?*
- *Can you show me on the plot?*
- *Why does the possibility of it happening by chance alone mean that it is not surprising?*
- *Is there another way to explain why it is not surprising?*
- *Is there any possibility that it* ***is surprising****?*

Contingency 3: Correct response that uses the sampling distribution, *but* attempts to calculate an approximate p-value and does so incorrectly. An example is "it wouldn't be very surprising to get a sample percentage of 24% or less just by chance if the population

percentage were actually 30%. The simulation shows that 22/50 = 44% probability of getting a sample percentage of 24% or lower, just by chance" (LOCUS website).

- *What do you mean by "just by chance"?*
- *Why did you include "if the population percentage were actually 30%"?*
- *How did you get that "the simulation shows that 22/50=44% …"? Can you show me?*
- *Why is such a result not "very surprising"?*
- *Is there another way to explain why it is not surprising?*
- *Is there any possibility that it **is surprising**?*

Contingency 4: Claiming no conclusions could be drawn because the sample size is too small. An example of this is "Because of the small sample size you cannot say exactly but the margin of error here seems to be of about 12 points and 24 falls well within that range" (LOCUS website).

**Note:* This kind of response also can occur in part (b).

- *What do you mean by "because of the small sample size"?*
- *What size sample is considered "small"?*
- *Why can the sample size not be smaller than such a value?*
- *Why does a small sample mean that you "cannot say exactly"?*
- *What if the sample size **was** large enough? Then what would you conclude?*

Contingency 5: Claiming that because the dotplot appears normally distributed, there must be about 16% of the data in the left, which is not surprising. Although this response is still somewhat viable, the sampling distribution is not being appropriately used since the exact proportion can be calculated. Moreover, it makes some assumptions about the location of the cutoff of 24%.

**Note:* This kind of response also can occur in part (b).

- *What do you mean by 16% and that the data appear normally distributed?*
- *Why is it important for you to refer to a normal distribution to answer this question?*
- *Is there another way to explain why it is not surprising?*
- *What if the data were **not** normally distributed? How would you then respond?*
- *[If response is something like "I wouldn't be able to"] Do you have enough information with the dotplot to find an exact percentage instead of a theoretical on based on a normal model?*
- *Is there any possibility that it **is surprising**?*

Task part (b) Based on her sample data, should Stella conclude that the percentage of students at the school who favor an extended school day is less than 30%? Explain why or why not.

**Note:* Similar responses could occur that are incorrect but still consistent with an incorrect response in part (a), making them internally consistent within an individual. Questions would be similar in such a case but language would be slightly altered to fit the incorrect (but consistent) response.

Contingency 1: Correct response claiming Stella *should not* conclude that the percentage is less than 30% and justifying it based on either the difference not being statistically significant (in which case an explanation should be provided for making such a decision) or that sampling variability could explain the difference. An example is "No, she should not conclude this. The probability of getting 24% or less is fairly high, so there is not sufficient evidence to prove that the students at her school who favor an extended school day is less than 24%. In addition, she only took one sample, so we expect to see sampling variability among the possible samples" (LOCUS website, emphasis in original).

- *What do you mean by "we expect to see sampling variability among the possible samples"?*
- *Why does expecting "to see sampling variability" mean that she should not conclude that the percentage is less than 30%?*
- *Why do you keep referring to the 24% value? Why is the 24% value important in your explanation?*
- *What does the dotplot represent? What does each dot represent?*
- *Is there another way to explain why Stella should not conclude this?*
- *Is there a possible reason why Stella **should** conclude that the percentage is less than 30%?*

Contingency 2: Assuming the population **is** centered around 30% instead of recognizing the sampling distribution as representing what might happen **if** it were actually 30%. Some examples of this are "No, because there are about an equal number of model sample percentages above 30%, and also 3/20 of the data is at 30%. This means there is about an equal or better chance the percent is at or above 30%", and "No the percent of students who favor the extended school day is right around 30% according to the dotplot" (LOCUS website).

**Note:* This error may occur even after responding correctly in part (a).

- *What do you mean by "an equal or better chance the percent is at or above 30%"?*
- *Why is the 30% important? Can you explain more?*
- *Why does the data being "right around 30%" mean that Stella should **not** conclude that the percentage is less than 30%?*
- *Can you show me on the dotplot?*
- *What does the dotplot represent? What does each dot represent?*
- *Is there another way to explain why Stella should not conclude this?*
- *Is there a possible reason why Stella **should** conclude that the percentage is less than 30%?*

Contingency 3: Perceiving the data points in the sampling distribution as representing individuals. An example is "No. There are 47 sample percentages below the 30% mark. This, then makes it unfair to conclude that most student at her school are under the 30% mark, when in fact only 47% are. So, she should not conclude this, there is no convincing evidence to back it up" (LOCUS website).

**Note:* In this particular response, there is also evidence of Contingency 2 above.

- *What do you mean by "when in fact only 47% are"?*
- *How did you get the 47% value?*

- *What do you mean by "most students at her school are under the 30% mark"? Can you say more about how you came to that conclusion?*
- *Can you show me on the plot?*
- *What does the dotplot represent? What does each dot represent?*
- *What do you think would constitute "convincing evidence"?*
- *Is there another way to explain why Stella should not conclude this?*
- *Is there a possible reason why Stella **should** conclude that the percentage is less than 30%?*

Contingency 4: Responding that Stella **should** conclude the percentage is less than 30% *because* most data points in the simulated sampling distribution are below 30%. Some examples are "Yes she should because that was her results" and "yes, because most of the data lies below 30%" (LOCUS website).

- *What do you mean by "because that was her results"?*
- *What do you mean by "most of the data lies below 30%"?*
- *Can you show me on the plot?*
- *Why is the 30% value important for your explanation?*
- *What does the dotplot represent? What does each dot represent?*
- *Is there another way to explain why Stella **should** conclude this?*
- *Is there a possible reason why Stella **should not** conclude this?*

Contingency 5: Claiming no conclusions can be drawn because the sample size is too small. Some examples are "No, she should conduct more extensive testing with larger sample sizes" and "Stella's sample size of 50 students is very small, so it may not accurately represent the school's preference. Furthermore, since her dotplot found that the sample percentages are pretty equally distributed on both sides of 30 (the mode) she can't say anything conclusively" (LOCUS website).
**Note 1:* This kind of response can also occur in part (a).
**Note 2:* The second example above also implies that the student does not understand what the sampling distribution represents. Questions from Contingency 2 above would also be asked.

