# ASSESSMENT OF COGNITIVE TRANSFER OUTCOMES FOR STUDENTS OF INTRODUCTORY STATISTICS

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Matthew Donald Beckman

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Joan Garfield, Advisor
Robert delMas, Co-Advisor

October 2015

**Acknowledgements**

Everything that matters in my life can be traced to God's amazing grace, my sweet wife Sarah, and our children Eden and Jack. I am deeply grateful to Sarah for her constant support, encouragement, and inspiration on so many days and nights when I am away working or studying. Although only one of us will be named on the diploma, her hard work, sacrifice, and perseverance have contributed every bit as much to this accomplishment as mine. I can see no possible way that I would have or could have completed this challenge without her.

I am thankful to my dad, mom, sister, brother, and grandparents for setting such fantastic examples of lives lived well, and for teaching me from a young age to set big goals and work hard toward them. I also thank my employers, Medtronic, for many years of tuition support and professional development. Especially Tom Keenan, Pat Zimmerman, Michael Soma, Jared Hanson, Nancy Figueroa, Michelle Nelson, and Daryle Peterson who have shown endless patience to help me balance my work schedule with the demands of classes, teaching, and research. I am extremely grateful to students and faculty in the Statistics Education cohort at the University of Minnesota and to Laura Le, Sandy Weisberg, Roxy Peck, Sashank Varma, Tim Jacobbe, Beth Chance and Marsha Lovett for thoughtful feedback supporting my dissertation research. I thank Don Richards and Michelle Everson for taking an interest in a student they didn't know and completely changing the trajectory of his life. Lastly, I thank my advisors, Joan Garfield and Bob delMas, for much guidance and encouragement, and most of all for taking a chance on a part-time student who just loves teaching statistics.

## Abstract

This study chronicles the creation of an assessment tool that quantifies cognitive transfer outcomes for introductory statistics students. Literature suggested that outcomes associated with cognitive transfer are closely aligned with statistical thinking and are indicative of students' ability to apply learning to novel scenarios beyond the classroom. No assessment tool had been developed and published for the purpose of measuring cognitive transfer outcomes among statistics students. The results of this study suggest that the Introductory Statistics Understanding and Discernment Outcomes (I-STUDIO) assessment tool may effectively serve this purpose.

The assessment tool was developed according to a rigorous protocol of expert feedback and iterative piloting. Data were collected and analyzed from a nationwide sample of nearly 2,000 students attending a wide variety of post-secondary institutions, and the I-STUDIO instrument was found to measure both forward-reaching and backward-reaching high road transfer outcomes with good psychometric properties.

Data analysis indicated high reliability and diverse validity evidence. This evidence included confirmatory factor analysis models with compelling alignment to the theoretical model and analysis of qualitative themes among expert feedback. Analysis of scoring consistency also showed strong inter-rater agreement. Although the sample size of the scored responses is somewhat small by convention for item response theory, a graded response model generally showed good item functioning. Furthermore, the data suggested that the I-STUDIO assessment estimated student ability with consistent precision across a wide range of above-average and below-average students.

Teachers and researchers can use I-STUDIO for comparing outcomes of alternative curricula. Additionally, the I-STUDIO instrument can be used to measure the effect of curriculum changes designed to improve transfer outcomes. Furthermore, the instrument and scoring rubric were designed to accommodate diverse curricula for the purpose of refining course outcomes.

## Table of Contents

## List of Tables

## List of Figures

# 1 Introduction

## 1.1 Rationale for the Study

Statistical thinking has been described in part to concern comprehension of "how, when, and why" a statistical framework can inform some inquiry (Ben-Zvi & Garfield, 2005). In learning and cognition research, an important mechanism by which students accomplish this sort of comprehension is sometimes referred to as cognitive transfer—or simply transfer. Singley and Anderson (1989) defined transfer to concern "how knowledge acquired in one situation applies (or fails to apply) in other situations." Similarly, Perkins and Salomon (1988) described transfer as "knowledge or skill associated with one context reach[ing] out to enhance another." Additionally, researchers noted a number of specific types of transfer including vertical transfer, near transfer, far transfer, and negative transfer (Bransford, Brown, & Cocking, 2000; Perkins & Salomon, 1988; Singley & Anderson, 1989).

Regardless of the distance or direction of transfer intended, successful outcomes require intentional effort (Bransford et al., 2000; Perkins & Salomon, 1988; Singley & Anderson, 1989). A ubiquitous theme among transfer researchers is perhaps stated most succinctly by Perkins and Salomon (1988, p. 22) that "transfer does not take care of itself." In fact, students without explicit intervention will struggle or fail to transfer even when problem sets are extremely similar (Butterfield & Nelson, 1991; Cooper & Sweller, 1987; Lovett & Greenhouse, 2000; Reed, Dempster, & Ettinger, 1985; Singley & Anderson, 1989; E. L. Thorndike & Woodworth, 1901a). Similarly, Garfield (2002) explained that statistics instructors often lay the groundwork of concepts and procedures

and expect students to develop statistical reasoning or thinking through opportunities to apply content with software and data sets, but it seems this is simply not enough. Without further coaxing, most students do not abstract and generalize content effectively enough to achieve the cognitive plasticity required to assimilate novel or advanced applications (Garfield, delMas, & Zieffler, 2012; Lovett & Greenhouse, 2000).

The challenge of assessing propensity to apply learned knowledge to a novel task is essentially rooted in the problem of measuring the magnitude of abstraction or generalizability achieved by a learner. Because there are no externally defined boundaries, the researcher is faced with difficult choices about appropriate target domain(s) and transfer distance. Moreover, propensity for transfer may vary by topic within a discipline such that a student may successfully accomplish a transfer task related to correlation but not comparison of group means (Budé, 2006). Several researchers have discussed approaches to assess propensity to transfer knowledge, although no published assessment currently exists designed to measure cognitive transfer outcomes for students of introductory statistics.

## 1.2 Problem Statement

Based on the literature reviewed, much needs to be done to promote and assess successful cognitive transfer outcomes for students of introductory statistics. However, no published assessment existed to measure this specific outcome, and the literature indicates uncertainty about whether cognitive transfer outcomes can be achieved and measured following an introductory statistics curriculum. The goal of this dissertation

was to develop an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students.

### 1.3 I-STUDIO Assessment Tool

If students are to benefit from their statistical training beyond the classroom in any context, the most basic requirements must be to identify relevant applications and demonstrate enough aptitude to begin working in the direction of a sensible solution. In this dissertation, a new assessment tool called the Introductory Statistics Understanding and Discernment Outcomes (I-STUDIO) instrument is introduced. The I-STUDIO assessment tool was designed to quantify evidence of transfer outcomes for use with diverse approaches to the introductory statistics curriculum.

The primary conceptual models anticipated to align with the I-STUDIO assessment tool would be dominated by the three major latent variable dimensions represented by Discernment, Forward-Reaching Transfer, and Backward-Reaching Transfer. A plausible configuration of the conceptual model appears in Figure 1. The dashed connectors associate latent variables that are correlated but not directly measureable. Solid connectors indicate manifest variables that are directly measureable by the I-STUDIO instrument, and are used to draw inference about associated latent variables.

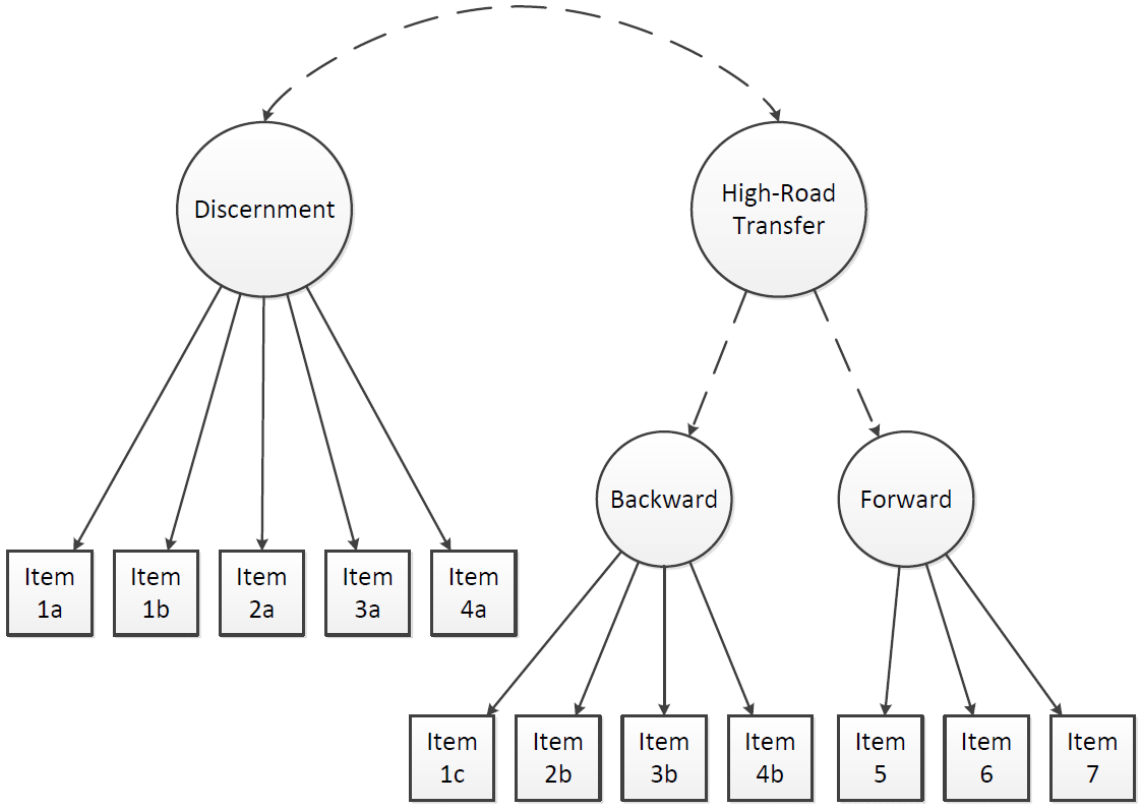


*Figure 1*. Conceptual model of I-STUDIO outcomes.

### 1.4 Structure of the Dissertation

Chapter 2 provides a review of the cognitive transfer literature including strategies that promote successful transfer in the classroom, and research studying the assessment of these outcomes. Chapter 3 explains the process of developing, refining, and implementing the I-STUDIO test blueprint and assessment tool. The chapter also describes the data collection and mixed methods approach to data analysis. These analyses estimate reliability, establish validity, and evaluate item response patterns using both quantitative and qualitative data.

In chapter 4, the results of the data analyses are reported. This includes expert feedback regarding the contribution of the I-STUDIO instrument in addition to scrutiny

of the test blueprint and draft assessment tool. Qualitative data were also summarized from student interviews during a pilot study prior to larger-scale field testing. Results of the field test were then reported including descriptive statistics, reliability analysis, additional validity evidence and preliminary analysis of item response modeling.

Chapter 5 synthesizes the results and offers interpretation of the findings. The chapter also discusses reliability, validity, and item analysis based on the field test data. Study limitations are presented and followed by implications for teaching and research. The chapter closes with a concise conclusion to the study. Following chapter 5 are appendices including raw materials and supporting documents. These include copies of the test blueprint, I-STUDIO instrument, and scoring rubric at various stages of their development.

# 2 Literature Review

## 2.1 Introduction to Literature Review

Many students only explicitly study statistics in one course during their academic career, and consequently, the vast majority of their use of statistics will be subject to what has been gleaned from that exposure (Giesbrecht, Sell, Scialfa, Sandals, & Ehlers, 1997). As such, the statistics education community carries a responsibility to set forth the most productive and effective introductory curriculum possible. However, Ben-Zvi and Garfield (2005) argued that "traditional approaches to teaching statistics focus on skills, procedures, and computations, which do not lead students to reason or think statistically… despite good performance in statistics courses."

Development of foundations for statistical thinking ought to be a key outcome of the introductory statistics curriculum (delMas, 2002; Shaughnessy, 2007). To this end, it is essential that the curriculum be optimized to impress the key ideals of statistical inference and probabilistic reasoning while preserving a flexibility that allows students to effectively apply these principles beyond the classroom in applications they encounter as students, professionals, citizens, and most any other domain in which information is aggregated and evaluated (Garfield et al., 2012). One theme that has emerged among researchers attempting to rethink the content of the introductory statistics curriculum is united by a reduction of emphasis on the traditional battery of procedures rooted in Normal distribution theory procedures by supplementing or replacing them with simulation-based methods (Ernst, 2004; Garfield et al., 2012; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011; Wild, Pfannkuch, & Regan, 2011).

This review summarizes research pertinent to cognitive transfer outcomes for introductory statistics students. In order to achieve this end, the review includes investigation of educational and psychological theories of cognitive transfer, including perspective related to instruction and assessment, especially as relating to introductory statistics education. Review of attributes that may differentiate the ability of a traditional or simulation-based introductory curricula to promote cognitive transfer are included.

### 2.1.1 Consensus of traditional approach based on Normal distribution theory.

Traditionally, a battery of procedures based on asymptotic approximations and Normal Distribution theory have been the tools of choice to serve these goals, and according to Scheaffer (1997) and D. S. Moore (2007), there has never been greater consensus on the content of the introductory statistics curriculum. However, instructors for such curricula may underestimate the complexity of included content and overestimate the capacity of students to succeed (Garfield et al., 2012; Wild et al., 2011). The result is an erosion of comprehension as students are expected to process and utilize too many concepts for them to manage (Wild et al., 2011). As a consequence, students are unable to retain what they have learned, much less transfer their knowledge to new applications beyond the classroom (Garfield et al., 2012).

> Using a metaphor introduced by Shoenfeld (1998), these types of statistics courses are teaching students how to follow ‘recipes’, but not how to really ‘cook’. That is, even if students leave these classes able to perform routine procedures and tests, they do not have the big picture of the statistical process that will allow them to solve

unfamiliar problems and to articulate and apply their understanding (Garfield et al., 2012, section 2.4, para. 1)

Even with the multitude of "recipes" to which students are exposed, other procedures with wide appeal and broad application such as ANOVA and basic nonparametric methods are frequently crowded out of the first course (Efron, 2000; Giesbrecht et al., 1997). The effect of such omissions may even impact the quality of academic literature in certain disciplines. For example, one study evaluating the use of nonparametric methods in an area of business research reviewed 1,102 papers from a group of 5 peer-reviewed academic journals related to organizational behavior and found 169 instances of nonparametric procedures out of 1,824 statistical applications (~9.3%) (Gaither & Glorfeld, 1985). In the cases where parametric procedures were used, the study authors found that most of the literature examined either omitted discussion of the assumptions underlying these methods or neglected to evaluate them (Gaither & Glorfeld, 1985). Although the study's scope is quite limited, it serves as a compelling illustration that some future researchers may conclude their statistics training either with little regard for evaluating parametric assumptions, or ill-equipped to draw inference when assumptions are not warranted (Gaither & Glorfeld, 1985).

### 2.1.2 Summary of efforts to retool the introductory curriculum.

Many reform efforts have been attempted, though often without consideration of alternative content or departure from the emphasis placed on computational mechanics (Ben-Zvi & Garfield, 2005). Teachers have attempted to promote software use so students can perform routine analyses more quickly, but absent other improvements,

students still may not truly achieve conceptual understanding (Brogan & Kutner, 1986; Mills, 2002). Another approach proposed to re-sequence introductory content to revisit material throughout the course in order to expand schema and broaden context (Malone, Gabrosek, Curtiss, & Race, 2010). The curriculum Malone et al. (2010) suggested was intended to shift away from a somewhat disjoint sequential model common among traditional introductory curricula toward a cyclical approach of visiting and revisiting topics in effort to more closely approximate the work of practicing scientists and statisticians. However, even if skillful integration of software tools and optimal sequencing were achieved within the consensus curriculum, Cobb (2007) and others argue that true progress requires critical consideration of the very content of the introductory curriculum. Indeed, for the introductory statistics curriculum to be truly reformed such that students actually understand how to apply statistical knowledge to new applications, the statistics education community must first study the very nature of cognitive transfer and evaluate how best to develop and assess introductory curricula in terms of this goal.

## 2.2 Cognitive Transfer Literature

In the emergent years of the discipline, psychologists gave considerable attention to the topic of cognitive transfer (Cox, 1997). The appeal of cognitive transfer among early and modern psychologists is associated with the age-old challenge of producing novel responses based on prior experiences, especially when there has not been (or cannot be) explicit training in the context of the new task (Cox, 1997; Singley & Anderson, 1989). This section defines cognitive transfer and related topics, and then outlines a brief history

of transfer research succeeded by discussion of modern research topics related to cognitive transfer including schema development, metacognition, cognitive load, as well as views pertaining to the assessment of transfer outcomes.

### 2.2.1 Definitions.

Several terms should first be defined in order to properly review a literature related to cognitive transfer, especially with the goal of connecting cognitive transfer to a concrete application such as outcomes of a particular curriculum. Although transfer is perhaps only one of many educational outcomes of value, it is worth noting that a number of other favorable outcomes bear remarkable overlap with cognitive transfer. Furthermore, cognitive transfer itself has been defined to take many forms. Also, it is pertinent that appropriate connections are made to the concept of statistical thinking in order to accomplish the larger purpose of this review.

#### *2.2.1.1 Topics analogous to cognitive transfer.*

A number of salient educational outcomes intersect with cognitive transfer, including development of expertise, synthesis of learning, integrated understanding, analogical reasoning, and statistical reasoning. Sternberg (1998) described the components of expertise to include broad, robust schema pertinent to the domain, as well as a well-refined ability to determine appropriate problem solving strategy and accurately characterize the difficulty of such tasks. Moreover, Sternberg (1998) noted that the expert is able to group routine operations and process them in an automated fashion with little need for controlled contemplation of the constituent tasks. Synthesis is often discussed in terms of achieving portability and flexibility with the abstract cognitive elements within a

schema network such that they can be organized and reorganized to accommodate novel extensions to which the existing schema can be applied (Bloom, 1956; Lovett & Greenhouse, 2000).

Analogical reasoning has also been described as a process closely linked to transfer in that both are predicated on linking understanding of a source domain to some appropriate parallel in a target domain (Alexander, Murphy, & Kulikowich, 1998; Alexander & Murphy, 1999; Gick & Holyoak, 1980). Garfield (2002, Summary section para. 2) discussed "integrated understanding" as a prerequisite for statistical reasoning, and in turn described statistical reasoning as the ability of an individual to recognize a statistical issue and assimilate it into the relevant schema in order to discern an appropriate strategy for solving, and evaluating a result in the original context of the task. There is much interactivity and even co-dependency among the concepts of transfer, expertise, synthesis, and understanding in the classroom, such that these terms are at times used and studied interchangeably in the literature. Although this review aimed to explore transfer outcomes as they relate to the introductory statistics curriculum, it is appropriate to draw on ideas related to some of these other concepts so far as they can be deemed applicable.

#### *2.2.1.2 Types of cognitive transfer.*

Singley and Anderson (1989) defined transfer to concern "how knowledge acquired in one situation applies (or fails to apply) in other situations." Similarly, Perkins and Salomon (1988) described transfer as "knowledge or skill associated with one context reach[ing] out to enhance another." Additionally, researchers noted a number of specific types of transfer including vertical transfer, near transfer, far transfer, and negative

transfer (Bransford et al., 2000; Perkins & Salomon, 1988; Singley & Anderson, 1989). Vertical transfer occurs when an application builds directly upon experience with a pre-requisite subset of the application (Bransford et al., 2000). For example, skills with the order of operations in mathematics are useful when learning to manipulate algebraic expressions. It is also possible that experience from one context actually interferes in another (Bransford et al., 2000; Perkins & Salomon, 1988; Singley & Anderson, 1989; E. L. Thorndike & Woodworth, 1901a). This phenomenon is commonly known as negative transfer in the literature. Near transfer takes place among applications that are highly similar, while far transfer is relevant to applications that differ among superficial attributes yet build upon common concepts (Alexander & Murphy, 1999; Bransford et al., 2000; Cox, 1997; Paas, 1992).

The different types of transfer described are generally discussed consistently in the literature, however, there is some room for interpretation regarding how “near” is near transfer or how “far” is far transfer. For example, Paas (1992) discusses an experiment conducted to evaluate factors influencing near and far transfer among classroom exercises, while Bransford et al. (2000) tended to reserve far transfer to bridge the gap from the educational context to settings outside of school. So, perhaps not surprisingly, “near” and “far” are relative terms in the transfer literature and, consequently, it is important that the reader acknowledge the author’s definition of each.

Salomon and Perkins (1989) described different means to produce near and far transfer outcomes through processes that the authors deemed “low road” and “high road” transfer. In essence, low road transfer relies on automaticity produced by repetition, while

high road transfer requires a deliberate appeal to abstract cognitive elements previously mastered (Cox, 1997; Salomon & Perkins, 1989). The nature of deliberation when promoting high-road transfer further partitions outcomes based on whether abstraction is conducted in order to generalize cognitive elements for an undetermined future use—forward-reaching transfer—or whether abstraction consists of an intentional search of available schema for relevant cognitive elements that may be applied to a task at hand—backward-reaching transfer (Salomon & Perkins, 1989).