- *Why do you say she should "conduct more extensive testing with larger sample sizes"? Why do you say her sample is "very small"?*
- *What size would you consider to be large enough?*
- *Why does the sample need to be at least this large?*
- *Why does having a small sample mean that she should not conclude this? Why does a small sample mean you can't draw any conclusions?*
- *What if the sample size **was** large enough? Then what would you conclude?*

Contingency 6: Misunderstanding the role of the simulation. For example, stating "No because that is extrapolation. She should collect actual data rather than use beads" (LOCUS website).

- *What do you mean by "extrapolation"?*
- *What do you mean "she should collect actual data rather than use beads?*
- *What did Stella use the beads to do?*
- *Why did Stella use the beads?*

- *How did Stella obtain the dotplot?*
- *What does the dotplot represent? What does each dot represent?*
- *Why would Stella believe the dotplot is helpful in this situation?*

Task parts (c) and (e) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

- All responses will receive the following questions in order to bring clarity:
  - *What is it that makes those responses appropriate or inappropriate?*
- Contingency 1: A non-response, such as "I don't know."
  - *Do you think students might respond in a similar way that you did?*
  - *What are some things you think students might place their attention that you would not want them to focus on?*
- Contingency 2: Responding only with one type
  - *Can you think of any other types of responses students might have?*

Task parts (d) and (f) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

- All responses will receive the following questions in order to bring clarity:
  - *Why did you choose that as the inappropriate response to talk about?*
  - *Why would you _____ (whatever they say they will do)?*
- Contingency: The inappropriate response is vague or does not address a common mistake students might make (e.g., students will struggle).
  - *What would you do if a student gave ____ (pick a response that was a common mistake) response?*

Jumping Distances Interview Protocol

Initial question before moving into specific parts:

- *Before we get started, is there anything about the problem or the context that you don't understand or need clarification about?*

Task part (a) Based on the boxplots, how do the lengths of the jumps compare for the two groups. Make sure to compare center, variability, and shape.

Contingency 1: Correct response that uses comparative language to describe differences in jumping differences for those using a target and those not, using center, variability and shape. Correct responses could include comparisons of median (or center more generally), general spread (or IQR), and general reference to shape, or they could include comparisons of proportion of data included within different quartiles, more general reference to spread within the middle 50% (without explicitly mentioning or calculating IQR), etc.

- *What do you mean by ____ (description of center/spread/shape)?*
- *Can you show me on the boxplot?*
- *How did you come up with _____ (description of center/spread/shape)?*
- *Why do you think ____ (description of center/spread/shape) is important to use when comparing the two groups?*
- *Is there another way you could compare ____ (center/spread/shape)?*

Contingency 2: Most common error was not addressing all of center, spread, and shape -- many did not address shape (LOCUS website). One example where shape is not mentioned (and spread is implicit) is "the center for target is less than the center for no target. The lowest for target is less than no target. The highest for target is less than no target" (LOCUS website).

- *What do you mean by (e.g., "center", "the lowest", "the highest")?*
- *Can you show me on the graph?*
- *Why do you think referring to center, lowest, and highest are important to use when comparing the two groups?*
- *Is there another way you could compare them?*

Contingency 3: Describing center/spread/shape but not comparing the two distributions. For instance, one student created a table with columns labeled "target" and "no-target" and included bullet points such as "median is low" and "median is high" without any comparative language.

- *What do you mean by (e.g., "median is low" and "median is high")?*
- *Can you show me on the graph?*
- *Why did you decide to use a table?*
- *Why do you think your description is important for comparing the two groups?*
- *Is there another way you could compare them?*

Contingency 4: Response focuses on confounding variables, similarity of groups, or some other feature that is not related to comparing center, spread, or shape of the two distributions.

- *What do you mean by that? Can you explain more?*
- *Why do you think that is important in this case?*
- *Suppose ____ (e.g., confounding variable, similarity of groups) was not an issue. How would you then compare the two groups in terms of center, spread, and shape?*

Task part (b) Write a concluding statement to address whether the distances the male students jumped were affected by having a target? Justify your conclusion.

Contingency 1: Response that makes a claim that is justified by the boxplot. Claiming that they were or that they were not are both acceptable (because the plots overlap) if the justification is strong. One example is "Jumpers who did not have a target tended to jump further than the jumpers who did have a target. This is so because the no target jumpers had a higher upper extreme, median, and lower extreme" (LOCUS website). Another example is "I don't think it effected anything because none of the plots showed a huge difference and both of the boxplots were pretty split up" (LOCUS website).

- *What do you mean by _____ (e.g., upper extreme, median, lower extreme)?*
- *Can you show me on the graph?*
- *Why do you think _____ (e.g., having a higher upper extreme, etc.) means that those without a target jumped further?*
- *Is there another way to explain why you think that is the case?*
- *Is it possible to conclude the opposite? That those **with** targets jump further than those without targets?*

Contingency 2: Not stating a conclusion. An example of not stating a conclusion is "there was a shorter range w/ target & the median was less than w/out the target" (LOCUS website). This statement could be justification for a conclusion, but the conclusion is missing.

- *What do you mean by ____ ? Can you say more about that?*
- *Why do you think referring to the range and median are important?*
- *What do you think your statement means about the use of a target when jumping?*
- *Is there another way to justify your conclusion?*
- *Is it possible to conclude something different?*

Contingency 3: Weak justification. One example of a weak justification is "yes, the boxplot shows that with no target the males were able to jump farther based on the size of the box and the ranges" (LOCUS website). Another example is "the distance a male student can jump is affected by having a target to jump to because the boxplot shows that the ones that had a target were closer together" (LOCUS website).

- *What do you mean by ___ (e.g., "based on the size of the box and the ranges", "closer together")?*
- *Why do you think ___ means that jumping distance was/was not affected?*
- *Is there another way to justify your conclusion?*
- *Is it possible to conclude something different?*

Contingency 4: No justification. Some examples are "yes, having a target is shown to decrease the length of jump", "they had to jump to a specific point/target so it affects the students jump by having a target", and "by having a target they went shorter distances than not having a target" (LOCUS website). The second example above also presents the possibility that the student is more focused on a study design issue as a justification than using data-based evidence. For instance, is the target beyond reach for everyone? If not, then perhaps the target caused people to not jump as far as possible. This issue can be addressed early on in the interview to make sure the participant is not focused on this.

- *What do you mean by that? Can you say more?*
- *Can you show me on the graph?*
- *Why do you think ___ (e.g., having a specific point/target) means that jumping distances were affected/shorter?*
- *[Assuming interpretation is that the target may have been within reach for some] What if the target was beyond reach for everyone? Then what would you say?*
- *Is there another way to justify your conclusion?*
- *Is it possible to conclude something different?*

Task parts (c) and (e) What responses would you expect from your students? Write down some appropriate and inappropriate responses. Indicate which are appropriate by placing a * next to them.

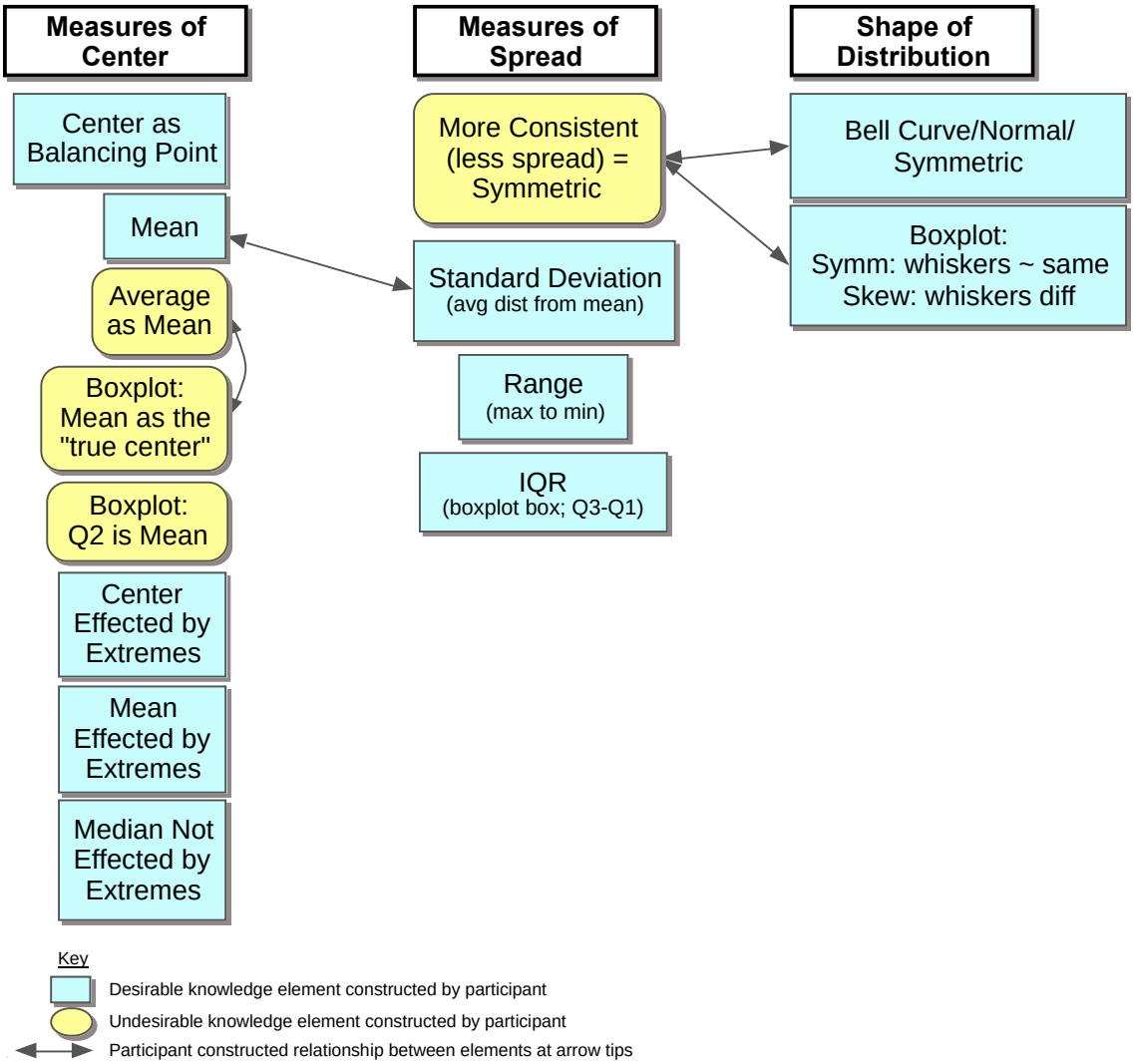
- All responses will receive the following questions in order to bring clarity:
  - *What is it that makes those responses appropriate or inappropriate?*
- Contingency 1: A non-response, such as "I don't know."
  - *Do you think students might respond in a similar way that you did?*
  - *What are some things you think students might place their attention that you would not want them to focus on?*
- Contingency 2: Responding only with one type
  - *Can you think of any other types of responses students might have?*

Task parts (d) and (f) Choose one of the inappropriate responses from above. What would you do if you gave this task to your students and one of them gave such a response?

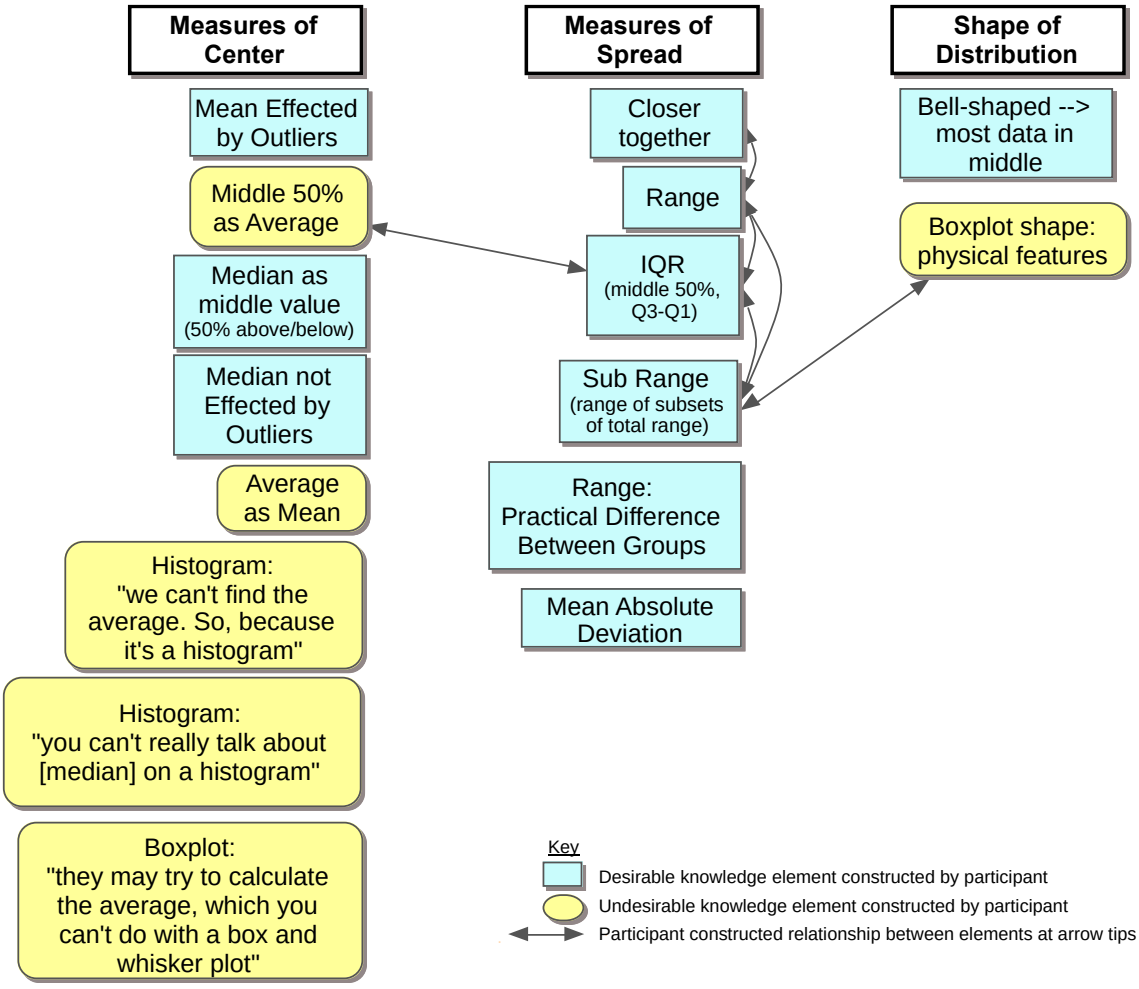
- All responses will receive the following questions in order to bring clarity:
  - *Why did you choose that as the inappropriate response to talk about?*
  - *Why would you _____ (whatever they say they will do)?*
- Contingency: The inappropriate response is vague or does not address a common mistake students might make (e.g., students will struggle).
  - *What would you do if a student gave ____ (pick a response that was a common mistake) response?*