#### *2.2.1.3 Statistical thinking and cognitive transfer.*

In the statistics education literature, the concept of "statistical thinking" is a relatively recent phenomenon similar to cognitive transfer as it has been described above (Pfannkuch & Wild, 2005). In short, statistical thinking has been described in part to concern comprehension of "how, when, and why" a statistical framework can inform some inquiry (Ben-Zvi & Garfield, 2005). Similarly, Wild and Pfannkuch (1999, p. 224) defined the core of statistical thinking as "complex thought processes involved in solving real-world problems using statistics with a view to improving such problem solving." delMas (2006) further grounded the idea by relating statistical thinking to elements described among the highest three levels of Bloom's (1956) hierarchical taxonomy of cognitive outcomes—analysis, synthesis, and evaluation. As mentioned previously, Garfield et al. (2012) describe the ability to think statistically as akin to the ability to "cook" rather than simply following "recipes", resulting in deep, agile understanding that can readily accommodate the nuances of unfamiliar applications. Each of these perspectives lend themselves to the idea that statistical thinking relates to an insight when

and how to inform one's approach to a real-world problem using appropriate statistical principles, which certainly agrees with the descriptions of cognitive transfer discussed above. However, some have cautioned that the phrase "statistical thinking" has been used too loosely in some academic literature, and therefore proposed abandoning use of the term and retreating to cognitive transfer as a more accurate characterization of the highest level of understanding for statistical inquiry (Budé, 2006).

Wild and Pfannkuch (1999; 2005) described five principle elements of statistical thinking including: acknowledgment of a need to collect data; conversion of raw data to meaningful graphical and numerical summaries (i.e., "transnumeration"); consideration of variability; construction of statistical representations (i.e., "models"); incorporation of statistical and contextual knowledge. As with cognitive transfer, students are unlikely to develop statistical thinking, or its fundamental elements, without explicit effort and targeted motivation (Pfannkuch & Wild, 2005). As quoted by Pfannkuch and Wild (2005), Gal, Ahlgren, Burrill, Landwehr, Rich, and Begg (1995, p. 25) summarized that students do not fully develop statistical thinking through simple participation in statistical inquiry, it is "both an issue of skill transfer, as well as the fact that a somewhat different set of cognitive skills and dispositions is called for." Moreover, students must be groomed to embrace imagination, skepticism, and consideration of problems from multiple perspectives in order to truly engage statistical thinking (Wild & Pfannkuch, 1999). If this is true, it would imply that the prevailing strategy of engaging students in project work in order to encourage statistical thinking, while arguably necessary, is perhaps not sufficient (Wild & Pfannkuch, 1999).

#### *2.2.1.4 Failure to transfer.*

Regardless of the distance or direction of transfer intended, successful outcomes require intentional effort (Bransford et al., 2000; Perkins & Salomon, 1988; Singley & Anderson, 1989). A ubiquitous theme among transfer researchers is perhaps stated most succinctly by Perkins and Salomon (1988, p. 22) that "transfer does not take care of itself." In fact, students without explicit intervention will struggle or fail to transfer even when problem sets are extremely similar (Butterfield & Nelson, 1991; Cooper & Sweller, 1987; Lovett & Greenhouse, 2000; Reed et al., 1985; Singley & Anderson, 1989; E. L. Thorndike & Woodworth, 1901a). The phenomenon of failed transfer applies to in-class tasks that differ only in appearance from previously mastered content as well as out-of-class scenarios for which students may not even think to apply their learning (Butterfield & Nelson, 1991; Lovett & Greenhouse, 2000; Salomon & Perkins, 1989). Bransford et al. (2000) discussed a compelling case study in which students achieved great mastery memorizing very long digit strings after much practice, but when the digits were replaced by letters the students were only capable of recalling the first few. Such examples are representative of the phenomenon of "skill specificity" which is operationally equivalent to failed transfer (Ackerman, 1990).

Similarly, Garfield (2002) explained that statistics instructors often lay the groundwork of concepts and procedures and expect students to develop statistical reasoning or thinking through opportunities to apply content with software and data sets, but it seems this is simply not enough. Unfortunately, students often resort to rote learning of statistical methods in order to get past required exams because the content

does not seem relevant to them, but rote learning of this nature is unlikely to produce integrated understanding (Broers, Mur, & Bude, 2004). Without further coaxing, most students simply do not abstract and generalize content effectively enough to achieve the cognitive plasticity required to assimilate novel or advanced applications (Garfield et al., 2012; Lovett & Greenhouse, 2000). Perkins and Salomon (1988) added that an important part of developing successful transfer outcomes requires not to simply introduce the component skills and concepts and hope for the right outcome, but to train students to be intentional about learning with transfer in mind. The desired integrated understanding cannot be received passively from an instructor; it must be constructed by the learners for themselves (Broers et al., 2004).

Salomon and Perkins (1989) explained that the phenomenon of failure to transfer observed in the context of academic experimentation should not be entirely surprising. They described that low road transfer outcomes in particular may be indiscernible via short-term experimentation since the propensity to produce these outcomes typically evolves gradually over a long period of time, often on the order of several years (Salomon & Perkins, 1989). If this is true, it would in part explain the remarkably divergent conclusions about transfer observed throughout the academic literature (Butterfield & Nelson, 1991), which includes everything from algorithmic approaches believed to practically ensure production of successful transfer (Salomon & Perkins, 1989) to denial that cognitive transfer even exists apart from a byproduct to general intelligence (Detterman, 1993).

### 2.2.2 Foundations of cognitive transfer research.

The prevailing doctrine at the turn of the 20th century believed that the mind could be developed broadly by exercising it, like a muscle, with esoteric subjects like Latin, Geometry and Chess—a theory with roots dating to Aristotle (Singley & Anderson, 1989). This so-called "doctrine of formal discipline" was built upon the premise that transfer occurs in very general terms, potentially among contexts with no commonalities (Cox, 1997; Singley & Anderson, 1989). The theory was actively supported by mainstream psychologists of the day including Alfred Binet (1899) who integrated these ideas into his work on intelligence testing (Singley & Anderson, 1989). The doctrine of formal discipline went largely unchallenged until a series of papers published by E. L. Thorndike and Woodworth (1901a; 1901b; 1901c) introduced a competing philosophy which became known as the "theory of identical elements" (Singley & Anderson, 1989). In contrast to the doctrine of formal discipline, E. L. Thorndike and Woodworth had conducted experiments that to show that success in cognitive transfer outcomes was a function of identical elements between learned tasks and novel tasks rather than a result influenced by exercising general faculties (Bransford et al., 2000; Cox, 1997).

Since the emergent years of psychology as a discipline, a number of approaches to explaining transfer outcomes have been developed that still bear remarkable similarity to the theory of identical elements and doctrine of formal disciplines (Cox, 1997). Researchers experimented with a Behaviorist emphasis on similarity between stimulus and response to produce transfer outcomes, but such emphasis rarely produces meaningful far transfer outcomes (Cox, 1997). Others have proposed modifications of the common elements theory, like the ACT* theory that rests upon commonality among

abstract procedural elements (Singley & Anderson, 1989). Furthermore, Gestalt ideas fuelled interest in teaching metacognitive processes a viable approach to facilitating transfer (Cox, 1997; Helfenstein, 2005).

Transfer outcomes tend to decline as the distance of transfer increases, but if the alternative to transfer is to teach content in exactly the contexts in which it should be applied, it would effectively amount to apprenticeship (Cox, 1997). While apprenticeship has its place, and is still prevalent in many forms—including the relationship between a graduate student and his advisor(s)—it is fair to say that Western education through at least high school is based almost entirely on the expectation of successful transfer (Cox, 1997).

### 2.2.3 Development of schema and cognitive elements.

Abstraction of cognitive elements (i.e., knowledge, skills, and combinations thereof) seems to have a compelling role in preparing students for successful transfer outcomes, but essential to this end is the development of a rich schema for the content area (Helfenstein, 2005; Lovett & Greenhouse, 2000; Perkins & Salomon, 1988). Mastery of a rich schema must precede successful transfer outcomes (Bransford et al., 2000; Cooper & Sweller, 1987; Paas, 1992), yet building and mastering schema is no small task. Schema tends to start small with a few similar problems, and then grows organically as the elements of the schema are repeatedly accessed and strengthened (Cooper & Sweller, 1987; Lovett & Greenhouse, 2000). Helfenstein (2005) argued that the functional development and interconnection of schema and abstraction of cognitive elements are rooted in Gestalt ideas related to insight during the course of problem solving. However,

this should not imply that development and organization of such insight require lightning-strike experiences; schema development can also be nurtured through careful direction of learning activities (Rittle-Johnson, 2006; Salomon & Perkins, 1989). For example, when content is taught with exposure to multiple contexts, students are more likely to abstract relevant features of the subject matter to more readily draw upon them in the future (Bransford et al., 2000; Salomon & Perkins, 1989). In time, the boundaries of a schema domain swell and overlap with other domains such that the problem solver becomes increasingly equipped to assimilate a new problem into existing schema because of a depth and breadth of associated content mastery (Cooper & Sweller, 1987). In fact, some believe that higher ability students distinguish themselves by virtue of achieving greater abstraction and the facility to call on more distant connections (Goska & Ackerman, 1996).

Bransford et al. (2000) assert that all learning requires transfer based on prior learning. For better or worse, students arrive with an existing network of schema that cannot be overlooked (Broers et al., 2004; Garfield, 1995; Lovett & Greenhouse, 2000). In fact, some have attributed the success of cognitive transfer to the degree of overlap between associated cognitive elements, whether concrete or abstracted (Bransford et al., 2000; Goska & Ackerman, 1996; Helfenstein, 2005; Singley & Anderson, 1989; Sternberg, 1998; E. L. Thorndike & Woodworth, 1901a). Probability topics, for example, notoriously suffer from challenges rooted in poor intuition and contradictory understanding of relevant terms due to inconsistencies with their conversational use (Garfield, 1995; Lovett & Greenhouse, 2000; Singley & Anderson, 1989) resulting in

negative transfer outcomes. Students must reconcile disparities between principles of statistical reasoning and the many fallacies that permeate common culture outside the classroom (Garfield, 1995). Unfortunately, simply providing contradictions to these fallacies is insufficient to depose them and rebuild intuition (Bransford et al., 2000; Garfield, 1995).

While cultivation of a deep and diverse schema is necessary for successful transfer outcomes, it is not sufficient. It should be acknowledged that all learning occurs within some context, and successful transfer outcomes require sufficient abstraction of cognitive elements to transcend the context in which content has been learned (Lovett & Greenhouse, 2000; Perkins & Salomon, 1988). However, students have difficulty recognizing problem structure and generalizing cognitive elements on their own, so the educator must engage strategic methods to facilitate the desired abstraction (Reed et al., 1985). One strategy to encourage successful transfer outcomes involves creating opportunities for students to receive feedback to help them understand when the content is applicable and when it is not applicable; absent this instruction, students tend to use inappropriate mnemonics such as chapter and textbook location in place of appropriate integration to their greater knowledgebase (Bransford et al., 2000).

Further complicating things, schools tend to emphasize abstraction of subject matter discussed, while the scenarios in the real-world that will demand use of that subject matter will almost certainly require contextualized reasoning (Bransford et al., 2000). Lovett and Greenhouse (2000) discuss a strategy implemented using “synthesis labs” in an introductory statistics course in which students are challenged on a regular basis

throughout the semester with tasks that leverage material that integrates cumulative content from the course to date. This creates practice with concepts like tool selection, which comes naturally to an expert statistician but is difficult for the novice to exercise in earnest; the labs give students an opportunity to combine cognitive elements in different ways during problem solving in order to promote synthesis and transfer (Lovett & Greenhouse, 2000). As summarized by Wild and Pfannkuch:

> Statistics is itself a collection of abstract models ('models' is used in a very broad sense) which permit an efficient implementation of the use of archetypes as a method of problem solution. One abstracts pertinent elements of the problem context that map onto a relevant archetypical problem type, uses what has been worked out about solving such problems, and maps the answer back to context domain. There is a continual shuttling between the two domains and it is in this shuttling or interplay, that statistical thinking takes place. (1999, p. 244)

To be clear, the process of schema development and abstraction of cognitive elements should start with specific examples from which the student is responsible for abstracting understanding. Instruction that is too general may become too vague to be useful for any specific application (Singley & Anderson, 1989). It is important to distinguish that the information should be available in the abstract, but must be usable for a particular situation (Singley & Anderson, 1989). For example, anyone who knows the rules of chess can theoretically execute the perfect game by generating all possible moves and counter-moves and thereby choose the optimal strategy in every scenario; however, this is not a

realistic outcome for a player that has only learned the rules abstractly (Singley & Anderson, 1989).

### 2.2.4 Metacognition.

Metacognitive strategies are believed to play a critical role in achieving successful transfer outcomes (Helfenstein, 2005; Perkins & Salomon, 1988), though they do not come naturally to students without intervention from an instructor (Sternberg, 1998). Sternberg (1998) points out that students become accustomed to the usual paradigm of passive learning and some persistent coaxing is needed to encourage them to engage metacognition and contemplation of the material on a deeper level. Moreover, Sternberg and others posit that learning effective metacognitive strategies ought to be at least as important as the subject matter of the course (Atkinson, Catrambone, & Merrill, 2003; Bransford et al., 2000; Georghiades, 2000; Perkins & Salomon, 1988; Sternberg, 1998). Decontextualized strategies for metacognition such as self-monitoring and feedback have been repeatedly demonstrated effective over several decades of research, but still have not enjoyed widespread use in the classroom (Cox, 1997). However, others have suggested that metacognitive strategies may be rooted in context just as schema development tends to be rooted in context, (e.g., Sternberg, 1998).

Several metacognitive strategies recommended based on positive research outcomes include active grouping of tasks within a problem solving context in order to make explicit an architecture of sub-goals and concept maps pertinent to a content area (Atkinson et al., 2003; Broers et al., 2004; Schau & Mattern, 1997), study of worked examples or expert solutions (Cooper & Sweller, 1987; Lovett & Greenhouse, 2000;

Paas, 1992; Reed et al., 1985; Renkl, 2002; Rittle-Johnson, 2006), and self-explanation (Broers et al., 2004; Chi, De Leeuw, Chiu, & LaVancher, 1994; Wong, Lawson, & Keeves, 2002).

#### *2.2.4.1 Developing sub-goals and concept maps for problem solving.*

Grouping of sub-goals and use of concept maps encourage students to organize their approach to problem solving (Atkinson et al., 2003; Broers et al., 2004; Schau & Mattern, 1997). Use of sub-goals seems to be effective since it is natural for problems within a common content area to share a consistent set of sub-goals, even though the individual steps for accomplishing these sub-goals will vary from one problem to the next (Atkinson et al., 2003). Internalizing sub-goal architecture provides a framework for the student to assimilate novel problems into their existing schema based on shared abstract elements and tangible benchmarks that will carry the student toward a solution (Atkinson et al., 2003). Similarly, concept mapping can be useful for instructional planning in order to show students how concepts are interconnected or as a learning tool compelling students to explicitly organize schema relevant to the task at hand (Schau & Mattern, 1997). In short, both strategies improve students' ability to connect novel aspects of a task to an established knowledgebase (Broers et al., 2004; Schau & Mattern, 1997).

#### *2.2.4.2 Study of worked examples and self-explanation.*

Study of worked examples may benefit students for similar reasons in that it makes accessible the key components required in a problem domain, so students can monitor their progress with continuous internal feedback as they work through the task mentally (Lovett & Greenhouse, 2000; Reed et al., 1985). Furthermore, study of worked examples

has been shown to be a very efficient strategy for achieving near transfer outcomes when compared to other strategies like completion exercises (Cooper & Sweller, 1987; Lovett & Greenhouse, 2000). In fact, Rittle-Johnson (2006) characterized study of worked examples as a form of direct instruction as contrasted with unaided problem solving. The act of comparing and contrasting tasks with analogous structure and superficial differences as advocated by Lovett & Greenhouse (2000) seems to profit the student by requiring that they strive for abstraction of the key elements and structure of the problem while acknowledging which attributes may be dismissed (Perkins & Salomon, 1988).

Self-explanation, for the purpose of this discussion, should be understood to mean the process of a student explaining correct material (e.g., a textbook passage) to herself rather than explaining her own solutions or interpreting explanations provided by others either of which may be incorrect to begin with (Rittle-Johnson, 2006). This definition makes clear that self-explanation is closely related to study of worked examples. Asking students to self-explain creates the opportunity for them to reconcile their existing schema with the idiosyncrasies of the novel task at hand (Broers et al., 2004; Chi et al., 1994). Proponents of a constructivist approach to learning theory reject the notion that understanding can be simply conveyed to a student by a teacher; the student must actively integrate new knowledge into their existing framework (Broers et al., 2004). Moreover, contradictions can be exposed and addressed in “real-time” before they can take root and further undermine the learning process (Chi et al., 1994).

Chi et al. (Chi et al., 1994) conducted a study to evaluate the effect of self-explanation on interpretation of a text passage about the circulatory system. They found

that the students who had been directed to self-explain had a better grasp of the content than students instructed to simply re-read the passage for a roughly equivalent amount of time. The authors also noted that within the self-explanation group, the most successful students also tended to record the largest number of comments and sketches while self-explaining, and described a causal relationship between number of self-explanation comments and improved performance outcomes (Chi et al., 1994). Other researchers have cautioned attributing causality to such conclusions (e.g., Rittle-Johnson, 2006), but the association has been corroborated (e.g., Salomon & Perkins, 1989). The analysis conducted does not address the converse argument that perhaps students generally capable of higher performance could be pre-disposed to construct more capable explanations of new content during a self-explanation exercise. The authors had collected standardized test scores for students studied, so that information could perhaps have been useful to help reconcile this confounding.

Rittle-Johnson (2006) studied whether learning and transfer gains are impacted in a discovery learning context as opposed to a direct instruction environment. Furthermore, Rittle-Johnson (2006) evaluated results involving elapsed time between intervention and production of the transfer task. The study was designed using a computer program involving pre-algebra tasks based on the associative property of addition. The program was described as a game and provided to 85 elementary students randomly assigned into four treatment groups. The treatment groups were divided by learning approach (discovery learning or direct instruction) and whether or not students were prompted to self-explain (Rittle-Johnson, 2006).

Students completed a pretest prior to intervention, and two posttests—one immediately following intervention and another after a two week delay (Rittle-Johnson, 2006). The transfer tasks were constructed such that the format was unfamiliar to the students, but could be solved by adapting procedures already learned. Typical tasks were of the sort 8 + 3 + 2 = ___ + 3; transfer tasks placed the unknown on the left, introduced subtraction, or did not include duplicate addend (Rittle-Johnson, 2006).

Rittle-Johnson (2006) reported that students prompted to self-explain improved more than their peers, however time delay and learning approach did not produce statistically significant differences among treatment groups. Thus, the improvement on transfer tasks was sustained over the two-week delay and was not impacted by instructional approach (Rittle-Johnson, 2006). Rittle-Johnson (2006) asserted that self-explanation likely prompted more active cognitive processing, which led to favorable results among students implementing self-explanation.

### 2.2.5 Cognitive load.

Active cognitive processing is an important consideration for learning and transfer outcomes, but the volume of information that one is capable of actively processing is considered finite (Deary, 2001; Lovett & Greenhouse, 2000; Sweller, 1994). Cognitive psychologists have frequently drawn upon a metaphor of computer processing speed and capacity in order to characterize elements of cognitive function (Deary, 2001). In essence, a person is constrained by some limited amount of cognitive capacity; this cognitive load describes the amount of burden imposed on the individual to process simultaneous demands (Cooper & Sweller, 1987; Lovett & Greenhouse, 2000).

Research has suggested that there exists an inverse relationship between cognitive load and the efficiency of learning and transfer (Lovett & Greenhouse, 2000). Consequently, consideration for cognitive load is necessary in order to promote positive transfer outcomes that might otherwise be compromised in the interest of maximizing the volume of content covered during a finite course (Paas, 1992). Although cognitive load is often discussed as a holistic concept, Sweller (1994) articulates a useful partition referred to as intrinsic and extraneous cognitive load. Also, automation of cognitive processes is discussed as a means to promote successful transfer outcomes by mitigating cognitive burden.

#### *2.2.5.1 Intrinsic and extraneous cognitive load.*

Sweller (1994) described intrinsic cognitive load as the genuine burden attributable to the nature of the content. In general, instructors are believed to have little or no control over the intrinsic cognitive load to which students are exposed (Sweller, 1994). If cognitive elements can be learned in succession, the intrinsic cognitive load is reduced because the need for interactivity is removed (Sweller, 1994). However, it is often the case that cognitive elements must be developed simultaneously because the most important outcomes lie in their interaction (Sweller, 1994).

Some areas require greater element interactivity than others, and this fact is largely a fundamental truth of the content area and cannot be greatly influenced by instruction or environment (Singley & Anderson, 1989; Sweller, 1994). Statistics has been implicated as a high cognitive load domain (Paas, 1992). If intrinsic cognitive load is relatively constant for a content area, the instructor has a responsibility to mitigate extraneous

cognitive load in order to provide students greatest opportunity for efficient learning outcomes.

Extraneous cognitive load can be described as an artificial burden directly attributed to the instructional methods (Sweller, 1994). Suboptimal instruction imposes inefficient demand on cognitive processing resulting in disproportionate extraneous cognitive load that hinders the potential for successful transfer (Paas, 1992; Singley & Anderson, 1989; Sweller, 1994). This paradigm frequently presents as a tendency to cover too much content too quickly for students to adequately process, and the resulting dysfunction is impaired learning and transfer because students perceive the content as a set of disjointed facts and have not organized and assimilated these concepts into usable schema (Bransford et al., 2000). Wild et al. (2011) and Garfield (1995) have echoed this phenomenon in statistics education, and suggested that many instructors are likely out of touch with the scale on which this problem affects their students.