APPENDIX H: KNOLWEDGE STRUCTURE MAPS

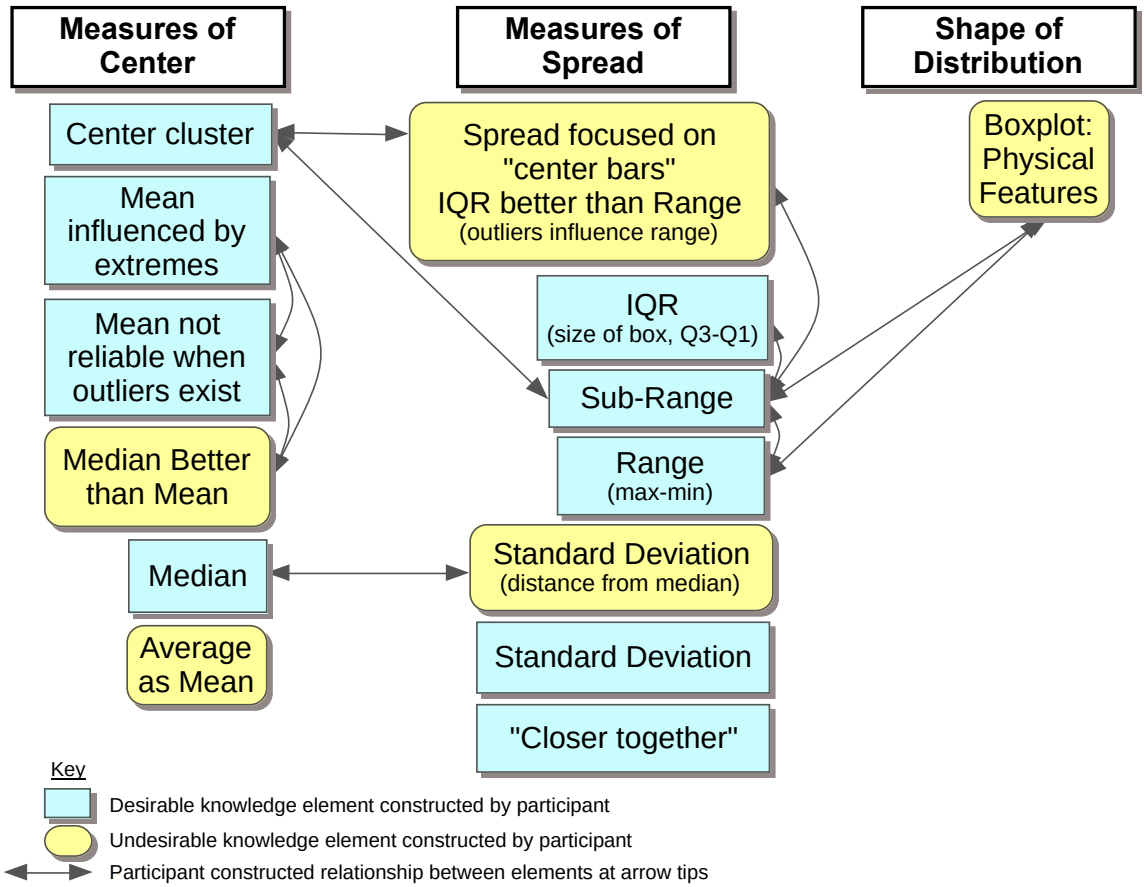
Amalia



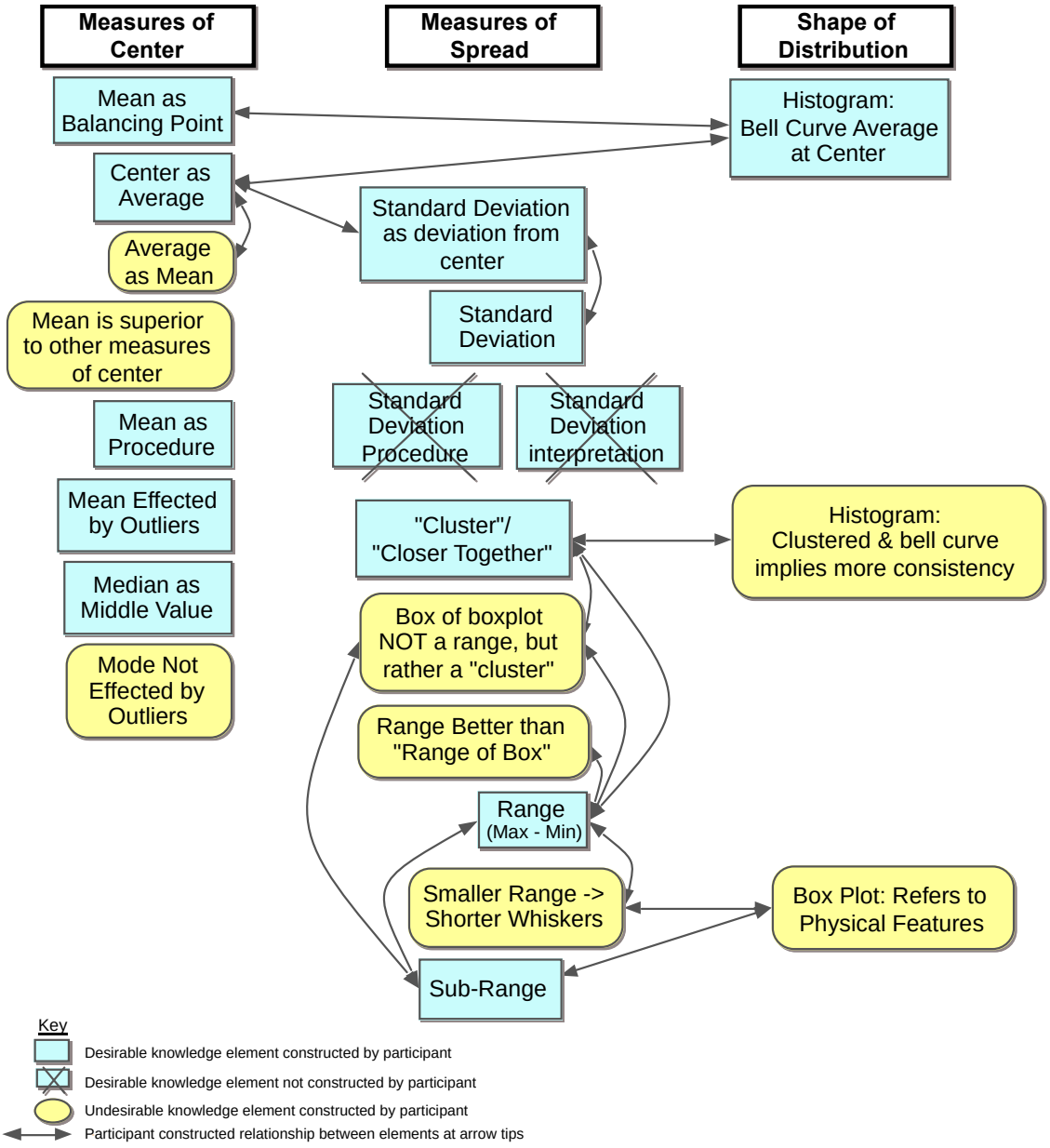
Kathy


Measures of Center
Mean Effected by Outliers
Middle 50% as Average
Median as middle value (50% above/below)
Median not Effected by Outliers
Average as Mean
Histogram: "we can't find the average. So, because it's a histogram"
Histogram: "you can't really talk about [median] on a histogram"
Boxplot: "they may try to calculate the average, which you can't do with a box and whisker plot"
Measures of Spread
Closer together
Range
IQR (middle 50%, Q3-Q1)
Sub Range (range of subsets of total range)
Range: Practical Difference Between Groups
Mean Absolute Deviation
Shape of Distribution
Bell-shaped --> most data in middle
Boxplot shape: physical features
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

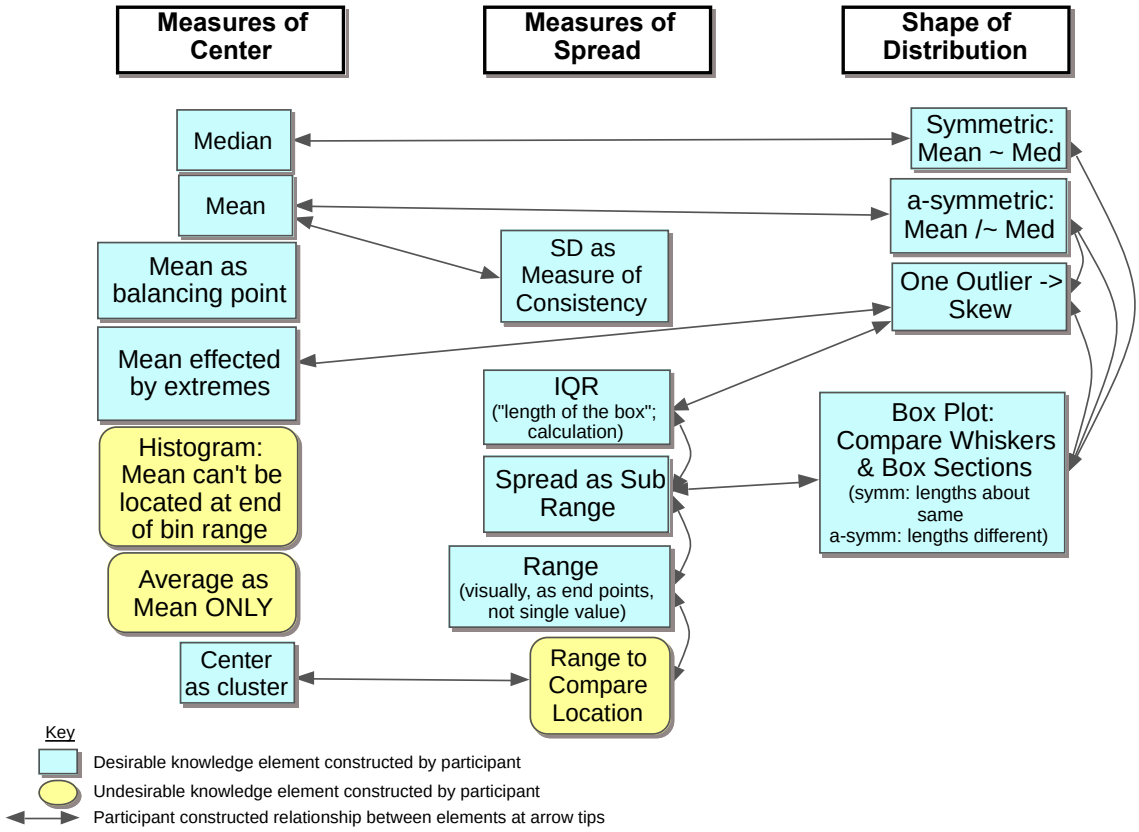
Ruby


Measures of Center
Measures of Spread
Shape of Distribution
Center cluster
Spread focused on "center bars" IQR better than Range (outliers influence range)
Boxplot: Physical Features
Mean influenced by extremes
IQR (size of box, Q3-Q1)
Mean not reliable when outliers exist
Sub-Range
Median Better than Mean
Range (max-min)
Median
Standard Deviation (distance from median)
Average as Mean
Standard Deviation
"Closer together"
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