#### *2.2.5.2 Automation of cognitive processes.*

While it is important to monitor and eliminate sources of extraneous cognitive load for learning, this does not completely address the management of cognitive load during problem solving. Development of a rich schema network and automation of cognitive processes have frequently been proposed as effective strategies to significantly reduce the working memory requirement for a given task (Rittle-Johnson, 2006; Sweller, 1994). Furthermore, schema acquisition and rule automation are also thought to facilitate effective transfer to other contexts (Cooper & Sweller, 1987). Sweller (1994) discusses that larger schema serves to abstract and expand cognitive elements, which leads to a

chunking effect by which closely related elements of schema are utilized as one, increasing available working memory. Note that this chunking behavior relates closely to the metacognitive strategies discussed by Atkinson et al. (2003) who emphasized grouping problem solving operations into sub-goals. If cultivating a larger, well-developed schema enables chunking to squeeze more knowledge into a limited cognitive capacity, automation of cognitive processes leverages an opposite strategy by allowing the working memory to be circumvented almost entirely (Cooper & Sweller, 1987; Sternberg, 1998). Recall the role of automation in Salomon and Perkins' (1989) discussion of low road transfer. Automation may serve as an effective catalyst for transfer and greater problem solving efficiency on future tasks because cognitive load can then be allocated towards planning, strategy and synthesis (Cooper & Sweller, 1987). In fact, Cox (1997) summarized that even early psychologists including Edward Thorndike and William James viewed such automaticity as a means to release the discerning faculties of consciousness to attend to these higher-level decisions.

Sternberg (1998) claimed that to the extent that automated functioning has been achieved, too much metacognitive processing will begin to hinder functioning. However, Sweller (1994) pointed out that as tempting as it is to treat controlled and automatic cognitive processing as though it is all-or-nothing, the transition between them almost always occurs on a continuum. Interestingly, Cooper & Sweller (1987) claimed that study of worked examples presents one of several promising strategies for facilitating the switch to automation.

### 2.2.6 Strategies for the assessment of cognitive transfer.

When considering the task of assessing transfer outcomes, it is most common to consider some learned body of knowledge then evaluate the learner's ability or propensity to apply that body of knowledge to a novel task (Bransford et al., 2000; Budé, 2006). Another incarnation of successful transfer, however, could be manifest as an increased speed in learning a new content area (Bransford et al., 2000). Both are desirable outcomes with important implications for how students carry what they learn in one context and transfer that knowledge to new applications.

#### *2.2.6.1 Assessing increased speed in learning a new content area.*

Examples of the increased speed in learning new content may include a student who is more successful in physics (or perhaps statistics) due to prior knowledge of calculus (Bransford et al., 2000), or learning a new text editor after having previously learned a different text editing software (Singley & Anderson, 1989). The text editor experiment conducted by Singley and Anderson (1989) provides a nice example in kind. The study included 24 women from a secretarial school, all of whom were naïve to computers but competent typists. Participants were evaluated for typing speed and performance on a standardized spatial memory test, and then assigned experimental groups that were approximately balanced with respect to these characteristics.

Table 1

*Experimental Treatment Groups in Text Editor Experiment*

| Group | Days 1 and 2 | Days 3 and 4 | Days 5 and 6 |
|---|---|---|---|
| Treatment 1 | LTE 1 | LTE 1 | STE |
| Treatment 2 | LTE 1 | LTE 2 | STE |
| Treatment 3 | LTE 2 | LTE 1 | STE |
| Treatment 4 | LTE 2 | LTE 2 | STE |
| Typing Control | Typing | Typing | STE |

| Group | Days 1 and 2 | Days 3 and 4 | Days 5 and 6 |
|---|---|---|---|
| STE Control | STE | STE | STE |

The experiment lasted six days, and studied influence of learning a line text editing (LTE) program on transfer to a screen text editing (STE) program. The treatment groups described in Table 1 shows the allocation of subjects during the first four days to practice one or both LTE programs, the STE program only, or typing only; on the fifth and sixth days all subjects used the STE program. Two control groups were included, such that the first was simply typing the manuscript at a terminal and the other used the STE for all six days. The assessment strategy for this experiment included analysis of transfer between LTE 1 and LTE 2 in addition to transfer from either or both LTEs to STE as measured by mean time per keystroke and number of keystrokes per trial.

For the text editor experiment, or similarly designed studies, the researchers have little interest beyond the context of the target content area. This casts the study objective in terms of assessing the strength and direction of transfer that has taken place, as opposed to measuring a propensity toward successful transfer outcomes when faced with novel tasks in the future and the threshold of transfer distance achieved.

#### *2.2.6.2 Assessing propensity to apply learned knowledge to a novel task.*

The challenge of assessing propensity to apply learned knowledge to a novel task is essentially rooted in the problem of measuring the magnitude of abstraction or generalizability achieved by a learner. Because there are no externally defined boundaries in this case, the researcher is faced with difficult choices about appropriate target domain(s) and transfer distance. Moreover, propensity for transfer may vary by topic

within a discipline such that a student may successfully accomplish a transfer task related to correlation but not comparison of group means (Budé, 2006). Several researchers have discussed approaches to assess propensity to transfer knowledge to novel tasks through use of concept maps, analogy, isomorphs, and graduated prompting.

*2.2.6.2.1 Concept maps.*

Schau and Mattern (1997) advocated use of concept maps to encourage as well as measure connected understanding of concepts that may otherwise appear isolated to students. Broers et al. (2004) added that use of concept maps effectively stimulates students to self-explain, and described development of a cognitive search algorithm similar to one described by Salomon and Perkins (1989) for the purpose of backward-reaching high road transfer. Schau and Mattern (1997) further argued that proficiency of statistical reasoning and problem solving is conditional upon such connected understanding of interrelated ideas; students who understand statistical concepts as isolated procedures are likely to persist as novices. The major benefit of concept mapping is the minimally filtered access to the mental representations that students have developed (Schau & Mattern, 1997).

In practice, the use of concept maps for assessment can be accomplished in a variety of ways. The most unfettered access to student understanding is achieved when students develop a complete concept map with no other prompting or intervention (Schau & Mattern, 1997). However, this approach is heavily dependent on the ability of students to effectively express their understanding in the form of a concept map, which itself takes training and practice, and the variability of outcomes may become untenable for the

instructor to interpret, evaluate, and score (Schau & Mattern, 1997). A proposed compromise to reduce the burden for the evaluator while attempting to preserve access to a representation entirely conceived by the student is to use essay descriptions provided by the student that are then translated into a concept map or compared against a reference concept map defined by the instructor (Ruiz-Primo & Shavelson, 1996; Schau & Mattern, 1997). Simultaneously, the strength and limitation of this approach is the influence of the instructor's interpretation when translating the essay to a concept map. This reduces the dependence on the skill of concept map generation, which could disadvantage students that may be expert with the subject matter though poor at organizing their understanding as a concept map. However, it also introduces the chance that the instructor might infer an unintended or incomplete meaning resulting in poor inter-rater reliability (Ruiz-Primo & Shavelson, 1996).

Concept maps may be further modified for the purpose of assessment by providing the student with a blank or partially blank network of concept nodes that they are prompted to complete with or without a word bank (Ruiz-Primo & Shavelson, 1996; Schau & Mattern, 1997). This approach enjoys much higher psychometric reliability, and seems very common in the literature (Ruiz-Primo & Shavelson, 1996). It is not clear how one might optimize the approach for the purpose of promoting and measuring cognitive transfer, rather than recall of definitions.

*2.2.6.2.2 Analogical reasoning.*

Due to parallels between analogical reasoning and cognitive transfer, analogy tasks—A:B::C:D—have been proposed as a simple method to evaluate transfer outcomes

(Alexander et al., 1998; Alexander & Murphy, 1999; Helfenstein, 2005). Also, this approach is amenable to conventional forced-choice and short answer item formats (Alexander et al., 1998). When designed carefully, Alexander et al. (1998) contend that even incorrect responses to analogy tasks can provide rich insights about the level of understanding and transfer achieved by the student.

Alexander et al. (1998) proposed seven distinct categories of response to analogical reasoning tasks such that students were asked to complete "A:B::C:____" tasks. The response categories from lowest level of achievement to highest were described as (1) no response, (2) repetition (usually of B), (3) non-domain response, (4) structural dependency, (5) domain response, (6), target variant, (7) correct response (Alexander et al., 1998). While some response categories are self-evident, the authors clarify several others as follows: non-domain response indicates an attempt at an original response, but is not relevant to the target domain; structural dependency shows some effort to produce a response within the proper domain but simply provides a variant on the C term; domain response is an original response within the target domain, though incorrect; target variant is nearly correct, but uses the wrong form—part of speech, conjugation, etc.—of the intended D term (Alexander et al., 1998).

Alexander et al. (1998) conducted two studies to demonstrate this method including 429 sixth grade students in the first and 329 university students in the second. Both studies were similarly designed and administered, such that students were evaluated using various baseline assessments, then given an assessment with a series of A:B::C:D analogy tasks that were scored using the seven categories described above (Alexander et

al., 1998). The authors asserted that the first study produced a non-random error pattern and described anecdotes of students whose errors were confined to a single category in order to bolster the credibility of the proposed hierarchy of errors (Alexander et al., 1998). A canonical correlation analysis employed in the first study showed statistically significant evidence of an effect, however, several categories were scarcely used and inspection of the canonical vector loading largely indicated that correct responses simply correlated with domain knowledge. Analysis of the second study included a partial credit item response model that showed poolability of responses into the following categories (1), (2-4), (5), (6-7), indicating that the theory may have credibility, but several of the defined response categories may be redundant.

*2.2.6.2.3 Isomorphs.*

Closely linked to assessment of transfer through analogical reasoning is isomorphic problem solving (Singley & Anderson, 1989). Two tasks, generally presented in narrative form, may be considered isomorphs when they are structurally the same but differ in superficial aspects such as "semantically distant" domains (Gick & Holyoak, 1980). One of several famous examples includes a military narrative in which students must discern that an army needs to divide it's forces and attack from different angles in order to conquer a fortified city, and it's isomorphic scenario of an oncologist seeking to destroy a tumor with radiation that must come from many directions in order to avoid unnecessary collateral damage to healthy tissue (e.g., Gick & Holyoak, 1980; Singley & Anderson, 1989). Bude (2006) has cautioned, however, that the researcher should be careful when designing such tasks because deviation too far from the target domain could confound

results due to the variability of familiarity with the target domain—radiation therapy in the latter case. In general, research has suggested that people struggle to tackle isomorphic tasks, but performance can be dramatically improved when prompted to consider the solution to a known isomorph (Gick & Holyoak, 1980; Singley & Anderson, 1989).

*2.2.6.2.4 Graduated prompting.*

In the ideal case, one might hope that students automatically recognize the need for transfer in order to accomplish the target task, but when this is not the case some amount of prompting can improve transfer outcomes substantially (Bransford et al., 2000). According to Singley and Anderson, "being reminded of the right problem is often more problematic than mapping the solution" (1989, p. 22). Consequently, tests that include graduated prompting have been suggested to access more detailed analysis of the state of learning and transfer present when compared to all-or-nothing tests of whether or not transfer has occurred (Bransford et al., 2000).

***2.2.6.3 Assessment of statistical thinking.***

Chance (2002, section 4 para. 5) summarizes that "evidence of statistical thinking lies in what students do spontaneously, without prompting or cue from the instructor. Students should be given opportunities to demonstrate their 'reflexes.'" A few suggestions have been proposed specifically for the purpose of assessing statistical thinking, though much of the guidance boils down to individual task recommendations and sample items rather than a dedicated assessment tool. One noteworthy exception is

the Models of Statistical Thinking (MOST) instrument described by Garfield et al. (2012).

*2.2.6.3.1 Item examples.*

A typical recommendation for developing a task to measure statistical thinking might include presenting students with data from a given study and ask how it might be analyzed in order to assess how readily students connect domain knowledge to novel applications (e.g., Budé, 2006; Garfield et al., 2012). Watson (1997) described another approach using statistics in the media as a tool to evaluate statistical thinking. Secondary education students were presented with media articles and asked several questions designed to probe interpretation. Watson (1997) discussed that some of the advantages to this approach including the fact that media consumption is truly a context that is encountered by all students and frequently warrants statistical thinking. Also, Watson (1997) claimed that since media often do not include source data, the items can access a higher level of abstract thinking than is typically the case if students become distracted by the mechanics of computation. However, little guidance was provided to aid the development of successful prompts to coax statistical thinking from students, except that they should be broad enough to allow various interpretations while still amenable to graduated prompting (Watson, 1997).

Alternatively, Chance (2002) describes a number of assessment items that have been modified from textbook exercises in order to assess statistical thinking outcomes. For example, Chance (2002, section 4) describes the following assessment item credited to Rossman and Chance (2001):

> The underlying principle of all statistical inference is that one uses sample statistics to learn something… about the population parameters. Convince me that you understand this statement by writing a short paragraph describing a situation in which you might use a sample statistic to infer something about a population parameter. Clearly identify the sample, population, statistic, and parameter in your example. Be as specific as possible, and do not use any example which we have discussed in class.

This item is interesting because a quality response could evince high road transfer of any distance, yet the overt nature of the task is somewhat surprising when compared to transfer tasks elsewhere proposed. Other assessment tasks proposed by Chance (2002) prompt students to consider confounding factors for a given scenario, critical analysis of outliers, critique of published methodology, and use of follow up questions to reveal the level of statistical thinking driving a student's solution to a statistical task.

*2.2.6.3.2 The MOST instrument.*

The MOST instrument is described as an assessment tool specifically created for the purpose of quantifying curriculum-independent statistical thinking outcomes (Garfield et al., 2012). The instrument was described to include eight items based on four real-world contexts (Garfield et al., 2012). The expectation is that students describe in detail how each scenario could be evaluated using statistical methods, but they were not asked to actually conduct the analysis (Garfield et al., 2012). Performance on each task was evaluated holistically using a rubric that included five facets deemed essential to complete statistical thinking (Garfield et al., 2012). The five facets defined by Garfield et al. (2012) are described in terms of a simulation-based approach, but essentially require

description of an acceptable chance model, accommodation for sampling variability, appropriate test statistic proposal, a method to calculate the associated *p*-value, and evaluation criteria for the *p*-value. Student responses were then evaluated to represent complete, partial, or incorrect statistical thinking based on the number of facets present (Garfield et al., 2012). Scoring each assessment task in this manner amounts to evaluation of a sub-goal architecture tailored to statistical problem solving in the same vein that Atkinson et al. (2003) recommended use of sub-goals as a metacognitive strategy to promote transfer.

To summarize, Chance (2002) described the goal in assessment of statistical thinking as a capability to measure plasticity of problem solving and critical thinking especially in the absence of explicit direction. Undoubtedly, these are challenging attributes to measure. However, tasks designed to promote high road transfer, use of media articles with graduated prompting, and the MOST instrument all have tremendous potential to inform the assessment of cognitive transfer within the context of statistics education.

## 2.3 Discussion of the literature.

Pfannkuch and Wild (2005) underscored that the development and implementation of instruction and assessment with the goal of promoting statistical thinking is critical for the development of the next generation of professional and citizen statisticians. This paper has synthesized several promising avenues for progress toward this goal. Review of the education research regarding optimization of cognitive transfer outcomes seems to align well with goals for developing statistical reasoning and statistical thinking.

### 2.3.1 Summary and critique.

#### *2.3.1.1 Promoting cognitive transfer.*

Clearly, transfer research has gone by many names in different circles of academic inquiry, yet its importance as an educational outcome is ubiquitous. The body of research has roots dating to Aristotle, yet modern psychologists seem to lend the topic of cognitive transfer as much significance and attention as ever (Bransford et al., 2000). If transfer outcomes are truly valued, then it is essential that mechanisms thought to promote them are studied and considered during evaluation of novel curricula.

One such mechanism includes the formation of a rich and interconnected network of schema consisting of abstracted cognitive elements with enough flexibility to easily assimilate novel content and contexts. Metacognitive techniques including problem solving frameworks like sub-goal architecture as well as abstraction strategies like study of worked examples and self-explanation have also demonstrated promise for promoting positive transfer potential. Management of cognitive load is also an important consideration, since increased efficiency may liberate cognitive resources for more strategic purposes essential to successful transfer. Strategies for assessment of cognitive transfer were considered and extended, where possible, by literature speaking directly to the context of statistics education. If an introductory statistics curriculum is to be optimized for production of successful transfer outcomes, careful consideration of each of these implications for instruction and assessment is necessary.

#### *2.3.1.2 Assessment of cognitive transfer outcomes.*

As discussed, attention to schema development, metacognitive strategies, and cognitive load enjoy fairly broad acceptance as considerations for instruction and

curriculum development for positive transfer outcomes. However, without clear methods and reliable tools to measure cognitive transfer outcomes, there is no way to quantify or even substantiate the benefit of these practices. If the desired outcome is vertical transfer, assessment may amount to evaluation of increased speed in learning a new content area. Arguably, this seems a more straight-forward task since otherwise vague parameters like distance of transfer are resolved, at least in part, by the requirements of the target content area. If the desired outcome is not vertical transfer, assessment seems to be a more complicated task because choices related to distance and target contexts become more subjective.

Concept maps and isomorphs have been discussed widely in the literature, though it is not clear how effective these strategies would be for the specific goal of evaluating transfer outcomes of introductory statistics curricula. Concept maps have the appeal of producing a physical representation of functional schema in the mind of each student, but it is unclear how to overcome some of the barriers to implementation and assessment of concept maps to access propensity for transfer. Isomorphs tend to be discussed as a relatively pure form of transfer assessment, though in the literature their use typically accompanies study of very general transfer outcomes as opposed to outcomes related to a particular content area.

Analogical reasoning seems to balance a natural conduit to cognitive transfer without onerous implementation or scoring concerns. Some research has suggested that error categorization associated with analogy tasks is capable of producing rich insights about the state of a student's knowledge on the subject matter. A careful analysis of the results

accompanying these studies indicates that the theory may have credibility, but several of the defined response categories may be redundant.

The literature pertaining to assessment of statistical thinking reveals that a strong starting point for assessment of transfer outcomes in statistics may be characterized by tasks that present students with data or scenarios and ask them to describe an appropriate method of analysis. This approach assesses how readily students connect domain knowledge to novel applications, and does so in a manner that closely relates to evaluation of high road transfer outcomes. For example, if students are directed by some specific research question(s), this paradigm aligns the assessment task with backward-reaching high road transfer, and if students are asked to propose new research questions that might be addressed by a described study design and provided data the task could assess forward-reaching high road transfer outcomes.

The MOST assessment is a particularly interesting tool because it currently appears to be the one of the only curriculum independent resources developed for the express purpose to evaluate statistical thinking outcomes. Its content invites high-road transfer and each task is flexible enough to invite a wide variety of responses from students including solutions predicated on parametric, nonparametric, and simulation-based approaches. This is an important step to lay groundwork for comparison of statistical thinking outcomes of different curricula. However, the scoring rubric designed to accompany the MOST assessment probably requires additional work before the tool can be used to evaluate or compare statistical thinking outcomes for students using non-simulation methods. The rubric described to accompany the MOST assessment is a

strong tool for consistent evaluation of statistical thinking outcomes among students in a simulation-based curriculum, but it is not obvious how the given rubric would accommodate responses that do not rely on simulation for equitable comparison. Another possible improvement could be availability of graduated prompts for students who do not succeed in demonstrating complete statistical thinking on their own.

Such prompting could effectively jump start high road transfer in order to draw the task within the student's zone of proximal development. A record of the graduated prompts provided to each student could then be used to inform the distance of transfer achieved for each student. With this capability to dynamically modify the transfer distance and difficulty of each task as needed, the assessment tool may produce more precise estimates of transfer potential for each student.

### 2.3.2 Implications for teaching.

If successful transfer is a desired outcome of the introductory statistics curriculum, then it is essential that mechanisms believed to improve transfer are carefully considered when developing and evaluating curricula. During instruction, cognitive load should be carefully managed to prevent students from becoming overwhelmed and to protect a portion of their cognitive capacity for more strategic purposes. Moreover, effective curricula should promote formation of a rich and interconnected schema and emphasize abstraction of cognitive elements so students are prepared to adapt and apply their learning in new situations. Once cognitive load is well-managed, and proper schema development are in place, metacognitive strategies should be added to impose intentional structure for the expressed purpose of transfer. Since it is difficult to improve and sustain

outcomes that are not measured, cognitive transfer outcomes should be explicitly evaluated as part of the assessment strategy for an introductory statistics curriculum.

### 2.3.3 Implications for research.

The literature reviewed and subsequent discussion in this paper prompts a number of implications worthy of future research. Of specific interest may be issues related to the impact of nonparametric and simulation-based introductory statistics curricula on cognitive transfer outcomes capable of reaching beyond the classroom. However, little experimental research has been published to address these issues at present.