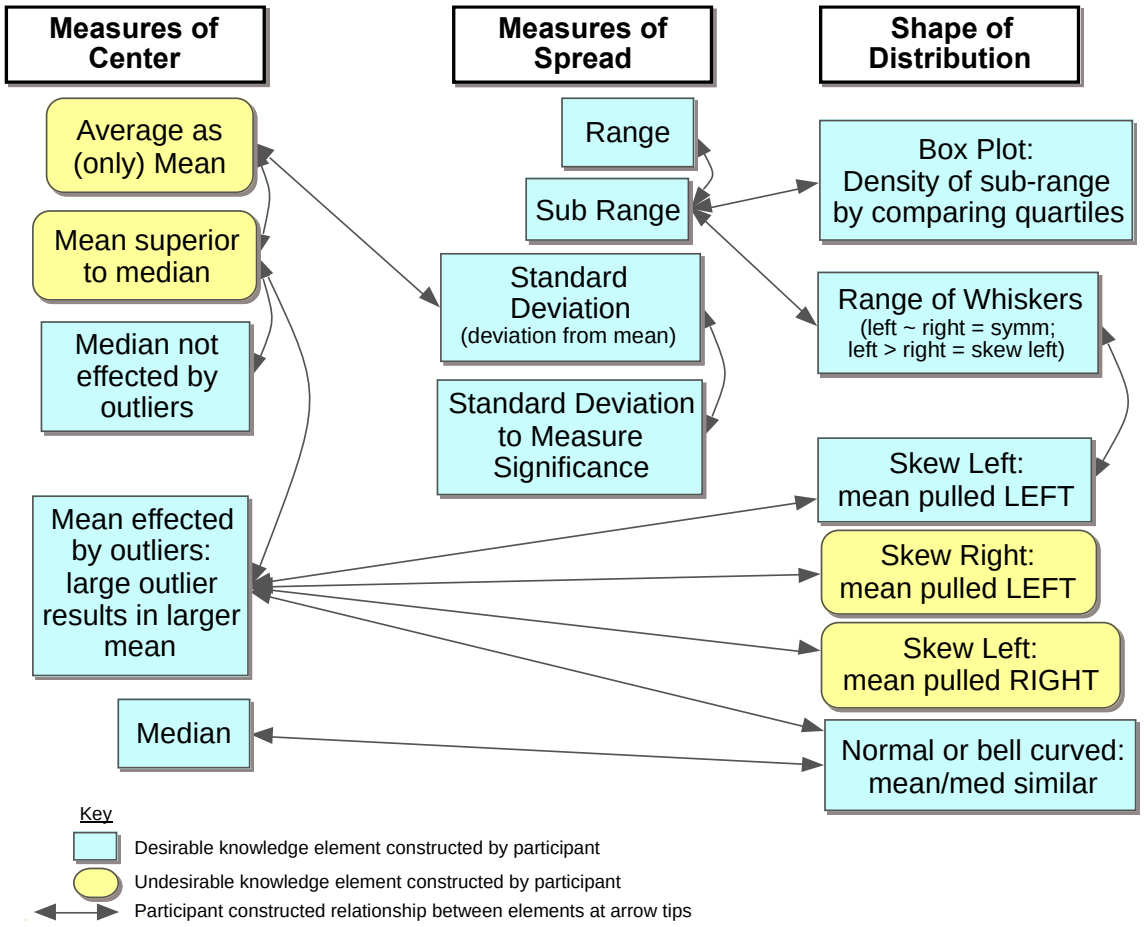
Ellie


Measures of Center
Measures of Spread
Shape of Distribution
Mean as Balancing Point
Histogram: Bell Curve Average at Center
Center as Average
Standard Deviation as deviation from center
Average as Mean
Mean is superior to other measures of center
Standard Deviation
Standard Deviation Procedure
Standard Deviation interpretation
Mean as Procedure
Mean Effected by Outliers
"Cluster"/ "Closer Together"
Histogram: Clustered & bell curve implies more consistency
Median as Middle Value
Box of boxplot NOT a range, but rather a "cluster"
Mode Not Effected by Outliers
Range Better than "Range of Box"
Range (Max - Min)
Smaller Range -> Shorter Whiskers
Box Plot: Refers to Physical Features
Sub-Range
Key
Desirable knowledge element constructed by participant
Desirable knowledge element not constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