Since cognitive load is believed to be relatively constant for a given subject matter, additional research is needed in order to investigate whether the presentation of introductory statistics through nonparametric and simulation-based methods is a departure from the traditional curriculum radical enough to alter the fundamental element interactivity required of learners. If intrinsic cognitive load accompanying simulation-based methods differs from non-simulation methods, it is necessary to understand the burden imposed by each curriculum so that characteristic can be properly weighed among other potential benefits when evaluating curricula. If one approach can be demonstrated to lower net cognitive burden during all or part of the curriculum, it could have implications for the speed that learners achieve abstraction of cognitive elements and ability to incorporate metacognitive strategies.

There is also a need for research evaluating the distance of transfer achieved by students after a particular curriculum, as well as comparisons of outcomes for students exposed to different curricula. The expectation is that students are likely to perform better

with near transfer tasks since they are closer to the form in which the content was learned, yet performance on far transfer tasks may approach a level of performance seen on near transfer tasks based on the extent that abstraction is achieved. Therefore, the nature of disparity between near transfer performance and far transfer performance may be an indicator of how well or how poorly students have abstracted the cognitive elements presented by a specific curriculum.

High road transfer outcomes may be of particular interest since it is better suited for conventional secondary and post-secondary curriculum design than low road transfer. To reiterate, forward-reaching high road transfer relates to the ability of the student to distill a given task down to its essential elements and think creatively about new tasks that might be similar, whereas backward-reaching high road transfer relates to the ability of the student to search his experience with other tasks in order to discern what schema may be useful for the problem at hand. Both influence potential to accomplish transfer tasks in novel scenarios. As such, it may be important to understand how simulation-based and traditional approaches to the introductory curriculum impact the ability to achieve forward-reaching and backward-reaching high road transfer outcomes.

There is also little research evaluating transfer outcomes after any appreciable delay, since successful transfer outcomes seem likely to decay with time. The research that has been conducted on this topic considers delays on the order of a few days or weeks, but research is needed to evaluate transfer outcomes after longer periods. Specifically, it may be interesting to investigate whether different models of the introductory curriculum

impact the rate at which propensity to transfer applied statistics content to novel problem scenarios decays over time.

Finally, additional research is needed to refine the assessment of cognitive transfer outcomes within the context of introductory statistics. The MOST assessment may represent a very promising foundation, but additional study may be warranted to learn whether the tool would benefit from graduated prompting and other item types. More importantly, research is needed in order to develop a scoring rubric for non-simulation approaches that can facilitate equitable comparison to simulation-based approaches. A curriculum independent assessment tool is critical for such comparisons, but it cannot be utilized to its greatest potential until it is accompanied by an equitable scoring strategy. Only then can researchers begin to make the reliable comparisons of statistical thinking outcomes that are needed to advance curriculum development and empower students to transfer statistical understanding to contexts beyond the introductory statistics course.

### 2.3.4 Problem statement

Based on the literature reviewed, much can be done to promote and assess successful cognitive transfer outcomes for students of introductory statistics. However, no published assessment existed to measure this specific outcome, and the literature indicates uncertainty about whether cognitive transfer outcomes can be achieved and measured following an introductory statistics curriculum. A new assessment tool should be developed for the purpose of quantifying cognitive transfer outcomes for introductory statistics students.

# 3 Methods

## 3.1 Research Question

The research question of this dissertation is: Can a new assessment tool with good psychometric properties be developed to quantify cognitive transfer outcomes for introductory statistics students?

## 3.2 Study Overview

The purpose of this study was to explore the feasibility of creating an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. In order to develop a high quality assessment of cognitive transfer outcomes following a first statistics course, the Introductory Statistics Understanding and Discernment Outcomes (I-STUDIO) instrument was developed and revised through an iterative process including expert feedback and piloting outlined in Table 2. Planning and development included generation and critique of a test blueprint to make explicit the structural organization of the instrument. Candidate tasks were then developed according to published standards and critiqued for inclusion according to their contribution based on criteria dictated in the test blueprint.

Expert feedback and cognitive interviews with student participants revealed improvements to instructions, tasks, and structural considerations that were addressed prior to finalizing the instrument for large-scale field testing. Instructors were recruited to participate in the field test through various methods in order to seek diversity in curriculum design and instructional practice. The instructors presented the instrument to students at or near the end of an introductory statistics course, and results were collected

electronically for analysis. Data analysis included evaluation of the reliability and validity of the instrument as well as abbreviated item analysis. Additionally, qualitative analysis of expert feedback provided insight about the contribution of the instrument for curriculum development decisions.

The development timeline for the study shown in Table 2 catalogues major project milestones and associated completion dates. Although many of the tasks in Table 2 have sequential dependencies, some were executed in parallel where possible. For example, IRB approval must be complete prior to cognitive interviews, but this work was concurrent to expert feedback and the resulting task or instrument revisions.

Table 2

*I-STUDIO Development Timeline*

| Task Name | Completion Date |
|---|---|
| Draft Test Blueprint | October 16, 2014 |
| Expert Feedback: Test Blueprint | November 11, 2014 |
| Final Test Blueprint | November 16, 2014 |
| IRB Approval | November 24, 2014 |
| First Draft I-STUDIO Instrument | November 28, 2014 |
| Expert Feedback: I-STUDIO Instrument | December 23, 2014 |
| Second Draft I-STUDIO Instrument | January 15, 2015 |
| Cognitive Interviews | January 27, 2015 |
| Final I-STUDIO Instrument for Field Test | April 4, 2015 |
| Instructor Participants Recruiting | April 4, 2015 |
| Field Test Data Collection | May 14, 2015 |
| Peer Review of Rubric | June 12, 2015 |
| Final Scoring Rubric | June 12, 2015 |
| Peer Validation of Rubric | July 1, 2015 |
| Score Field Test Data | July 20, 2015 |
| Data Analysis | August 2015 |
| Synthesis of Study Results | September 2015 |

## 3.3 Instrument Development Cycle

### 3.3.1 Defining the construct for measurement.

The construct measured in the study was defined as the ability to transfer conceptual understanding of statistical inference for use in novel problem settings. This construct was thought to require both the ability to identify novel problem scenarios that warrant application of statistical inference, and the ability to achieve forward-reaching and backward-reaching transfer of statistics knowledge. Specifically, backward-reaching transfer tasks included discernment of whether a given problem setting would or would not benefit from application of statistical inference and demonstration of an appropriate solution strategy. The desired evidence was more conceptual than procedural, so students were encouraged to frame solutions as though they were giving advice to a classmate rather than producing computations or formulas in order to demonstrate the target construct. Forward-reaching transfer tasks described a conceptual model or problem solving archetype and asked students to generate a context or scenario and map specific components of the conceptual model to the scenario they have chosen.

### 3.3.2 Test blueprint.

The test blueprint embodied the explicit plans that framed development of the I-STUDIO assessment tool. This included the definition of assessment outcomes, relative weight of each target outcome, the types of items used, the total number of items intended, and the distribution of these items with respect to target outcomes. The primary cognitive outcomes for consideration include discernment of whether statistical inference

is appropriate for a problem setting (i.e. discernment) and demonstration of high-road transfer (i.e. backward reaching and forward-reaching transfer).

The assessment was expected to include primarily (though perhaps not exclusively) open-ended tasks. Such tasks are labor intensive for students to complete, so the initial framework anticipated that each of the primary outcomes would be assessed by two or three contexts (e.g. data sets, data stories, etc.) with one or more open-ended prompts accompanying each. Consequently, the item allocation was relatively simple since the assessment consisted of relatively few open-ended tasks. Table 3 reproduces the table of item allocation found in the test blueprint (Appendix C: Final Test Blueprint). Note that since the instrument is intended to evaluate forward and backward high road transfer of statistics knowledge, the high road transfer tasks were nested within (not crossed with) the tasks assessing discernment of benefit from statistical approach.

Table 3

*Example of items classified by assessment goals*

| | Transfer Mechanism | | |
|---|---|---|---|
| Discernment Required? | Forward-Reaching | Backward-Reaching | Column Total |
| Yes, statistical inference appropriate | 2 | 2 | 4 |
| Yes, no statistical inference required | 1 | 1 | 2 |
| No | 0 | 1 | 1 |
| Row Total | 3 | 4 | 7 |

The test blueprint then defined the five item characteristics evident in the row and column labels of Table 3:

- forward-reaching high road transfer;

- backward-reaching high road transfer;
- discernment required – statistical inference appropriate;
- discernment required – statistical inference not appropriate;
- no discernment required.

The test blueprint also provided detailed descriptions of the six possible item types represented by each cell in the body of the table:

- backward-reaching high-road transfer with discernment—statistical inference appropriate;
- backward-reaching high-road transfer with discernment—statistical inference not appropriate;
- backward reaching high-road transfer with no discernment;
- forward-reaching high-road transfer with discernment—statistical inference appropriate;
- forward-reaching high-road transfer with discernment—statistical inference not required;
- and forward-reaching high-road transfer with no discernment.

It is relevant to point out that there were no items assigned to the cell corresponding to forward-reaching high-road transfer with no discernment required. The reason for this omission is predicated on the operational definitions of those two characteristics described in the test blueprint. In short, forward-reaching high-road transfer requires that students are given one or more abstract principles, and then they are instructed to invent and describe a novel application. When the student is asked to describe an application of

statistical inference, then the student has exercised discernment in choosing an appropriate application to describe. The same is true if the student is asked to describe an application that does not require statistical inference. Consequently, forward-reaching high-road transfer by definition must include some measure of discernment. Therefore, the I-STUDIO assessment cannot include any forward-reaching high-road transfer item with no discernment required.

Lastly, a draft item or description of a draft item was included in the test blueprint to illustrate each of the possible item types described. The initial test blueprint was developed under the supervision of the graduate advisors for the project. A panel of expert reviewers then reviewed a complete draft of the test blueprint. Each panel member had expertise in at least one of the following faculties: statistics; education; educational measurement; cognitive transfer. Reviewers were provided a questionnaire designed to direct specific feedback and recommendations for critical components of the test blueprint (Appendix B: Expert Feedback Questionnaire Accompanying Test Blueprint).

The final roster of expert reviewers for the test blueprint was:

- Sanford Weisberg (University of Minnesota – Statistics Dept.)
- Roxy Peck (California Polytechnic State University – Statistics Dept.)
- Sashank Varma (University of Minnesota – Educational Psychology Dept.)
- Beth Chance (California Polytechnic State University – Statistics Dept.)
- Marsha Lovett (Carnegie Mellon University – Psychology Dept.)

### 3.3.3 Item writing.

Amendments to the test blueprint were then taken into consideration during item writing and development of the first draft of the I-STUDIO instrument. The draft items were developed under the supervision of the graduate advisors for the project. Candidate items were designed and developed according to published standards and best practices. Items were then organized to produce a draft I-STUDIO instrument submitted to the panel of expert reviewers that had reviewed the test blueprint for feedback. Minimally acceptable responses were drafted to accompany the draft instrument so that expert reviewers could have a general sense of acceptable responses for each item. However, the final rubric was informed by actual student responses, and therefore most rubric development took place following collection of field test data.

#### *3.3.3.1 Item design.*

In order to achieve the goals of the I-STUDIO assessment, the tasks were written in an open-ended format (i.e. constructed/produce response). According to Thorndike and Thorndike-Christ (2010), "the major advantage of the produce-response, or essay, type of question lies in its potential for measuring examinees' abilities to organize, synthesize, and integrate their knowledge; to use information to solve novel problems; and to demonstrate original or integrative thought." Such remarks corroborate the appropriateness of open-ended tasks since the aforementioned outcomes align closely with the definition of cognitive transfer outcomes the I-STUDIO assessment intended to measure.

A drawback of open-ended tasks is that content knowledge may be confounded with the ability to organize and synthesize a coherent response (R. M. Thorndike &

Thorndike-Christ, 2010). While perhaps true in general, since the target construct could be described as an interaction between the content knowledge and organization of schema, the concern of the stated drawback is minimally (if at all) problematic for the goals of I-STUDIO assessment. Another challenge presented by open-ended tasks is the time required to produce thoughtful responses. In order to mitigate this challenge, instructors were permitted to offer the assessment tool for use outside of class although the constraint of student fatigue was still present.

#### *3.3.3.2 Item development.*

The items included in the I-STUDIO assessment were selected from a pool including tasks adapted or adopted from published sources as well as original tasks developed by the author. Item development adhered to published guidance and best-practices in the literature (e.g. AERA, APA, & NCME, 1999; Haladyna & Rodriguez, 2013; R. M. Thorndike & Thorndike-Christ, 2010). These guidelines include attention to suitability of content presented in each item with respect to pertinence to the target domain, appropriate cognitive demand, and consistency of expectations among similar tasks (Haladyna & Rodriguez, 2013). Haladyna and Rodriguez (2013) also recommend that careful attention be paid to write specific instructions that include information about the desired format of a quality response. Additionally, item development attended to cultural diversity and appropriate level of language sophistication in order to mitigate these sources of construct-irrelevant variance (Haladyna & Rodriguez, 2013).

#### *3.3.3.3 Draft instrument review.*

The draft I-STUDIO instrument was developed under the supervision of the graduate advisors for the project. As with the test blueprint, the complete draft I-STUDIO instrument was then reviewed by the panel of expert reviewers (Appendix D: Draft I-STUDIO Version Prior to Expert Feedback). The final roster of expert reviewers for the draft I-STUDIO instrument was:

- Sanford Weisberg (University of Minnesota – Statistics Dept.)
- Roxy Peck (California Polytechnic State University – Statistics Dept.)
- Sashank Varma (University of Minnesota – Educational Psychology Dept.)
- Tim Jacobbe (University of Florida – Teaching & Learning Dept.)
- Beth Chance (California Polytechnic State University – Statistics Dept.)
- Marsha Lovett (Carnegie Mellon University – Psychology Dept.)

As with the test blueprint, reviewers were again provided a questionnaire designed to direct specific feedback and recommendations for critical components of the draft I-STUDIO instrument (Appendix E: Expert Feedback Questionnaire Accompanying Draft I-STUDIO Assessment Tool). The instrument was then updated to reflect recommended changes before use with students, and then revised again following observations extracted from the cognitive interview data.

#### *3.3.3.4 Item scoring.*

Since all I-STUDIO tasks were open-ended, scoring decisions required careful consideration. Depending on the nature of the task and the expectations for task performance, an open-ended task may accommodate objective as well as subjective

scoring criteria. Objective scoring was used where possible in order to reduce the dependence on subject matter expertise and subjective judgments that may differ between raters or even within a rater over time (Haladyna & Rodriguez, 2013). For example, the rubrics for item 5 (matched pairs study design), item 6 (underlying principle of inference), and item 7 (inference not required) compare the response against a checklist of target characteristics. A subjective scoring approach using a pre-defined rubric was used for item 1 (ATC preparation), item 2 (note identification), item 3 (display screen inspection), and item 4 (Walleye fishing), and their constituent subtasks.

Rubrics for items scored subjectively were designed to assess different levels of quality in response (e.g., essentially correct; partially correct; incorrect). Examples were provided to describe work commensurate with each score in the rubric and illustrate detail that is irrelevant to the target domain. Minimally acceptable responses were created to accompany the draft I-STUDIO instrument during expert review, but final item rubrics were developed and tuned using a sample of actual student responses.

The student responses used for rubric development were selected as a stratified random sample from the pool of usable responses collected in the field test. Three randomly selected students were chosen from 13 unique courses that participated in the field test for a total of 24 complete student responses. For each subtask, the 24 responses were ranked by desirability and noted for exceptional features. Themes among responses were then translated into rubric criteria for the subtask. Model responses were selected for inclusion in the rubric as exemplars of each scoring level. The draft rubric was developed under the supervision of graduate advisors to the project, and then the rubric

and a small number of actual student responses were provided to a Statistics Education PhD candidate for additional feedback that was used to further refine the rubric.

### 3.3.4 Iterative piloting process.

Following expert review of the draft assessment tool, the I-STUDIO instrument was updated to accommodate reviewer feedback and prepared for use with students. The first iteration of piloting with students consisted of cognitive interviews with five student volunteers. The second iteration was then a large-scale field test of the final instrument once observations from the cognitive interviews had been incorporated. All student data gathered was stored in a password-protected location on a local hard-drive or cloud storage service, and de-identified prior to data analysis and reporting. A more detailed description of each iteration cycle follows.

#### *3.3.4.1 Cognitive interviews.*

Following IRB approval (#1411E55223) the first stage of instrument piloting consisted of cognitive interviews—sometimes called "think-aloud" exercises—during which the student was asked to complete the assessment tool while attempting to verbalize their stream of consciousness with a silent "interviewer" present. The interview adhered to a consistent protocol with each participating student, only occasionally deviating to remind a student to verbalize their train of thought or probe to better understand emergent thinking patterns. One goal of the exercise was to glean information from students about whether tasks or instructions caused confusion or misinterpretation. Additionally, the interviewer captured the actual electronic responses to each task, which could then be mapped to the thinking patterns captured on an audio recording of each

student as he or she responded rather than attempting to infer that information from a completed response *post hoc*. Informative aspects of the problem solving process such as false starts or failed initial attempts could then be observed as they occurred where they would have otherwise gone undetected in a completed response if text had been erased or deleted.

Five introductory statistics students were recruited from the University of Minnesota to participate in cognitive interviews. Students were each compensated with a $20 Amazon gift card funded by the author for completing the exercise. All interviews were conducted in person on the University of Minnesota campus. Since the current version of the I-STUDIO assessment tool was intended for use with students that had completed (or nearly completed) at least one course in statistics, interview participants were recruited among students that had recently completed an introductory statistics course.

The recruiting effort attempted to include students with experience from different types of introductory statistics courses (e.g., simulation-based, non-simulation-based, and hybrid). No students from the non-simulation-based course volunteered, but two students had come from a simulation-based curriculum and three students had come from a hybrid course including roughly equal treatment of both simulation-based and non-simulation-based methods.

Following completion of the cognitive interviews, the completed assessments and interview notes were compiled for qualitative data analysis. The data were reviewed for patterns and themes that were incorporated into the instrument prior to the large-scale field test (Appendix G: I-STUDIO Version for Field Test).

#### *3.3.4.2 Large-scale field test.*

The large-scale field test aimed to collect data from 6 to 10 class sections including approximately 120 to 180 students using the final version of the I-STUDIO assessment tool. A secondary goal was to represent curriculum diversity so the I-STUDIO instrument and rubric development could be tested by a variety of students, courses, and use scenarios. This section describes the process of recruiting instructor participants followed by a description of the sample actually obtained.

##### *3.3.4.2.1 Recruiting.*

In order to solicit participation of introductory statistics instructors, announcements were broadcast through the Isolated Statisticians of the American Statistical Association (ISOSTAT), American Statistical Association (ASA) Section on Statistical Education, and Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), list serve outlets. These list serves are likely to have considerable overlap among membership, but were believed to provide several access points to a nation-wide target audience.

The ISOSTAT list serve has around 275 members who often, though not exclusively, teach statistics at small colleges or universities and are frequently the only statistician in their department (J. Witmer, personal communication, October 11, 2012). Officially, ISOSTAT is considered a subgroup of the ASA, however, neither ASA membership nor dues are required for ISOSTAT participation (ISOSTAT charter. n.d.). Similarly, the ASA Section on Statistical Education has approximately 1200 nationwide members represented largely by university and liberal arts college faculty with a few members

from community colleges, high schools, and industry (R. Nichols, personal communication, October 11, 2012). All members are enrolled in the list serve automatically, but are provided an opportunity to opt-out of email communications. CAUSE also hosts a list serve outlet that may be employed to advertise the instrument and recruit collaborators. This activity is well aligned with the goals outlined in the CAUSE charter (2006), which include connecting collaborators, sharing resources, and cultivate research visibility among undergraduate statistics educators.

*3.3.4.2.2 Sample.*

A total of 33 introductory statistics instructors responded to the recruiting effort. This initial set of instructors was then contacted with further information about the I-STUDIO assessment and the goals of the research study. Instructors were recommended to offer students course credit of some kind (e.g. homework, final exam review) in exchange for submitting a response to the assessment. Also, instructors were invited to make the assessment available to students outside of class with any resources they wish as long as they agreed to complete the assessment independently.

Fallout among instructor contacts had varied reasons. Some instructors were not able to accommodate the requested conditions of the study (e.g., course credit), others were teaching high school rather than post-secondary students, several more backed out of the study for unrelated personal reasons, and a handful were simply lost to follow-up. Finally, the I-STUDIO assessment was implemented by fourteen (14) instructor participants representing a total of 29 class sections for 16 unique courses at 15 institutions. A roster of participating institutions is shown in Table 4.

Table 4

*Distribution of usable responses by institution*

| Institution | Course | Usable Responses |
|---|---|---|
| California State University – Fullerton | MATH 120 | < 30 |
| College of Staten Island – CUNY | MTH 113 | < 30 |
| Community College of Vermont | MAT-2021 | < 30 |
| Florida State University | STA 2023 | 439 |
| Florida State University | STA 2122 | 375 |
| Heartland Community College | MATH 141 | 53 |
| Indiana University | PSY-K 300 | 35 |
| Iowa State University | STAT 101 | 261 |
| Maastricht University | | 39 |
| Marist College | MATH 130L | 39 |
| Mount Saint Mary College | MTH 2070 | < 30 |
| St Ambrose University | MATH 300 | < 30 |
| SUNY Buffalo State | MAT 311 | < 30 |
| University of Kentucky | BST 330 | 68 |
| University of Vermont | STAT 141 | 129 |
| Valdosta State University | BUSA 2100 | 59 |

Raw data included 1995 respondents, which was reduced to 1975 after removing responses submitted by instructors. A total of 1935 students consented to participation in the research study. Responses were submitted using a web-based application. Students were required to type a response to each item before they were permitted to move on to the next item. Omitting responses that abandoned the instrument (i.e., closed the web browser without submitting a complete response) resulted in a sample size of 1614 complete cases. Some responses were submitted with apparent complete data, however a subset of these were submitted in an unreasonably short amount of time. In order to restrict the data to more earnest attempts, attempts submitted in fewer than 10 minutes were omitted from the final data set. The resulting sample size included 1566 unique

student participants. Maximum course enrollment aggregated across all participating institutions was estimated as 2265, indicating a total response rate of 87% and total usable response rate of 69%. These estimates assume that every course ended the semester with maximum enrollment, and that no student submitted more than one response to the instrument. It is very unlikely that all courses had maximum enrollment at the end of the semester, and although administration of the I-STUDIO assessment was configured to block multiple responses from the same IP address students in one class were discovered to circumvent these measures.

Since the sample obtained was much larger than anticipated, a representative random sample from each course was selected for scoring and data analysis. The sampling method chose a maximum of 12 students from each course; all students were selected from courses with fewer than 12 complete submissions. The resulting sample for data analysis included 178 respondents. Note that a similar strategy was used to select students for rubric development. The data analysis sample of 178 students was selected first, and then the rubric development sample was selected from the remaining pool of unselected students. This strategy was used in order to preserve submissions from small classes for use in the primary analysis, while preventing any submission used to create the rubric from inclusion in the primary data set and analysis.

## 3.4 Data Analysis

### 3.4.1 Contribution of the instrument.

The contribution of the instrument as a tool to inform general curriculum and instruction outcomes was evaluated through mixed methods data analysis. The primary

data source was feedback provided by expert reviewers via a questionnaire that accompanied the test blueprint (Appendix B: Expert Feedback Questionnaire Accompanying Test Blueprint). Ordinal scale responses were tabulated, while open-ended feedback was aggregated and summarized thematically using a qualitative approach. Quantitative data analyses were conducted using the R statistical computing platform (R Core Team, 2014).

### 3.4.2 Rubric consistency.

The same Statistics Education PhD candidate who participated in the rubric development also participated in an independent scoring exercise in order to assess inter-rater consistency of rubric application to a randomly selected set of the student responses. Discrepancies were discussed in order to discern how the content of the rubric, scoring levels, or training for scorers might be refined for future use. Intra-rater consistency was also evaluated in order to capture evidence of drifting rubric interpretation over time by the author rescoring randomly selected responses. The delay between rescoring these randomly selected responses was no more than a few hours because all responses for a given item (e.g. item 4b) were scored within the space of a single day, often within a single sitting without interruption. Consequently, rubric validation evidence was derived from peer review evidence that the rubric appropriately reflects incremental response quality relevant to the target construct as well as analysis of rubric inter-rater and intra-rater consistency of rubric interpretation applied to actual student responses collected from the I-STUDIO field test.

### 3.4.3 Descriptive statistics.

Descriptive statistics of I-STUDIO total scores included summary statistics for the sample of 178 scored responses. This was followed by summary statistics by scoring element as well as summary statistics of total score for each unique course represented in the sample.

### 3.4.4 Reliability of the instrument.

The reliability of an assessment tool could be summarized as the ability of the instrument to generate repeatable and reproducible measurements of the desired construct. Repeatability reflects the precision of measurements gleaned from the same person if she were evaluated over and over without the influence of practice, learning, or fatigue. Reproducibility pertains to consistency of results gleaned from theoretically equivalent persons. It follows that the extent to which assessment results differ among individuals indicates how much of the trait to be measured is possessed by each individual. Therefore, the reliability of an instrument relates to the precision of its measurements, which effectively bounds its utility. As Thorndike & Thorndike-Christ (2010, p. 133) put it, “A test must measure *something* before it can measure what we want it to measure.”

Three paths toward demonstration of instrument reliability include testing and retesting the same students with the same assessment tool, testing and retesting the same students with different but equivalent assessment tools, or dividing an assessment tool into equivalent subsets after a single test-administration (R. M. Thorndike & Thorndike-Christ, 2010). Since only one form of the assessment tool was created, the second path is not pertinent. The first path involving a re-test would be complicated by practice and

learning effect among students, and an appropriate washout period to mitigate these effects was not feasible for this study though may have value in future research. In any case, the test blueprint and instrument development were aligned to support single-administration reliability analysis.

Cronbach's alpha and mean inter-item correlations were estimated, although these methods are conventionally predicated on the assumption that all items are intended to measure a unidimensional characteristic. Since the construct involves more than one underlying trait, reliability may be more appropriately estimated using a method based on a judicious partition of the instrument to produce halves that are carefully matched based on item type and difficulty. One such method uses the Spearman-Brown formula: (Equation 1)

$$\hat{r}_{tt} = \frac{2r_{AB}}{1 + r_{AB}}$$

such that $\hat{r}_{tt}$ is the reliability of the total test and $r_{AB}$ is the correlation between the two halves of the instrument. The process of partitioning the two halves for estimation of Spearman-Brown reliability coefficient was conducted by matching homogenous scoring elements as grouped in Table 5.

Table 5

*Homogeneous item groups for split-half reliability estimation*

| Group | Homogeneous Tasks | Description |
|---|---|---|
| 1 | 1a & 1b | Research question proposal |
| 2 | 2a, 3a, & 4a | Discernment when inference is warranted |
| 3 | 1c, 2b, 3b, & 4b | Data analysis strategy proposal |
| 4 | 5, 6, & 7 (context scores) | Create context for forward-reaching transfer |
| 5 | 5, 6, & 7 (component scores) | Map task components to context |

The split for groups 1 and 3 were straightforward, but groups 2, 4, and 5 required additional attention. The strategy ultimately employed was to randomly select one of the three elements for the first half and similarly assign another randomly selected element to the second half, and then drop the third element. Groups 4 and 5 were further constrained such that the two elements selected from group 5 corresponded to the two elements selected from group 4. For example, if the item 5 context was selected for inclusion, then the item 5 component was selected to the same half. This resulted in six item pairs including 12 of the 15 scoring elements. The process was then repeated 500,000 times to simulate a distribution of Spearman-Brown reliability estimates and then the mean, median, and a 95% confidence interval were reported. The confidence interval was estimated using the 0.025 and 0.975 quantiles of the bootstrap sampling distribution constructed from the 500,000 simulations.

The simulated Spearman-Brown coefficient would be expected to underestimate the actual reliability of the I-STUDIO assessment tool since reliability is affected by the number of items in an instrument and only 12 of the 15 I-STUDIO scoring elements were included for each simulated estimate. The following adjustment projects a calculated reliability estimate to a form with a different number of items:

$$\hat{r}_{KK} = \frac{kr_{tt}}{1 + (k-1)r_{tt}}$$

where $\hat{r}_{KK}$ is the projected reliability of an instrument with *k* times as many items (i.e.

including 15 elements), and $r_{tt}$ is the reliability of the original test (i.e. including 12 elements).

Based on calculated estimates of the total variability of test scores, $SD_X$, and the reliability of the total test indicated by Equation 1, the standard error of measurement for the instrument is estimated by Equation 2.

(Equation 2)

$$SD_e = SD_X\sqrt{(1 - \hat{r}_{tt})}$$

The standard error of measurement for the instrument was calculated from the distribution of simulated Spearman-Brown reliability estimates, and reported along with a 95% confidence interval. The confidence interval was again estimated using the 0.025 and 0.975 quantiles of the bootstrap sampling distribution constructed from 500,000 simulations. Since Spearman-Brown reliability was estimated based on 12 scoring elements for the simulation and projected to estimate reliability of all 15 scoring elements, the standard error of measurement was calculated and reported in both cases as well.

### 3.4.5 Validity of the instrument.

#### *3.4.5.1 Overview of validity evidence.*

As a holistic concept, validity can be summarized as the degree to which assessment outcomes provide information that is relevant to the inferences to be made from them. Instrument reliability, as discussed previously, is a necessary condition of validity, but it is not sufficient. To paraphrase Thorndike and Thorndike-Christ (2010), instrument reliability provides evidence that the assessment tool is measuring *something* and validity

ensures that the instrument is measuring the *right* thing. An instrument that measures the "wrong thing" with great precision would still have poor validity since it cannot be used to support the intended inferences.

Thorndike and Thorndike-Christ (2010) describe different approaches to characterizing validity evidence including a segmented approach and a unified approach. The three primary constituents of the segmented validity perspective include content-related, criterion-related, and construct-related evidence. Content-related evidence of validity is concerned with both the factual domain knowledge demonstrated and the cognitive processes employed by students as they engage that knowledge (R. M. Thorndike & Thorndike-Christ, 2010). Criterion-related evidence of validity often includes some statistical analysis of how closely successful performance on the instrument correlates with empirical outcomes that the assessment was designed to predict or approximate (R. M. Thorndike & Thorndike-Christ, 2010). Lastly, construct-related evidence of validity reflects whether the instrument produces results that are consistent with theoretical predictions (R. M. Thorndike & Thorndike-Christ, 2010).

By contrast, the unified approach to validity evidence is predicated on the integrity of the inferences as opposed to characteristics of the instrument (R. M. Thorndike & Thorndike-Christ, 2010). In this way, construct-validity has been said to subsume the holistic notion of validity in which the task for test validation becomes a process of developing the most compelling case possible for the inferences that we intend to make (R. M. Thorndike & Thorndike-Christ, 2010). These inferences then are characterized based on possible interpretation of assessment scores and actions warranted from them.

Validity evidence will defer to the unified approach for the purpose of this paper. Even with a unified concept of validity, the evidence required to substantiate validity claims may be quite diverse. Such evidence included peer review provided by individuals with demonstrated expertise in the subject matter of interest and careful analysis of field test data.

#### *3.4.5.2 I-STUDIO validity evidence.*

Expert reviews of the test blueprint, draft I-STUDIO instrument, and accompanying questionnaires provided validity evidence of the contribution of the instrument and the degree to which the assessment outcomes align with the intended construct. Peer review of draft scoring rubrics and evaluation of inter-rater consistency were used to establish confidence that the rubric adequately characterizes the continuum of possible responses and provides sufficient detail to evoke consistent judgments from qualified raters (AERA et al., 1999; R. M. Thorndike & Thorndike-Christ, 2010).

Confirmatory factor analysis (CFA) models were evaluated on the basis of both statistical and conceptual fit. Statistical models were executed using the *lavaan* package in R (Rosseel, 2012). The initial conceptual model tested was based on the model outlined in *Figure 1*. Alternative configurations of this model were also evaluated which reconfigured the I-STUDIO discernment component as subordinate to backward-reaching transfer, as well as studied the effect of accommodating correlated items directly and by aggregating scoring elements into testlets. Comparisons to reduced and alternative models were conducted using likelihood ratio tests and fit diagnostics.

Model fit diagnostics included Chi-square test for multivariate normality, the ratio of Chi-square test statistic to its degrees of freedom, goodness of fit, adjusted goodness of fit, proportion of residual item correlations greater than 0.1, proportion of residual item correlations greater than 0.05, root mean square error of approximation (RMSEA), 90% confidence interval for RMSEA, McDonald's noncentrality index, and Hoelter's critical N (Beaujean, 2014). The Chi-square test for multivariate normality is the conventional diagnostic to measure how closely the convariances calculated based on parameter estimates of the model are to those calculated from the sample directly. In practice, goodness-of-fit tests of this sort tend to reject given a large enough sample size, so alternative diagnostics were used to provide additional perspective for judging model fit.

The goodness of fit, and adjusted goodness of fit diagnostics are analogous to familiar $R^2$ counterparts, and values closer to 1.0 indicate better fit. Proportion of residual item correlations greater than 0.1 (or 0.05) is simply a screen to suggest whether important factors or correlation structure have been overlooked; lower proportion indicates better fit. RMSEA (and corresponding 90% confidence interval) is designed to assess whether a model reasonably approximates the data (as opposed to assessing exact fit); lower than 0.05 is desirable with values closer to zero indicating better fit. McDonald's noncentrality index is a function of the scaled noncentrality parameter for the model of interest, such that values closer to 1.0 indicate better fit. Hoelter's critical N is an estimate of the sample size at which the Chi-square statistic associated with the model of interest would reject the null hypothesis; values greater than 200 are considered desirable. All of the model diagnostics described here are considered "absolute fit indexes" because they do

not compare models or evaluate the improvement over a base model. This is significant because one of the CFA models studied required aggregation of the data into testlets, so it did not have the same base model as non-testlet CFA models. Therefore, any fit diagnostic that involved improvement over a base model would not be a fair comparison among all candidate CFA models studied.

Lastly, analysis of the reliability of the instrument was taken into consideration as a prerequisite to validity. An instrument with very low reliability has little utility as a tool to support any kind of inference. Validity evidence was therefore comprised of judgment by expert reviewers to measure the domain of interest, appropriateness of the test blueprint, confirmatory factor analysis, and demonstration of adequate reliability.

### 3.4.6 Item analysis.

The response patterns of each item were analyzed using quantitative as well as qualitative methods. Quantitative analysis based on multidimensional item response theory was conducted using the *mirt* package in R (Chalmers, 2012). Since the available sample size (178) is fairly small, IRT results were expected to provide little more than preliminary test information and general item functioning. The analysis compared partial credit and graded response models. The partial credit (PC) model shows the conditional probability of an individual with ability $\boldsymbol{\theta}$ achieving a score of $x_j$ (Masters, 1982):

$$\text{Partial Credit: } P\left(x_j|\boldsymbol{\theta}, \delta_{jh}\right) = \frac{\exp\left[\sum_{h=0}^{x_j}\left(\boldsymbol{\theta} - \delta_{jh}\right)\right]}{\sum_{k=0}^{m_j}\exp\left[\sum_{h=0}^{k}\left(\boldsymbol{\theta} - \delta_{jh}\right)\right]}$$

such that $j$ is the number of items, $\delta_{jh}$ represents the difficulty achieving a score of $h$ over a score of $(h-1)$, and $m_j$ is the maximum score for item $j$. The graded response (GR)

model is conceptualized somewhat differently and estimates the probability of an individual with ability $\boldsymbol{\theta}$ achieving a score of $x_j$ or better (Samejima, 1969):

$$\text{Graded Response: } P_{x_j}(\boldsymbol{\theta}) = \frac{\exp\left[\boldsymbol{\alpha}'_j\left(\boldsymbol{\theta} - \delta_{x_j}\right)\right]}{1 + \exp\left[\boldsymbol{\alpha}'_j\left(\boldsymbol{\theta} - \delta_{x_j}\right)\right]}$$

such that *j* is the number of items, $\delta_{x_j}$ represents the category boundary location, and $\boldsymbol{\alpha}_j$ is the vector of discrimination parameters for item *j* on each dimension.

Models were compared based on Akaike Information Criterion (AIC), and item fit was evaluated using S-X2 (Kang & Chen, 2008). In general, AIC is useful to compare models fit to the same data set; lower AIC is desired. S-X2 can be interpreted as analogous to a Chi-square test statistic over each score category of a polytomous item and groups of respondents with approximately homogeneous ability; a statistically significant result indicates evidence of poor fit for the corresponding item. Test information, item information, factor loadings, and model coefficients were evaluated for the final graded response model. Qualitative analysis consisted of noting unusual or unexpected response patterns.

## 3.5 Chapter Summary

This chapter explained the process of developing, refining, and implementing the I-STUDIO test blueprint and assessment tool. The chapter also described the data collection and methods used to estimate reliability, establish validity, and evaluate item responses. The next chapter reports the results of the study.

# 4 Results

## 4.1 Introduction

This chapter explains the results of the instrument development process, field test, and data analysis. Content includes summary of expert feedback relating to the contribution of the I-STUDIO assessment tool, the test blueprint, and the content of the instrument. The chapter also includes results of cognitive interviews and a summary of changes to the I-STUDIO instrument. The field test data are described and analyzed to estimate reliability, report validity evidence, and evaluate item response data.

## 4.2 Expert Reviewer Feedback

Expert feedback was solicited in order to evaluate the contribution of the I-STUDIO instrument, and critique both the test blueprint and draft assessment tool. This section summarizes the data from two iterations of feedback, and catalogues subsequent changes to the test blueprint and draft I-STUDIO assessment tool.

### 4.2.1 Contribution of the instrument.

The contribution of the instrument as a tool to inform gross curriculum and instruction outcomes was evaluated through mixed methods data analysis. The primary data were survey responses provided by subject matter experts upon reviewing the test blueprint. The test blueprint provided to the expert reviewers and the feedback questionnaire are available as appendices (Appendix A: Test Blueprint Prior to Expert Feedback; Appendix B: Expert Feedback Questionnaire Accompanying Test Blueprint). Specifically, questions 1, 2, and 16 of the feedback questionnaire prompted the most productive feedback relevant to the contributions of the I-STUDIO assessment.

Question 1 of the test blueprint questionnaire asked the following: “After completing an introductory statistics course, how important or unimportant is it that students be able to discern when a problem setting outside of class would or would not benefit from a statistical approach?” Reviewers were expected to provide a rating response and were invited to explain their answers. The distribution of rating endorsements among the 5 expert reviewers that responded is summarized in Table 6.

Table 6

*Distribution of expert feedback for questions 1 and 2 (blueprint questionnaire)*

| Frequency | Category |
|---|---|
| Question 1 | |
| 4 | Important |
| 1 | Somewhat Important |
| 0 | Somewhat Unimportant |
| 0 | Not Important |
| Question 2 | |
| 2 | Important |
| 2 | Somewhat Important |
| 1 | Somewhat Unimportant |
| 0 | Not Important |

All five reviewers that completed the survey agreed that it is “Important” or “somewhat important” that “students be able to discern when a problem setting outside of class would or would not benefit from a statistical approach.” Additionally, some reviewers claimed that this should be a main objective of the introductory statistics course even if the students are unable to do the analysis themselves (e.g. students of a statistics literacy course). Two reviewers commented on additional aspects related to discernment which include recognizing the need to collect data, as well as using data to

make decisions and test assumptions. One reviewer additionally cautioned the emphasis on “inference” in favor of a more general appeal to “statistical approach.”

Question 2 of the test blueprint questionnaire asked the following: “After completing an introductory statistics course, how important or unimportant is it that students be able to apply the statistical knowledge they have learned to novel problem settings outside of class?” Reviewers were expected to provide a rating response and were invited to explain their answers. The distribution of rating endorsements among the 5 expert reviewers that responded is summarized in Table 6.

Four reviewers that completed the survey agreed that it is “Important” or “Somewhat Important” that “students be able to apply the statistical knowledge they have learned to novel problem settings outside of class.” The two reviewers that endorsed this idea most strongly added comments to the effect that an introductory statistics course is of little value if students are only able to do well on exams or to reproduce the examples that they have seen upon completion.

The two reviewers that endorsed “Somewhat Important” for question 2 provided views that students need not necessarily be able to carry out the specific statistical methods and inferential procedures often discussed in the introductory curriculum. Rather, students should understand how to reasonably interpret graphs and summaries, use data in decision making, and distinguish between outcomes that are likely/unlikely to have occurred by chance. Similarly, the reviewer that endorsed “Somewhat Unimportant” explained that he felt students should understand the big concepts, yet need not necessarily be able to apply tools on their own to real-world problems.

Question 16 of the test blueprint questionnaire prompted reviewers to respond to the following: “Please share your overall evaluation of the test blueprint as well as any general comments that you have about this project.” No rating scale accompanied this prompt, so reviewers were simply invited to share their overall impressions about the test blueprint and the project. In response, three reviewers commented that creating and validating an assessment for the transfer of statistics knowledge is a worthwhile pursuit. Two reviewers cautioned against over-emphasis on inference and setting forth expectations that are too rigid.

### 4.2.2 Test blueprint.

The test blueprint provided to the expert reviewers and the feedback questionnaire are available as appendices (Appendix A: Test Blueprint Prior to Expert Feedback; Appendix B: Expert Feedback Questionnaire Accompanying Test Blueprint). Specifically, questions 3 through 15 of the test blueprint feedback questionnaire were most relevant to establish validity evidence supporting the I-STUDIO assessment. A summary of qualitative themes observed from the expert feedback and the resulting changes to the test blueprint follow.

#### *4.2.2.1 Summary of feedback.*

##### *4.2.2.1.1 Definitions described in the test blueprint*

Questions 3-7 of the test blueprint questionnaire asked a question about whether the definition was clear, and then reviewers were expected to choose “yes” or “no” and explain how the definition could be improved. Question 3 of the test blueprint questionnaire asked if the definition of “Forward-Reaching High Road Transfer” was

clear. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 7. All five reviewers agreed that the definition was clear, however, one reviewer recommended avoiding the use of jargon and another remarked that forward-reaching transfer seems to be an interesting exercise as a learning or instructional manipulation to facilitate future 'backward-reaching' transfer.