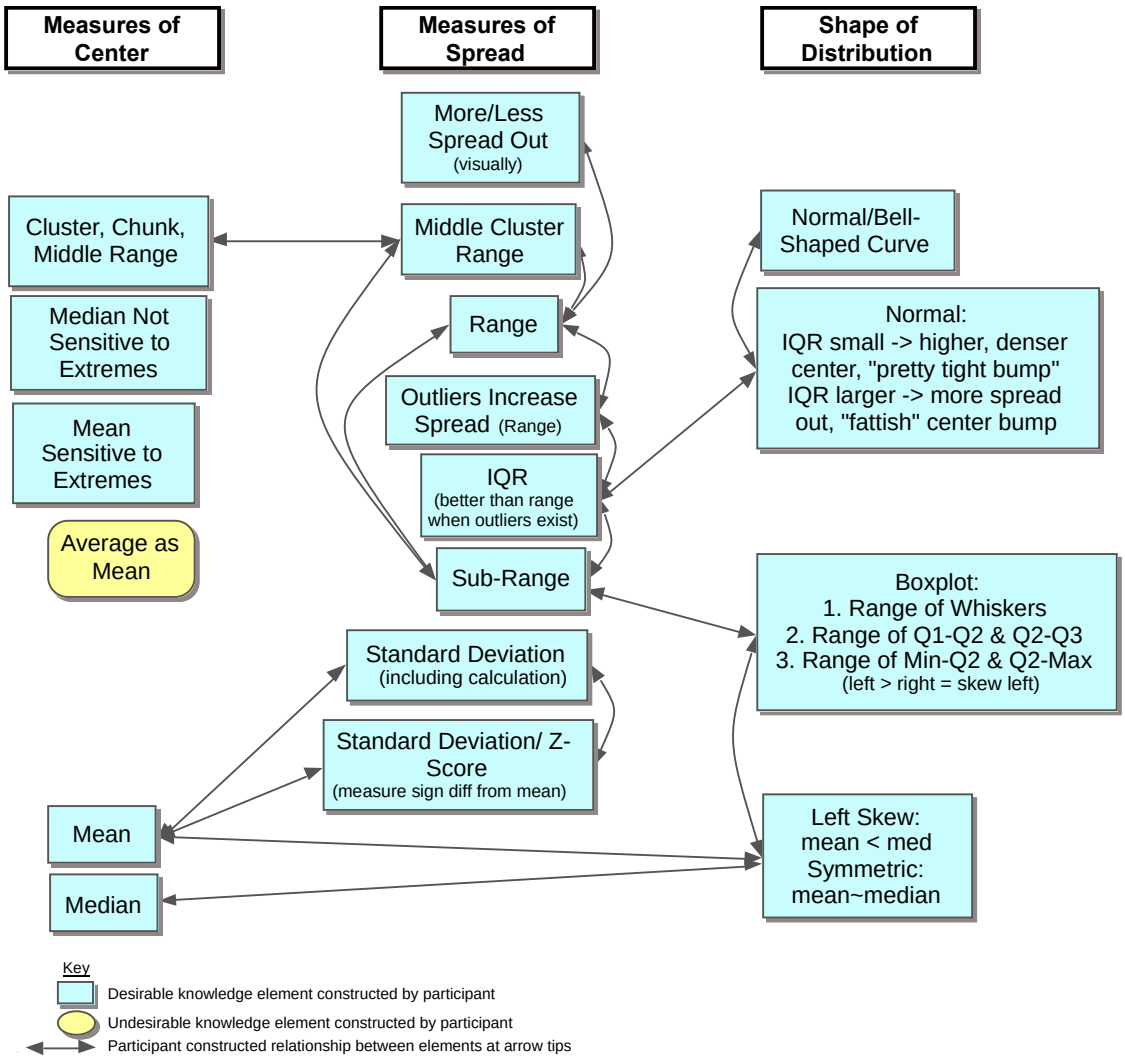
Harrison


Measures of Center
Measures of Spread
Shape of Distribution
Median
Mean
Mean as balancing point
Mean effected by extremes
Histogram: Mean can't be located at end of bin range
Average as Mean ONLY
Center as cluster
SD as Measure of Consistency
IQR ("length of the box"; calculation)
Spread as Sub Range
Range (visually, as end points, not single value)
Range to Compare Location
Symmetric: Mean ~ Med
a-symmetric: Mean /~ Med
One Outlier -> Skew
Box Plot: Compare Whiskers & Box Sections (symm: lengths about same a-symm: lengths different)
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

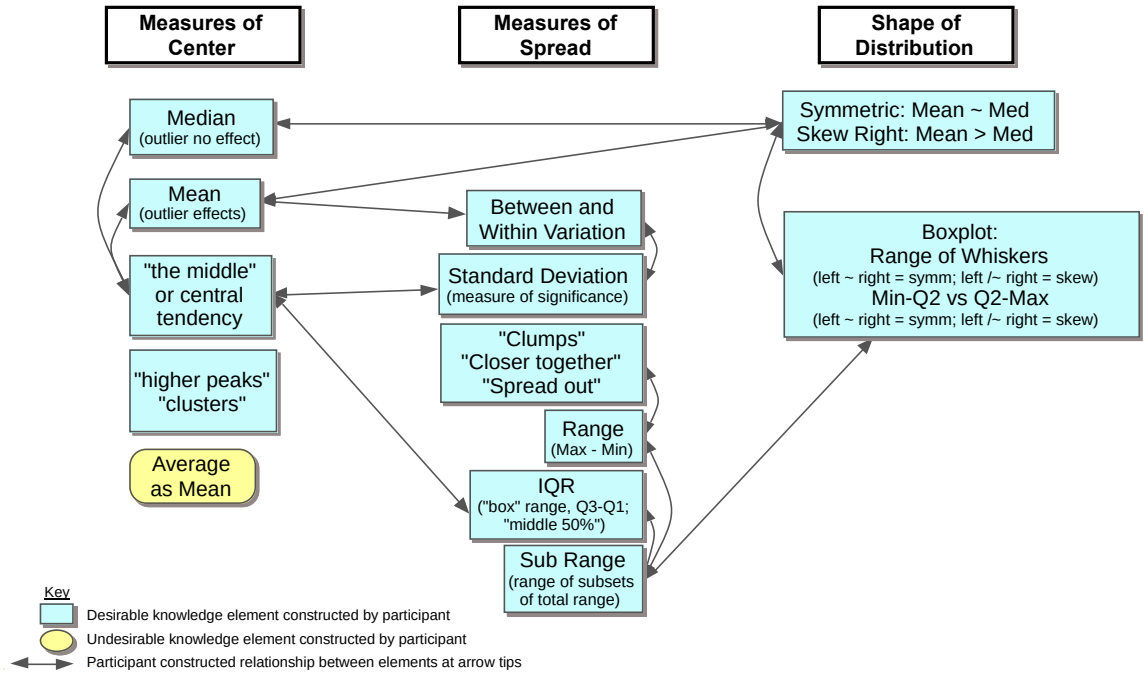
Michaela


Measures of Center
Measures of Spread
Shape of Distribution
Average as (only) Mean
Mean superior to median
Median not effected by outliers
Mean effected by outliers: large outlier results in larger mean
Median
Range
Sub Range
Standard Deviation (deviation from mean)
Standard Deviation to Measure Significance
Box Plot: Density of sub-range by comparing quartiles
Range of Whiskers (left ~ right = symm; left > right = skew left)
Skew Left: mean pulled LEFT
Skew Right: mean pulled LEFT
Skew Left: mean pulled RIGHT
Normal or bell curved: mean/med similar
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