Table 7

*Distribution of expert feedback for questions 3-7 (blueprint questionnaire)*

| Frequency | Response |
|---|---|
| Question 3 | |
| 5 | Yes |
| 0 | No |
| Question 4 | |
| 5 | Yes |
| 0 | No |
| Question 5 | |
| 5 | Yes |
| 0 | No |
| Question 6 | |
| 5 | Yes |
| 0 | No |
| Question 7 | |
| 2 | Yes |
| 3 | No |

Question 4 of the test blueprint questionnaire asked if the definition of "Backward-Reaching High Road Transfer" was clear. The distribution of endorsements among the 5 expert reviewers that responded is again summarized in Table 7. All five reviewers agreed that the definition was clear; one reviewer remarked that the definition was not completely clear at first, but explained that the examples were helpful and asked whether "applying their knowledge" is the same as "demonstrating abstract principles."

Question 5 of the test blueprint questionnaire asked if the definition of “Discernment Required – Statistical Inference Appropriate” was clear. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 7. All five reviewers agreed that the definition was clear; one reviewer also asked whether students should be able to generate such situations in addition to recognizing them.

Question 6 of the test blueprint questionnaire asked if the definition of “Discernment Required – No Statistical Inference Required” was clear. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 7. All five reviewers agreed that the definition was clear. No additional comments were provided.

Question 7 of the test blueprint questionnaire asked if the definition of “No Discernment Required” was clear. The distribution of rating endorsements among the 5 expert reviewers that responded is summarized in Table 7. Two reviewers felt that the definition was clear, and three reviewers felt the definition was unclear.

One of the three reviewers that felt the definition was unclear suggested making it more explicit that students will be told to use inference. The second of the two reviewers that felt the definition was unclear explained that his understanding from the definition would suggest that irrelevant questions be included. The third explained confusion between ‘discernment required—no statistical inference’ and ‘no discernment required’ though he explained that he did eventually grasp the distinction after reviewing the example items.

One of the reviewers that felt the definition was clear as written remarked that the definition could be improved by noting that some problem types by their nature do not allow discernment, and other problem types (e.g., backward reaching) may or may not allow discernment depending on how they are composed. Furthermore, the reviewer explained that the definition remarks that these items contribute to measurement of high-road transfer only, but this should be more specific to include only backward-reaching high-road transfer.

*4.2.2.1.2 Item types described in the test blueprint*

Questions 8-13 of the test blueprint questionnaire prompted expert reviewers to study the description of each possible item type. Reviewers were expected to respond "yes" or "no" whether the description is clear as well as explain their choice. Similarly, reviewers were then expected to respond "yes" or "no" whether the item type seems important and then explain their choice.

Question 8 of the test blueprint questionnaire prompted reviewers to critique the description and importance of the "Forward-reaching high-road transfer with discernment—statistical inference appropriate" item type. The distribution of rating endorsements among the 5 expert reviewers that responded is summarized in Table 8. All five reviewers agreed that the description is clear.

One reviewer additionally commented that the examples were very helpful. Three of the five reviewers felt that the "Forward-reaching high-road transfer with discernment—statistical inference appropriate" item type is important, though one remarked that backward-reaching transfer items was more important by comparison since they are more

consistent with real-world problems that may be faced outside of class. The fourth reviewer, who marked this item type as unimportant, explained that it would be important for a statistical methods course but not for a statistical literacy course. The fifth reviewer expressed uncertainty about the concept of forward-reaching transfer as a tool for assessing what is learned in an introductory statistics course. During a follow-up meeting with the fifth reviewer, we agreed that forward-reaching transfer could be a useful outcome, but may be challenging to measure because many students may simply leave the item blank if they aren't able to come up with a novel response.

Table 8

*Distribution of expert feedback for questions 8-13 (blueprint questionnaire)*

| Frequency | Response |
|---|---|
| Question 8 | |
| Clear Description? | |
| Yes | 5 |
| No | 0 |
| Important? | |
| Yes | 3 |
| No | 2 |
| Question 9 | |
| Clear Description? | |
| Yes | 3 |
| No | 2 |
| Important? | |
| Yes | 3 |
| No | 0 |
| Question 10 | |
| Clear Description? | |
| Yes | 2 |
| No | 2 |
| Important? | |
| Yes | n/a |
| No | n/a |
| Question 11 | |

| Frequency | Response |
|---|---|
| Clear Description? | |
| Yes | 5 |
| No | 0 |
| Important? | |
| Yes | 5 |
| No | 0 |
| Question 12 | |
| Clear Description? | |
| Yes | 5 |
| No | 0 |
| Important? | |
| Yes | 4 |
| No | 0 |
| Question 13 | |
| Clear Description? | |
| Yes | 2 |
| No | 1 |
| Important? | |
| Yes | 2 |
| No | 2 |

Question 9 of the test blueprint questionnaire prompted reviewers to critique the description and importance of the “Forward-reaching high-road transfer with discernment—no statistical inference required” item type. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 8. Three out of five reviewers agreed that the description is clear. The first reviewer that felt the description was unclear explained that it would be important to explore issues beyond census data. The second reviewer that felt the description was unclear provided feedback as such in the survey, then explained in a comment on his returned copy of the blueprint that he did understand, but would like to see a few more example items.

Three out of five reviewers agreed that “Forward-reaching high-road transfer with discernment—no statistical inference required” is an important item type, and the fourth said that he had “no opinion” and did not respond, neither did the fifth. One reviewer remarked that this is an important type of item, but the instrument should focus on the questions where inference is appropriate so the stated balance of twice as many items with inference appropriate is good. Another reviewer suggested that this be reworded to “not benefiting from statistical approach” especially when inferential methods can be done, but are just not useful or efficient.

Question 10 of the test blueprint questionnaire prompted reviewers to critique the description and importance of the “Forward-reaching high-road transfer with no discernment required” item type. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 8. Two out of five reviewers agreed that the description is clear, and one remarked that it seems reasonable to exclude items of this type. A third felt the description is not clear, and the fourth said that he had “no opinion” and did not respond. The reviewer that felt the description is unclear was confused by what was meant by “discernment” since it seemed like the blueprint used discernment to mean students choose between inference and no inference, but it also seemed that the blueprint may use the term to include discernment of application. The fifth reviewer did not respond to the item.

Question 11 of the test blueprint questionnaire prompted reviewers to critique the description and importance of the “Backward-reaching high-road transfer with discernment—statistical inference appropriate” item type. The distribution of

endorsements among the 5 expert reviewers that responded is summarized in Table 8. All reviewers consulted claimed that the description is clear and this item type is important. None of the reviewers shared additional comments expanding on their endorsement.

Question 12 of the test blueprint questionnaire prompted reviewers to critique the description and importance of the "Backward-reaching high-road transfer with discernment—no statistical inference required" item type. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 8. All reviewers consulted claimed that the description is clear and this item type is important. One reviewer did not select "yes" on the questionnaire although his intent was clear from written comments provided. Another reviewer added that it will be important to make sure the scenarios are not too artificial and make use of more than just the census issue.

Question 13 of the test blueprint questionnaire prompted reviewers to critique the description and importance of the "Backward reaching high-road transfer with no discernment" item type. The distribution of endorsements among the 5 expert reviewers that responded is summarized in Table 8. Two out of five reviewers agreed that the description was clear, a third felt the description is not clear, and the fourth had "no opinion" and did not respond, neither did the fifth. The reviewer that felt the description is unclear wanted to know what criteria will be used to decide whether a response is a "viable research question" and the prompt should encourage students to provide more detail than "I will find a p-value."

Two out of five reviewers agreed that the "Backward reaching high-road transfer with no discernment" item type is important, a third felt the item type is not important, and the

fourth said that he had “no opinion” and did not respond. The reviewer that rated this item type as ‘not important’ remarked that it was rated as such by comparison to the importance of discernment tasks, however, the reviewer understood the desire to include items of this type in order to isolate students’ ability to transfer their statistical knowledge when told to do so. One of the reviewers that marked this item type as important added the opinion that “being able to write a testable research question is more important than describing the analysis.”

Question 14 of the test blueprint questionnaire prompts the following: “Do the item examples generally seem to align well with the definitions, descriptions, and intended learning outcomes? Please explain by referencing specific examples.” No rating scale accompanied this prompt, so reviewers were simply invited to share free-form feedback. In response, two reviewers commented on example item 4 (display screen inspection), suggesting that students may be tempted to use a sampling approach to decide whether to accept or reject the display screens. They recommend emphasizing that the engineer will conduct testing on all 50 displays and changing the question text to something like “should statistical inference be used.”

Another reviewer commented that it’s difficult to evaluate the instrument without defining what we intend by ‘introductory statistics course.’ The reviewer suggested that courses with no forward-reaching transfer may have their place as a literacy curriculum and may become more common as Big Data gains momentum, and it’s not clear that meaningful evidence of forward-reading transfer is expected among students taking their first course of many. The reviewer further suggested that backward-reaching transfer

seems to be the critical component of a literacy course. With respect to the vernacular used the reviewer added that the term “backward” could be interpreted to have negative connotation, and “discernment” may be a loaded term.

The fourth reviewer remarked that the instrument focuses on whether students can use what they learned in class in new situations, but that it is also important to include new structures and frameworks as well. Perhaps a class didn’t cover comparing multiple means (e.g. ANOVA); it may be valuable to study whether students are able to transfer what they have learned to this new situation.

Question 15 of the test blueprint questionnaire prompts the following: “Do you feel that anything is incomplete or missing from the test blueprint?” No rating scale accompanied this prompt, so reviewers were simply invited to share free-form feedback. In response, one reviewer explained that the rubrics are critical, and it is not yet clear how details like “whether or not they are sufficiently acknowledging randomness” will be decided. Also, it may be important to include topics beyond just inference and make sure that there are good “non-inference” items. Another reviewer commented that Example 5 was more of a question template and suggested that a specific example would have been nice there.

#### *4.2.2.2 Summary of changes to the test blueprint.*

In light of the expert feedback summarized in Section 4.2.2.1, a number of changes to the I-STUDIO test blueprint were warranted. The draft blueprint presented to the expert reviewers for feedback is shown in Appendix A: Test Blueprint Prior to Expert Feedback, and the improved blueprint showing changes in response to their feedback is available in

Appendix C: Final Test Blueprint. Changes to the test blueprint included modification of terms and definitions as well as organization of content.

The most substantial change to the content of the test blueprint was a revision to the definition of "No Discernment Required." The definition was completely rewritten to emphasize that students must still demonstrate problem-solving skills indicative of high-road transfer, although the task would preclude a discernment task by explicitly dictating whether or not statistical inference should be used. Another minor update to the terminology throughout the test blueprint changed all references of "Discernment Required—No Statistical Inference Required" to "Discernment Required—Statistical Inference Not Appropriate." Lastly, the organization of the test blueprint was modified to describe item types associated with backward-reaching transfer before item types associated with forward-reaching transfer.

### 4.2.3 Draft I-STUDIO assessment tool.

The draft assessment tool provided to the expert reviewers and the feedback questionnaire are available in an appendix (Appendix D: Draft I-STUDIO Version Prior to Expert Feedback; Appendix E: Expert Feedback Questionnaire Accompanying Draft I-STUDIO Assessment Tool). Specifically, questions 2 through 12 of the feedback questionnaire are most relevant to establish validity evidence supporting the I-STUDIO assessment. Note that the order of the assessment tasks in the draft assessment scrutinized by the reviewers differs from the order of tasks used in the final version of the I-STUDIO assessment used for the large-scale field test. A summary of qualitative themes observed from the expert feedback and the resulting changes to the instrument follow.

#### *4.2.3.1 Summary of feedback.*

Each of questions 2-8 of the draft assessment questionnaire asked the reviewers to critique a specific I-STUDIO item, and remark whether the item aligned with specific characteristics described in the test blueprint. Reviewers were expected to choose "yes" or "no" and then explain how the item could be improved.

Question 2 of the draft assessment questionnaire asked whether item 1 (Walleye fishermen) aligned with the characteristics described in the test blueprint for "Backward-reaching high-road transfer with discernment—statistical inference appropriate." The distribution of rating endorsements among the 6 expert reviewers that responded is summarized in Table 9. All six reviewers that completed the survey agreed that item 1 (Walleye fishermen) aligned with the characteristics described in the test blueprint for backward-reaching high-road transfer with discernment—statistical inference appropriate. Two reviewers commented that they "really like" the item, and other reviewers recommended improvements that describe how the data were collected. The type of issues that reviewers mentioned reveal a great deal of statistics knowledge successfully transferred to the context, which is the goal. Other comments generally critiqued the example solutions provided to represent minimally acceptable student responses. These comments can be addressed with the rubric.

Table 9

*Distribution of expert feedback for questions 2-10 (draft assessment questionnaire)*

| Response | Frequency |
|---|---|
| Question 2 | |
| Yes | 6 |

| Response | Frequency |
|---|---|
| No | 0 |
| Question 3 | |
| Yes | 4 |
| No | 2 |
| Question 4 | |
| Yes | 6 |
| No | 0 |
| Question 5 | |
| Yes | 6 |
| No | 0 |
| Question 6 | |
| Yes | 5 |
| No | 0 |
| Question 7 | |
| Yes | 5 |
| No | 1 |

Question 3 of the draft assessment questionnaire asked whether item 2 (note identification) aligned with the characteristics described in the test blueprint for "Backward-reaching high-road transfer with discernment—statistical inference appropriate." The distribution of endorsements among the 6 expert reviewers that responded is summarized in Table 9. Four reviewers felt that item 2 (note identification) aligns with the characteristics described in the test blueprint for backward-reaching high-road transfer with discernment—statistical inference appropriate, while two reviewers said it does not. Two reviewers commented that the question implies that a determination be made based on a single note, and suggested clarification that the test be repeated.

Question 4 of the draft assessment questionnaire asked whether item 3 (display screen inspection) aligned with the characteristics described in the test blueprint for "Backward-reaching high-road transfer with discernment—statistical inference NOT appropriate."

The distribution of endorsements among the 6 expert reviewers that responded is summarized in Table 9. All six reviewers that completed the survey agreed that item 3 (display screen inspection) aligned with the characteristics described in the test blueprint for backward-reaching high-road transfer with discernment—statistical inference not appropriate.

Question 5 of the draft assessment questionnaire asked whether item 4 (air traffic control) aligned with the characteristics described in the test blueprint for "Backward reaching high-road transfer with no discernment." The distribution of endorsements among the 6 expert reviewers that responded is summarized in Table 9. All six reviewers that completed the survey agreed that item 4 (air traffic control) aligned with the characteristics described in the test blueprint for backward reaching high-road transfer with no discernment. Two reviewers commented that the question should clarify what the pretest measures. Other comments generally critiqued the example solutions provided to represent minimally acceptable student responses.

Question 6 of the draft assessment questionnaire asked whether item 5 (underlying principle of inference) aligned with the characteristics described in the test blueprint for "Forward-reaching high-road transfer with discernment—statistical inference appropriate." The distribution of rating endorsements among the 6 expert reviewers that responded is summarized in Table 9. All six reviewers that completed the survey agreed that item 5 (underlying principle of inference) aligns with the characteristics described in the test blueprint for forward-reaching high-road transfer with discernment—statistical

inference appropriate. Comments generally critiqued the example solutions provided to represent minimally acceptable student responses.

Question 7 of the draft assessment questionnaire asked whether item 6 (inference not appropriate) aligned with the characteristics described in the test blueprint for "Forward-reaching high-road transfer with discernment—statistical inference NOT appropriate." The distribution of rating endorsements among the 6 expert reviewers that responded is summarized in Table 9. Five out of six reviewers that completed the survey agreed that item 6 (inference not appropriate) aligns with the characteristics described in the test blueprint for forward-reaching high-road transfer with discernment—statistical inference NOT appropriate. The sixth reviewer did not respond to the yes/no portion, and shared an article published by Freedman and Lane (1983) in which the authors argue that significance testing and confidence intervals are appropriate even when the sample is equated to represent the entire population of interest. After reviewing the article, the authors propose an interpretation that characterizes "significance level [as] a descriptive statistic rather than a probability" (p. 293). Therefore, it seems fair to conclude that although the authors argue utility of p-values for non-stochastic processes, they did not advocate for inference in these cases. Most comments generally critiqued the example solution provided to represent minimally acceptable student responses.

Question 8 of the draft assessment questionnaire asked whether item 7 (matched pairs study design) aligned with the characteristics described in the test blueprint for "Forward-reaching high-road transfer with discernment—statistical inference appropriate." The distribution of rating endorsements among the 6 expert reviewers that responded is

summarized in Table 9. Five out of six reviewers that completed the survey agreed that item 7 aligns with the characteristics described in the test blueprint for forward-reaching high-road transfer with discernment—statistical inference appropriate. The sixth reviewer did not respond to the yes/no portion. One reviewer suggested a clarification to preserve the definition of the matched pairs design as a randomized complete block design with 2 experimental units per block. Other comments generally critiqued the example solution provided to represent minimally acceptable student responses.

Questions 9 and 10 each asked the reviewers a question about whether the I-STUDIO assessment tool measures a target construct. Reviewers were expected to rate their agreement on a four-point scale of Agree, Somewhat Agree, Somewhat Disagree, Disagree, and then explain their answer.

Question 9 of the draft assessment asked the following: “Think about a student that has completed an introductory course in statistical methods, to what extent do you agree or disagree that the I-STUDIO assessment measures whether students would be able to discern whether statistical inference is appropriate for problem settings outside of class?” The distribution of rating endorsements among the 6 expert reviewers that responded is summarized in Table 10. All six reviewers that completed the survey said that they “agree” or “somewhat agree” that the I-STUDIO assessment measures whether students would be able to discern whether statistical inference is appropriate for problem settings outside of class. Three reviewers alluded to comments that they made about previous items and said that the instrument would be improved if those are corrected. Other

comments and concerns discussed here seem to challenge the boundaries of the scope intended for this instrument.

Table 10

*Distribution of expert feedback for questions 9 and 10 (draft assessment questionnaire)*

| Response | Frequency |
|---|---|
| Question 9 | |
| Agree | 2 |
| Somewhat Agree | 4 |
| Somewhat Disagree | 0 |
| Disagree | 0 |
| Question 10 | |
| Agree | 4 |
| Somewhat Agree | 2 |
| Somewhat Disagree | 0 |
| Disagree | 0 |

Question 10 of the draft assessment asked the following: “Think about a student that has completed an introductory course in statistical methods, to what extent do you agree or disagree that the I-STUDIO assessment measures whether students would be able to demonstrate high-road transfer in novel problem settings outside of class?” The distribution of rating endorsements among the 5 expert reviewers that responded is summarized in Table 10. All six reviewers that completed the survey said that they “agree” or “somewhat agree” that the I-STUDIO assessment measures whether students would be able to discern whether statistical inference is appropriate for problem settings outside of class. One reviewer explained a desire to see students actually conduct the analysis to complete the transfer, but said that the assessment is well done for its purposes. Another reviewer stated that the “assessment items ask students to complete

tasks that involve the skills very highly overlapping with ‘demonstrating high-road transfer’ and occur in novel problem settings” which alludes to the definition provided in the test blueprint.

Question 11 of the draft assessment questionnaire requested the following: “Please share anything you feel is missing from the I-STUDIO assessment.” No rating scale accompanied this prompt, so reviewers were simply invited to share free-form feedback. Only one out of the six reviewers who completed the survey chose to comment here. The reviewer felt that the instrument is missing “the clearly badly collected data situations and realize they shouldn’t use inference when they don’t have randomness.”

Question 12 of the draft assessment questionnaire requested the following: “Please share any general comments that you have about this project.” No rating scale accompanied this prompt, so reviewers were simply invited to share their overall impressions about the test blueprint and the project. In response, five out of six reviewers that completed the survey volunteered a comment here. One comment touched on sufficiency of informal inference, and another reiterated that some items need to pay more attention to data collection issues. A third comment remarked that the instrument may “penalize students who know too much, and reward students who know just enough.” The same commenter also asked whether it is fair to remove mathematics from the assessment which could put non-native English speakers at a disadvantage. Lastly, another reviewer remarked that he found it interesting that the assessment is “directing transfer, and not depending on spontaneous transfer.”

#### *4.2.3.2 Summary of changes to the instrument.*

In light of the expert feedback summarized in Section 4.2.3.1, a number of changes to the I-STUDIO assessment tool were warranted. The draft instrument presented to the expert reviewers for feedback is shown in Appendix D: Draft I-STUDIO Version Prior to Expert Feedback, and the improved instrument showing changes in response to their feedback is available in Appendix F: I-STUDIO Version for Cognitive Interviews. Changes to the instrument spanned the entire assessment tool including the consent form, instructions, vignettes, and item prompts.

Within the consent form, the order of the "confidentiality" and "risks" sections were switched in order to present the confidentiality section first since the primary risk of participating in the study is a breach of confidentiality. In the instructions to the student at the beginning of the instrument, a clause was added to be clear that statistical inference is not appropriate for some of the questions in the assessment.

The vignette accompanying item 1 (Walleye fishing) was modified to place the brothers on an extended fishing trip together in order to resolve several data collection issues identified by reviewers. The resulting description provides each brother with comparable equipment and resources so they can fish independently. The final text was chosen carefully to provide enough detail to support the study design without using obvious terminology that would undermine the discernment task in part A.

The vignette accompanying item 2 (note identification) was modified to change several references to “note identification test” to “method of note identification” in order to mitigate interpretations that the test consists of exactly one note. Students are expected

to recognize the need for data collection by repeating the test many times, so the item should not lead them to believe that the test includes only one note, but the resulting item avoids obvious terminology that would undermine the discernment task in part A.

Minor modifications were made to items 3, 4, and 5. For item 3 (display screen inspection), the company names were made to were made to be generic in order to avoid using a name too similar to an active company of any kind. Text was added in item 4 (air traffic control) to clarify what was measured in the pretest. Item 5 (underlying principle of inference) had previously specified that the student write one or two paragraphs, but that guidance was removed. There were no noteworthy changes applied to item 6, and only a minor change was made to item 7 (matched pairs study design) which added contexts of medicine & psychology to further ground terminology such as "participant" and "treatment" for the reader. Language was chosen such that acceptable response in an unrelated context would certainly not be penalized.

### 4.3 Student Cognitive Interviews

The draft assessment tool provided to the cognitive interview participants is available in Appendix F: I-STUDIO Version for Cognitive Interviews. Note that the order of the assessment tasks in the draft assessment cognitive interviews and the large-scale field test differs from the order of tasks previously scrutinized by the expert reviewers. Cognitive interviews were conducted about 6 weeks after fall semester had ended, and none of the students brought notes or other resources with them to the interview. A summary of qualitative themes observed during cognitive interviews and the resulting changes to the instrument follow.

### 4.3.1 Summary of feedback.

Students did not share any questions or comments on the consent form. Three out of five students skipped the directions. No students had questions or comments about the directions. In item 1 (air traffic control), some students apparently did not recognize the ellipses to indicate that this is simply an excerpt of the data; they described specific cases in the data set. Most students seemed to follow the question well enough to propose reasonable research questions and solution strategy; no major revisions to item were necessary. One student remarked that it felt intimidating to propose a detailed strategy on the spot as opposed to an environment of a take-home exam or homework which would be more similar to the context in which he had done things like this in the past.

For item 2 (note identification), several students assumed that many participants would be involved in the note identification test, or they became distracted by generalizing to a population of students. In item 3 (display screen inspection), most students described reasons why statistical inference would not be required in this scenario, but sometimes remarked that they felt like they were tempted to overthink things. With item 4 (Walleye fishing), most students seemed to follow the question well enough and propose reasonable solution strategies. No major revisions to the item seemed necessary although some students struggled to incorporate both the length and the weight data in order to determine which brother catches larger fish on average.

Item 5 (matched-pairs study design) was a difficult item for several of the students. Several students had false starts, and decided to reread the prompt and start over. Some students described a comparison for two independent samples. In item 6 (underlying

principle of inference), several students began by describing a hypothetical sample, extended the sample to some plausible population, summarized the appropriate statistic, and then determined the corresponding parameter without having first described a realistic scenario. Overall, item 6 functioned quite well, but several students never actually typed a description of their scenario into their response when incrementally building up to their response as described. Lastly, four out of five students reread the stem of item 7 (statistical inference NOT appropriate) at least once, but they generally seemed to think about the task appropriately and arrived at a reasonable solution.

Upon completion of the instrument, two out of five students reviewed their solutions. The total time in minutes taken for each student to complete the assessment was 26, 45, 50, 55, and 63. One student finished the assessment much more quickly than the other four, but did not appear quite as invested in her responses by comparison to the others. Another one of the five students began to describe test fatigue at the end of the assessment; that student spent a total of 50 minutes to complete the assessment tool.

### 4.3.2 Summary of changes to the instrument.

In light of the student feedback summarized in Section 4.3.1, a number of changes to the I-STUDIO assessment tool were warranted. The draft instrument presented during the cognitive interviews is shown in Appendix F: I-STUDIO Version for Cognitive Interviews, and the improved instrument showing changes in response to their feedback is available in Appendix G: I-STUDIO Version for Field Test. Changes to the instrument spanned the entire assessment tool including the instructions, vignettes, and item prompts.

Since several students skipped the directions in the cognitive interviews, the directions were modified prior to the field test such that each of the key instructions were listed and accompanied by checkboxes that the student must acknowledge before advancing to the first item. For item 1 (air traffic control), the ATC preparation data table header was modified to state "Example ATC Preparation Data (Showing five of the nineteen students)" in order to underscore that the data shown are only an excerpt. Item 2 (note identification) was modified to specify an individual student by name in order to make clear that the problem setting is concerned with evaluating the results for only one student.

The prompt for item 3 (display screen inspection), part b was improved to clarify that the decision to accept or reject the bulk order is based on the data gathered by the engineer in order to suggest that the student need not invent a new method in order to invoke statistical inference. A modification was made to the item 4 (Walleye fishing) vignette to state that only the length of each fish was recorded so students would not complicate their analysis proposal by incorporating more than one measurement of fish size. Part b of item 4 (Walleye fishing) was also updated to clarify that the comparison is based on the data collected on the two week fishing trip.

No noteworthy changes were made to the content of item 5 (matched pairs study design). Item 6 (underlying principle of inference) was modified to add an additional bullet point prompting students to "briefly describe your chosen scenario and state the question of interest you would explore using statistical inference in that scenario" since that information was omitted from several of the responses submitted by cognitive

interviews participants. The prompt for item 7 (inference is not required) was simplified by removing comments about deterministic results and cleaning up lengthy sentences. Also, the prompt for item 7 was modified to show bullet points listing each component the response should address as paralleled by item 6.

## 4.4 Field Test Data Analysis

Rubric development, expert feedback, and inter-rater reliability were evaluated once the student response data was obtained from the field test. With the rubric in place and 178 student responses scored, analysis of student data followed. Data analysis included descriptive statistics of total scores, and scoring reliability statistics were calculated. Confirmatory factor analysis models were fitted to the data in order to evaluate dimensionality of latent characteristics measured by I-STUDIO, followed by appropriate item response modeling and qualitative analysis of student responses.

### 4.4.1 Scoring rubric.

The draft scoring rubric tool provided to the Statistics Education PhD candidate is available in Appendix H: I-STUDIO Draft Scoring Rubric. A summary of qualitative themes observed and the resulting changes to the rubric follow. The final version of the rubric used to score the field test data is available in Appendix I: I-STUDIO Final Scoring Rubric for Field Test.

#### *4.4.1.1 Summary of feedback.*

Comments related to the rubric for item 1 (ATC preparation data) included a recommendation to clarify whether students needed to “name” a specific method of statistical inference, as well as a remark that the criteria for demonstrating sufficient

understanding appeared unclear. The reviewer remarked for item 2a, that it is possible a student may not explicitly declare a “yes” or “no” position, and that such a response would not be accommodated by the rubric. For item 2b, the reviewer commented that there are relatively few appropriate inferential methods and recommended that they be explicitly listed in the rubric to drive consistent use.

Feedback on the item 3a rubric suggested that the partial credit (P) criteria be simplified for easier use. In item 3b, feedback suggested that the rubric should address the possibility of a student advocating for statistical inference in 3a and then recommending a reasonable inferential strategy in 3b (e.g. one proportion z-test/confidence interval). In item 6, the reviewer asked for clarification in the rubric describing whether and how to score inferred sample bias. There were no major comments recommending changes to the rubric for item 4, 5, or 7.

#### *4.4.1.2 Summary of changes to the scoring rubric.*

The item 1 (air traffic control) rubric was updated to require that the student name the specific statistical method of choice in their response for part c. This comment was carried forward to items 2b and 4b. Also, the criterion for an essentially correct response (E) was re-written to explain that the response should not indicate a flawed understanding of the chosen method. In other words, students were not necessarily expected to explain the method in detail, but were penalized if they volunteered incorrect understanding of the chosen method. For item 2 (note identification), the rubric accompanying part 2a was updated to award partial credit (P) to a student that does not clearly declare a “yes” or “no” position, but provides an otherwise satisfactory response. Item 2b was modified to

elaborate on specific methods that the student should name or paraphrase in order to earn full credit for an essentially correct (E) response. Furthermore, a note was included specifying how a student may earn an E for item 2b regardless of their score on 2a.

The partial credit (P) logic for part 3a of the item 3 (display screen inspection) rubric was reordered to clarify expectations, but the content remained unchanged. Item 3b was modified to include a note and grant partial credit (P) for a response that recommends a reasonable inferential strategy given that the student advocated for statistical inference in 3a. The rubric for item 4 was updated to accommodate broader characterization of fish size (i.e. size, length, weight) among essentially correct (E) responses.

The rubric for item 5 (matched pairs study design) was largely unchanged with the exception of a note added describing the treatment of responses that clearly label each intended element. In short, labels can be implied if the response clearly demonstrates understanding (e.g. a bullet list that corresponds to the order of components requested), but if labels are made explicit the response should be scored accordingly. This comment was applied to items 6 and 7 as well.

A note was added to the rubric for item 6 explaining that the penalty for a biased sample should only be applied when the response explicitly describes a sampling method that introduces bias. Lastly, no significant changes were made to the rubric for item 7 with the exception of the note regarding the use of labels in the response.

#### *4.4.1.3 Consistency of rubric application.*

Rubric use was evaluated for both inter-rater agreement, and intra-rater agreement. The inter-rater agreement was based on evaluation of 5 students across 31 individual

scores per student. The two independent reviewers agreed on 140 of 155 individual scores to produce inter-rater agreement on 90.3% of scoring decisions. Among the 15 scoring conflicts, 9 were scored higher by the author and 6 were scored higher by the Statistics Education PhD candidate. No score discrepancies were more than 1 point. That is no element was scored essentially correct (E) by one rater and incorrect (I) by the other. Table 11 shows the proportion of scores agreed upon by both raters. One scoring element (item 3a) had 40% (2/5) agreement between raters, and a total of three scoring elements associated with items 5 and 7 each had 60% (3/5) agreement between raters. All other scoring elements had at least 80% agreement between raters, including 21 scoring elements with perfect agreement.

Table 11

*Inter-rater agreement by scoring element*

| Scoring Element | Rater Agreement (with 5 responses) |
|---|---|
| Item 3a | 40% |
| Item 5 (pairing) | 60% |
| Item 5 (treatments) | 60% |
| Item 7 (data) | 60% |
| Item 1b (redundancy-penalty) | 80% |
| Item 1b | 80% |
| Item 3b | 80% |
| Item 4a | 80% |
| Item 5 (response) | 80% |
| Item 6 (sample) | 80% |
| Item 1a | 100% |
| Item 1c | 100% |
| Item 2a | 100% |
| Item 2b | 100% |
| Item 4b | 100% |
| Item 5 (analysis) | 100% |
| Item 5 (interpretation) | 100% |
| Item 5 (lacks replication-penalty) | 100% |
| Item 5 (participants) | 100% |

| Scoring Element | Rater Agreement (with 5 responses) |
|---|---|
| Item 5 (scenario) | 100% |
| Item 6 (biased sample-penalty) | 100% |
| Item 6 (parameter) | 100% |
| Item 6 (population) | 100% |
| Item 6 (question) | 100% |
| Item 6 (scenario) | 100% |
| Item 6 (statistic) | 100% |
| Item 7 (analysis) | 100% |
| Item 7 (parameter) | 100% |
| Item 7 (population) | 100% |
| Item 7 (research question) | 100% |
| Item 7 (scenario) | 100% |

The intra-rater agreement was based on evaluation of 10 students across 31 individual scores per student. Second attempt scoring of the ten responses agreed with the original decision for 306/310 individual scoring decisions to produce intra-rater agreement of 98.7%. Among the 4 scoring conflicts, 2 were scored higher on the first scoring attempt and 2 were scored higher on the second scoring attempt. No score discrepancies were more than 1 point. That is no element was scored essentially correct (E) for one scoring attempt and incorrect (I) for the other scoring attempt. One scoring element (item 1a) had 80% (8/10) scoring agreement, a total of two scoring elements associated with items 4a and item 5 (interpretation) each had 90% (9/10) agreement, and all other elements had 100% (10/10) intra-rater agreement.

### 4.4.2 Descriptive statistics.

As shown in *Figure 2*, the distribution of total scores among 178 randomly selected student responses evaluated appears unimodal and somewhat positively skewed. The mean and median scores were 16.07 and 16 points, respectively, out of a total of 37 possible points. The standard deviation and interquartile range of the I-STUDIO total

scores were 7.05 points and 10 points, respectively. Summary statistics by item and testlet shown in Table 12 include the total points possible as well as the associated mean and standard deviation.

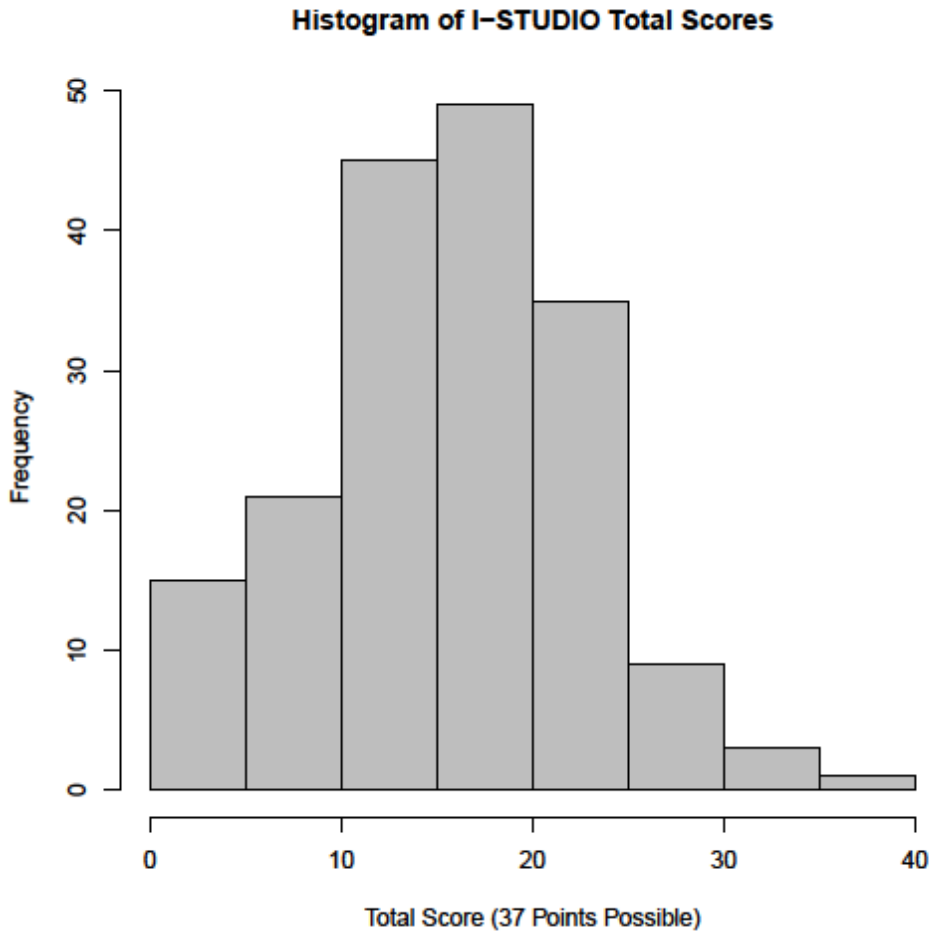


*Figure 2*. Histogram of I-STUDIO total scores.

Table 12

*I-STUDIO summary statistics by item and testlet.*

| Scoring element | Possible Points | Mean (SD) | Mean Percent Points Earned |
|---|---|---|---|
| Testlet 1 subtotal | 6 | 2.79 (1.61) | 47% |
| Item 1a | 2 | 1.29 (0.72) | 65% |
| Item 1b | 2 | 1.12 (0.82) | 56% |
| Item 1c | 2 | 0.38 (0.69) | 19% |
| Testlet 2 subtotal | 4 | 1.63 (1.15) | 41% |
| Item 2a | 2 | 0.77 (0.84) | 38% |
| Item 2b | 2 | 0.87 (0.55) | 43% |
| Testlet 3 subtotal | 4 | 1.20 (1.39) | 30% |
| Item 3a | 2 | 0.39 (0.71) | 20% |
| Item 3b | 2 | 0.80 (0.94) | 40% |
| Testlet 4 subtotal | 4 | 1.52 (1.32) | 38% |
| Item 4a | 2 | 0.76 (0.79) | 38% |
| Item 4b | 2 | 0.76 (0.71) | 38% |
| Testlet 5 subtotal | 7 | 3.04 (2.15) | 43% |
| Item 5 context | 2 | 1.01 (0.75) | 51% |
| Item 5 component | 5 | 2.03 (1.54) | 41% |
| Testlet 6 subtotal | 6 | 3.26 (1.68) | 54% |
| Item 6 context | 2 | 1.75 (0.61) | 88% |
| Item 6 component | 4 | 1.51 (1.34) | 38% |
| Testlet 7 subtotal | 6 | 2.62 (2.05) | 44% |
| Item 7 context | 2 | 1.25 (0.75) | 63% |
| Item 7 component | 4 | 1.37 (1.43) | 34% |
| I-STUDIO Total | 37 | 16.07 (7.05) | 43% |

Students earned the lowest percentage of possible points on item 1c and item 3a, and the highest on the context portion of item 6 and research question proposal in item 1a. Testlet scores ranged from 30% of points earned for testlet 3 (display screen inspection) to 54% of points earned for item 6 (underlying principle of inference).

Total scores for each student organized by course and accompanied by 95% confidence intervals for the mean are shown in *Figure 3* and in Table 13. In *Figure 3,* several pairwise differences among courses are evident. For example, course 9 averaged a

higher total score than most of the others, while course 2 averaged a lower total score than several others. Only two students from course 12 completed the I-STUDIO assessment. Both students were included in the data analysis, but the standard deviation and the associated confidence interval were not informative.

Table 13

*Summary statistics of I-STUDIO scores*

| Course ID | N | Mean (SD) | 95% CI |
|---|---|---|---|
| 1 | 12 | 16.9 (6.9) | [12.5, 21.3] |
| 2 | 12 | 8.4 (5.5) | [4.9, 11.9] |
| 3 | 8 | 20.0 (4.6) | [16.1, 23.9] |
| 4 | 12 | 17.2 (3.3) | [15.1, 19.3] |
| 5 | 12 | 15.2 (5.9) | [11.5, 19.0] |
| 6 | 12 | 18.6 (5.5) | [15.1, 22.0] |
| 7 | 12 | 16.1 (6.8) | [11.8, 20.4] |
| 8 | 12 | 17.3 (6.7) | [13.1, 21.6] |
| 9 | 12 | 27.8 (6.0) | [24.1, 31.6] |
| 10 | 12 | 11.8 (5.8) | [8.1, 15.4] |
| 11 | 12 | 11.4 (4.9) | [8.3, 14.5] |
| 12 | 2 | 19.5 | |
| 13 | 12 | 11.9 (5.2) | [8.6, 15.2] |
| 14 | 12 | 18.1 (6.4) | [14.0, 22.2] |
| 15 | 12 | 17.4 (7.0) | [13.0, 21.8] |
| 16 | 12 | 13.6 (3.8) | [11.2, 16.0] |
| Total | 178 | 16.1 (7.0) | [15.0, 17.1] |

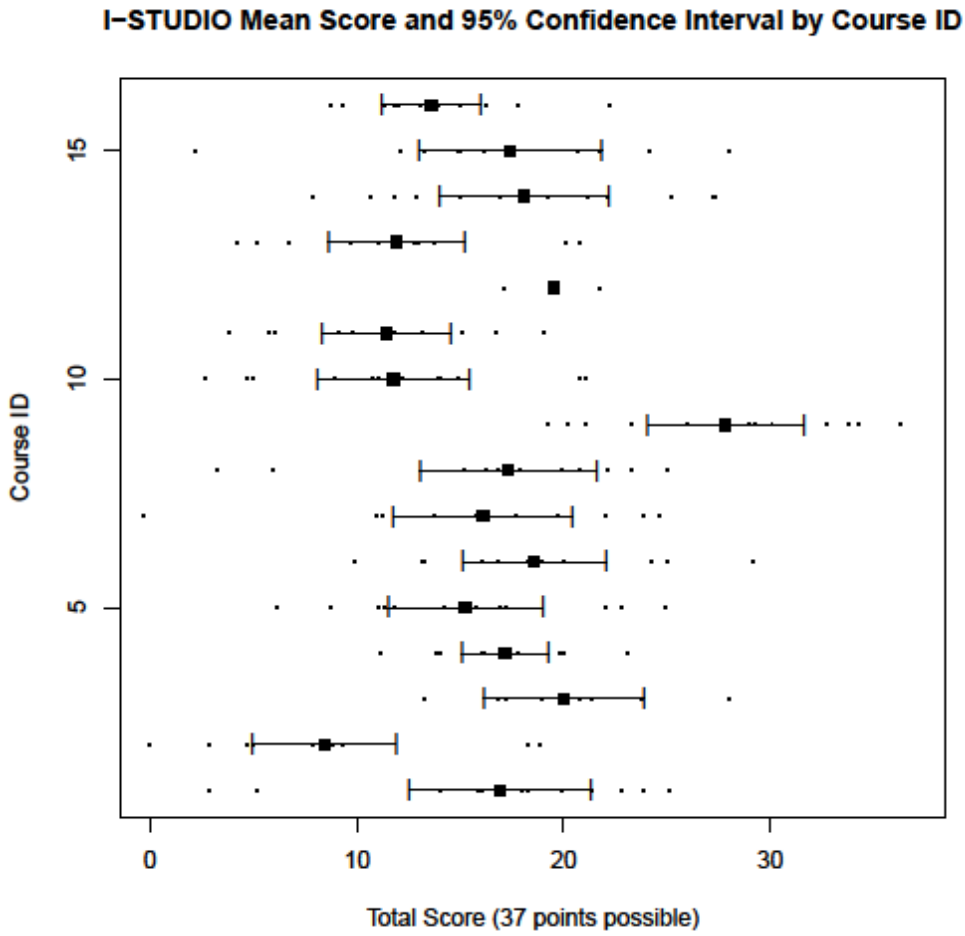


*Figure 3*. Mean scores with 95% confidence intervals by course ID.