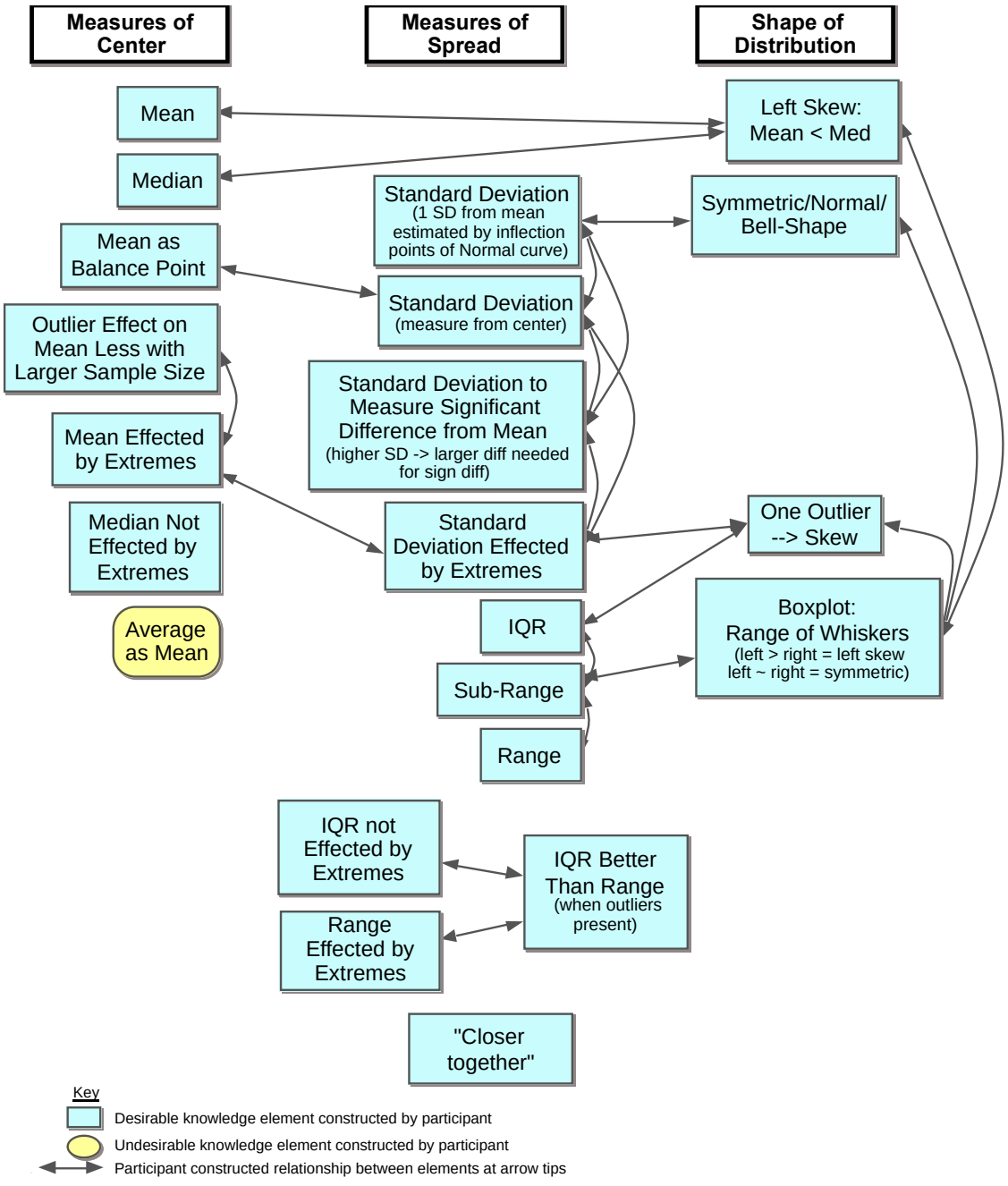
Mike


Measures of Center
Measures of Spread
Shape of Distribution
More/Less Spread Out (visually)
Cluster, Chunk, Middle Range
Middle Cluster Range
Normal/Bell-Shaped Curve
Median Not Sensitive to Extremes
Range
Normal: IQR small -> higher, denser center, "pretty tight bump" IQR larger -> more spread out, "fattish" center bump
Mean Sensitive to Extremes
Outliers Increase Spread (Range)
IQR (better than range when outliers exist)
Average as Mean
Sub-Range
Boxplot: 1. Range of Whiskers 2. Range of Q1-Q2 & Q2-Q3 3. Range of Min-Q2 & Q2-Max (left > right = skew left)
Standard Deviation (including calculation)
Standard Deviation/ Z-Score (measure sign diff from mean)
Mean
Median
Left Skew: mean < med Symmetric: mean~median
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

Rosalynn


Measures of Center
Measures of Spread
Shape of Distribution
Median
(outlier no effect)
Mean
(outlier effects)
"the middle" or central tendency
"higher peaks" "clusters"
Average as Mean
Between and Within Variation
Standard Deviation
(measure of significance)
"Clumps" "Closer together" "Spread out"
Range
(Max - Min)
IQR
("box" range, Q3-Q1; "middle 50%")
Sub Range
(range of subsets of total range)
Symmetric: Mean ~ Med
Skew Right: Mean > Med
Boxplot:
Range of Whiskers
(left ~ right = symm; left /~ right = skew)
Min-Q2 vs Q2-Max
(left ~ right = symm; left /~ right = skew)
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

Tim


Measures of Center
Measures of Spread
Shape of Distribution
Mean
Median
Mean as Balance Point
Outlier Effect on Mean Less with Larger Sample Size
Mean Effected by Extremes
Median Not Effected by Extremes
Average as Mean
Standard Deviation (1 SD from mean estimated by inflection points of Normal curve)
Standard Deviation (measure from center)
Standard Deviation to Measure Significant Difference from Mean (higher SD -> larger diff needed for sign diff)
Standard Deviation Effected by Extremes
IQR
Sub-Range
Range
IQR not Effected by Extremes
Range Effected by Extremes
IQR Better Than Range (when outliers present)
"Closer together"
Left Skew: Mean < Med
Symmetric/Normal/ Bell-Shape
One Outlier --> Skew
Boxplot: Range of Whiskers (left > right = left skew left ~ right = symmetric)
Key
Desirable knowledge element constructed by participant
Undesirable knowledge element constructed by participant
Participant constructed relationship between elements at arrow tips

VITA

Christopher Engledowl was born on January 27, 1985 in Springdale, Arkansas, the son of Brian and Barbara Engledowl. He completed his K-12 education in Springdale and graduated from Springdale High School in 2003. He then completed a Bachelor of Arts in Mathematics and a Master of Arts in Teaching Secondary Education–Mathematics degree from the University of Arkansas in 2009 and 2010. He then taught various algebra courses and AP Statistics at Fayetteville High School from 2010–2013, before entering the mathematics education PhD program at the University of Missouri.

During his PhD program at the University of Missouri, he worked as a Graduate Research Assistant on the Understanding and Implementing the CCSSM Practices project with Dr. Samuel Otten, the Centers for Learning and Teaching: Research to Identify Changes in Mathematics Education Doctoral Preparation and the Production of New Doctorates project with Dr. Robert E. Reys, and the Studying Teacher Expertise & Assignment in Mathematics (STEAM) project with Drs. Corey Webel, James E. Tarr, and Barbara Reys. He also was the instructor of record for LTC 4581—Teaching Mathematics in Secondary Schools: Algebra and LTC 4370—Teaching and Modeling Middle School Mathematics.

Christopher Engledowl is married to Gretchen Engledowl and has one daughter, Leela Engledowl. He has accepted an Assistant Professor position at New Mexico State University in the Curriculum & Instruction department in the College of Education, effective August 2017.