### 4.4.3 Reliability.

A distribution of 500,000 simulated reliability estimates using the Spearman-Brown formula for split-half reliability is shown in *Figure 4*. The mean and median of the distribution of simulated reliability estimates are both 0.74 and a 95% confidence interval for the Spearman-Brown reliability is [0.66, 0.82]. Since the simulated reliability calculation only included 12 of the 15 items for each iteration, the projected median reliability of the 15 item instrument is estimated as $(15r_{tt})/(1+(14)\ r_{tt})= 0.78$ and the

transformed 95% confidence interval became [0.71, 0.85]. The standard error of measurement calculated from the same 500,000 simulations has a mean and median of 3.56 points with a 95% confidence interval of [3.0, 4.1]. The standard error of measurement estimated from the median projected reliability of the 15 item instrument is 3.30 points.

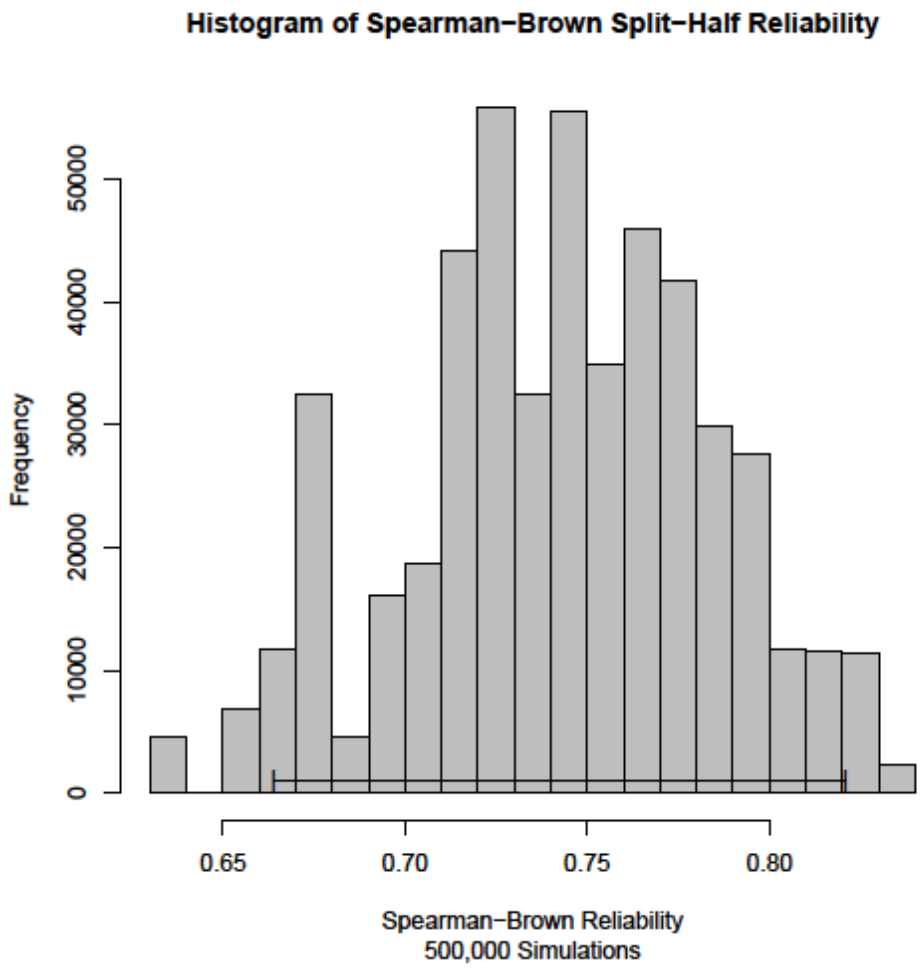


*Figure 4.* 500,000 simulated Spearman-Brown split-half reliability estimates with 95% confidence interval based on 0.025 and 0.975 quantiles.

Cronbach's alpha calculated based on a testlet representation of the response data (*i.e.* 7 items summed over subparts) of the I-STUDIO instrument is 0.71 with a 95% confidence interval of [0.62, 0.81]. The mean inter-item correlation of the testlet data is 0.27, with a 95% confidence interval of [0.20, 0.33] based on the bootstrap percentile method using 500,000 simulations.

### 4.4.4 Confirmatory factor analysis.

#### *4.4.4.1 Independent item modeling.*

Several confirmatory factor analysis (CFA) models were evaluated on the basis of conceptual and statistical fit. Conceptual models aligned with the test blueprint suggest that I-STUDIO scores would be dominated by the three major latent variable dimensions represented by Discernment, Forward-Reaching Transfer, and Backward-Reaching Transfer. Organization of such a model could conceivably be manifest in two ways CFA-1 (*Figure 1*) or CFA-2 (*Figure 5*). Note that each case is equivalent when used to accommodate all three latent dimensions (3LV), and similarly when collapsed to a unidimensional (1LV). The difference between CFA-1 and CFA-2 is manifest when they are defined to model two latent dimensions (2LV).

The first group of CFA models evaluated were based on 15 scores including 9 scoring elements aligned to the subparts of items 1-4, and 6 scoring elements for items 5-7 based on a context score and a component score for each (e.g. *Figure 1*). Items were partitioned to represent two latent variables of Discernment and Transfer (2LV-Discernment model). Scoring elements 1a, 1b, 2a, 3a, and 4a represented the Discernment ability trait; scoring elements 1c, 2b, 3b, 4b, 5-context, 5-component, 6-context, 6-component, 7-context, and

7-component represented the Transfer ability trait. A likelihood ratio test comparing the 2LV-Discernment model to a unidimensional model resulted in marginal evidence that the 2LV-Discernment model provided a statistical improvement over the unidimensional model ($p = 0.0578$).

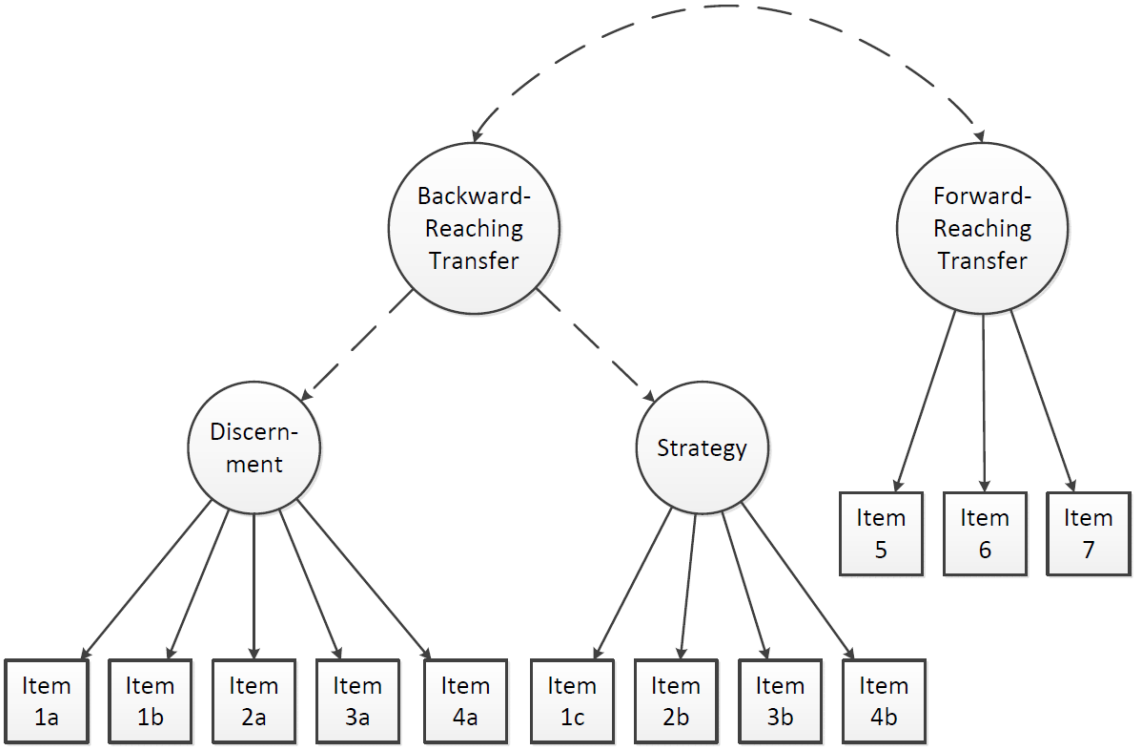


*Figure 5*. Confirmatory factor analysis model CFA-2.

The second group of CFA models evaluated were based on 15 scores including 9 scoring elements aligned to the subparts of items 1-4, and 6 scoring elements for items 5-7 based on a context score and a component score for each (*Figure 5*). Items were partitioned the 15 items as they align with Backward-Reaching and Forward-Reaching transfer outcomes (2LV-Transfer model). Scoring elements 1a, 1b, 1c, 2a, 2b, 3a, 3b, 4a, and 4b represented the Backward-Reaching transfer tasks, while scoring elements 5-

context, 5-component, 6-context, 6-component, 7-context, and 7-component represented Forward-Reaching transfer tasks. A likelihood ratio test comparing the 2LV-Transfer model, to the unidimensional model resulted in highly significant evidence that the 2LV-Transfer model provided a statistical improvement over the unidimensional model ($p < 0.0001$).

The 2LV-Transfer model was then extended to a 3LV-model, such that items were partitioned to represent latent variables of Discernment, Backward-Transfer, and Forward-Transfer (3LV model). The 3LV model appears to converge, however the covariance matrix among the three latent variables is non-positive definite, undermining confidence in the estimates produced by the fit. Furthermore, a likelihood ratio test comparing the 3LV model to the 2LV-Transfer model did not show strong evidence of a statistically significant improvement ($p = 0.3632$). Fit diagnostics for independent item models are shown in Table 14. The correlation between Backward-Reaching transfer ability and Forward-Reaching transfer ability is 0.648 as estimated by the 2LV-Transfer model. Parameter estimates for the 2LV-Transfer model are shown in Table 15. Note the inclusion of standardized estimates, which may have a slightly more convenient interpretation in the context of latent variable modeling. The standardized estimate represents the change in the latent variable on a standardized scale (i.e. standard deviation) for each standard deviation of improvement on the associated task.

Table 14

*Confirmatory factor analysis (CFA) fit diagnostics for independent item models*

| CFA Model Fit Diagnostic | 1LV | 2LV-Disc-ernment | 2LV-Transfer | 3LV |
|---|---|---|---|---|

| CFA Model Fit Diagnostic | 1LV | 2LV-Disc-ernment | 2LV-Transfer | 3LV |
|---|---|---|---|---|
| AIC | 6237.2 | 6235.6 | 6190.5 | 6192.5 |
| Chi-Square Test (p-value) | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| Chi-Square Ratio (Stat./DF) | 4.693 | 4.706 | 4.199 | 4.272 |
| Goodness of fit | 0.780 | 0.780 | 0.796 | 0.798 |
| Adjusted goodness of fit | 0.706 | 0.703 | 0.725 | 0.721 |
| Residual item corr. > 0.1 | 19.0% | 20.0% | 21.9% | 21.0% |
| Residual item corr. > 0.05 | 47.6% | 52.4% | 59.0% | 54.3% |
| RMSEA* | 0.144 | 0.144 | 0.134 | 0.136 |
| 90% CI for RMSEA | (0.13, 0.16) | (0.13, 0.16) | (0.12, 0.15) | (0.12, 0.15) |
| McDonald Noncentrality Idx | 0.393 | 0.396 | 0.449 | 53.575 |
| Hoelter's Critical N | 48.679 | 48.611 | 54.361 | 0.45 |

* RMSEA = Root mean square error of approximation.

Table 15

*Parameter estimates for 2LV-Transfer model fit*

| | Estimate (SE) | Z-value | P-value | Standardized Est. |
|---|---|---|---|---|
| Item 1a | 0.265 (0.058) | 4.545 | < 0.001 | 0.371 |
| Item 1b | 0.269 (0.067) | 4.014 | < 0.001 | 0.330 |
| Item 1c | 0.395 (0.053) | 7.406 | < 0.001 | 0.574 |
| Item 2a | 0.247 (0.069) | 3.551 | < 0.001 | 0.294 |
| Item 2b | 0.326 (0.042) | 7.774 | < 0.001 | 0.598 |
| Item 3a | 0.211 (0.058) | 3.613 | < 0.001 | 0.299 |
| Item 3b | 0.243 (0.078) | 3.120 | 0.002 | 0.259 |
| Item 4a | 0.412 (0.062) | 6.656 | < 0.001 | 0.524 |
| Item 4b | 0.516 (0.052) | 9.910 | < 0.001 | 0.732 |
| Item 5-context | 0.521 (0.054) | 9.677 | < 0.001 | 0.696 |
| Item 5-component | 1.186 (0.107) | 11.081 | < 0.001 | 0.772 |
| Item 6-context | 0.325 (0.046) | 7.045 | < 0.001 | 0.537 |
| Item 6-component | 0.818 (0.099) | 8.231 | < 0.001 | 0.611 |
| Item 7-context | 0.415 (0.057) | 7.325 | < 0.001 | 0.555 |
| Item 7-component | 0.723 (0.110) | 6.597 | < 0.001 | 0.507 |

#### *4.4.4.2 Correlated item modeling.*

Correlated item models were also evaluated in order to acknowledge the correlation structure among item sub-parts (e.g. 2a and 2b). The 2LV-Corr model was defined as a

modification to the 2LV-Transfer model such that correlation among the following item pairs was free to vary: (1a, 1b); (2a, 2b); (3a, 3b); (4a, 4b); (5-context, 5-component), (6-context, 6-component), and (7-context, 7-component).

A likelihood ratio test comparing the 2LV-Corr model, to the analogous unidimensional correlated item (1LV-Corr) model resulted in significant evidence that the 2LV-Corr model provided a statistical improvement over the 1LV-Corr model ($p = 0.0060$). Similarly, likelihood ratio tests showed that both the 1LV-Corr model and the 2LV-Corr model resulted in highly significant evidence of a statistical improvement over the base unidimensional model without accommodations for item correlation structure ($p < 0.0001$ in both cases). Fit diagnostics for correlated item models are shown in Table 16. The correlation between Backward-Reaching transfer ability and Forward-Reaching transfer ability is 0.806 as estimated by the 2LV-Corr model. Parameter estimates for the 2LV-Corr model are shown in Table 17, and item correlation estimates are shown in Table 18. Again, the standardized estimate indicates the change in the latent variable on a standardized scale (i.e. standard deviation) for each standard deviation of improvement on the associated task.

Table 16

*Confirmatory factor analysis (CFA) fit diagnostics for correlated item models*

| CFA Model Fit Diagnostic | 1LV-Corr | 2LV-Corr |
|---|---|---|
| AIC | 5961.7 | 5956.1 |
| Chi-Square Test | < 0.001 | 0.001 |
| Chi-Square Ratio (Statistic / DF) | 1.601 | 1.528 |
| Goodness of fit | 0.907 | 0.911 |
| Adjusted goodness of fit | 0.865 | 0.870 |
| Residual item corr. greater than 0.1 | 10.5% | 10.5% |
| Residual item corr. greater than 0.05 | 36.2% | 35.2% |

| CFA Model Fit Diagnostic | 1LV-Corr | 2LV-Corr |
|---|---|---|
| RMSEA* | 0.058 | 0.054 |
| 90% CI for RMSEA | (0.039, 0.076) | (0.034, 0.073) |
| McDonald Noncentrality Index | 0.869 | 0.885 |
| Hoelter's Critical N | 142.031 | 148.936 |

* RMSEA = Root mean square error of approximation.

Table 17

*Parameter estimates for 2LV-Corr model fit*

| | Estimate (SE) | Z-value | P-value | Standardized Est. |
|---|---|---|---|---|
| Item 1a | 0.263 (0.060) | 4.407 | < 0.001 | 0.369 |
| Item 1b | 0.261 (0.069) | 3.784 | < 0.001 | 0.320 |
| Item 1c | 0.406 (0.055) | 7.444 | < 0.001 | 0.591 |
| Item 2a | 0.176 (0.074) | 2.373 | 0.018 | 0.209 |
| Item 2b | 0.343 (0.043) | 8.003 | < 0.001 | 0.630 |
| Item 3a | 0.190 (0.060) | 3.178 | 0.001 | 0.270 |
| Item 3b | 0.209 (0.080) | 2.609 | 0.009 | 0.223 |
| Item 4a | 0.277 (0.068) | 4.070 | < 0.001 | 0.352 |
| Item 4b | 0.459 (0.055) | 8.311 | < 0.001 | 0.459 |
| Item 5-context | 0.398 (0.062) | 6.436 | < 0.001 | 0.532 |
| Item 5-component | 1.017 (0.120) | 8.454 | < 0.001 | 0.662 |
| Item 6-context | 0.324 (0.051) | 6.366 | < 0.001 | 0.535 |
| Item 6-component | 0.873 (0.107) | 8.126 | < 0.001 | 0.652 |
| Item 7-context | 0.400 (0.060) | 6.629 | < 0.001 | 0.534 |
| Item 7-component | 0.701 (0.116) | 6.025 | < 0.001 | 0.492 |

Table 18

*Correlation estimates for 2LV-Corr model fit*

| | Estimate (SE) | Z-value | P-value | Standardized Est. |
|---|---|---|---|---|
| Item 1a and 1b | 0.195 (0.044) | 4.469 | < 0.001 | 0.379 |
| Item 2a and 2b | 0.100 (0.031) | 3.201 | 0.001 | 0.287 |
| Item 3a and 3b | 0.229 (0.051) | 4.497 | < 0.001 | 0.369 |
| Item 4a and 4b | 0.187 (0.040) | 4.721 | < 0.001 | 0.473 |
| Item 5-context and 5-component | 0.426 (0.083) | 5.106 | < 0.001 | 0.583 |
| Item 6-context and 6-component | 0.045 (0.053) | 0.851 | 0.395 | 0.086 |
| Item 7-context and 7-component | 0.521 (0.082) | 6.354 | < 0.001 | 0.664 |

#### *4.4.4.3 Testlet modeling.*

Testlet models were evaluated such that all item sub-parts are aggregated into a single score as an alternative adjustment to accommodate the correlation structure among item sub-parts. For example, the item 1 testlet score is the sum of 1a, 1b, and 1c scores. The 2LV-Testlet model was defined as a modification to the 2LV-Transfer model such that the testlet items 1, 2, 3, and 4 represent Backward-Reaching transfer ability, and testlet items 5, 6, and 7 represent Forward-Reaching transfer ability.

A likelihood ratio test comparing the 2LV-Testlet, to the analogous unidimensional testlet model (1LV-Testlet) resulted in significant evidence that the 2LV-Testlet model provided a statistical improvement over the 1LV-Testlet model ($p = 0.0122$). Fit diagnostics for the testlet models are shown in Table 19. The correlation between Backward-Reaching transfer ability and Forward-Reaching transfer ability is 0.790 as estimated by the 2LV-Testlet model. Parameter estimates for the 2LV-Testlet model are shown in Table 20. Again, the standardized estimate estimates the change in the latent variable on a standardized scale (i.e. standard deviation) for each standard deviation of improvement on the associated task.

Table 19

*Confirmatory factor analysis (CFA) fit diagnostics for testlet models*

| CFA Model Fit Diagnostic | 1LV-Testlet | 2LV-Testlet |
|---|---|---|
| AIC | 4527.6 | 4523.4 |
| Chi-Square Test | 0.162 | 0.462 |
| Ratio of Chi-Square Statistic / DF | 1.363 | 0.986 |
| Goodness of Fit | 0.970 | 0.980 |
| Adjusted Goodness of Fit | 0.940 | 0.958 |

| CFA Model Fit Diagnostic | 1LV-Testlet | 2LV-Testlet |
|---|---|---|
| Residual item corr. greater than 0.1 | 0.0% | 4.8% |
| Residual item corr. greater than 0.05 | 33.3% | 28.6% |
| RMSEA* | 0.045 | 0.000 |
| 90% CI for RMSEA | (0.000, 0.091) | (0.000, 0.073) |
| McDonald Noncentrality Index | 0.986 | 1.001 |
| Hoelter's Critical N | 221.858 | 311.667 |

* RMSEA = Root mean square error of approximation.

Table 20

*Parameter estimates for 2LV-Testlet model fit*

| | Estimate (SE) | Z-value | P-value | Standardized Est. |
|---|---|---|---|---|
| Item 1 | 0.909 (0.140) | 6.516 | < 0.001 | 0.567 |
| Item 2 | 0.562 (0.100) | 5.597 | < 0.001 | 0.489 |
| Item 3 | 0.471 (0.123) | 3.833 | < 0.001 | 0.341 |
| Item 4 | 0.726 (0.115) | 6.319 | < 0.001 | 0.550 |
| Item 5 | 1.434 (0.172) | 8.315 | < 0.001 | 0.670 |
| Item 6 | 1.201 (0.135) | 8.895 | < 0.001 | 0.717 |
| Item 7 | 1.076 (0.168) | 6.409 | < 0.001 | 0.526 |

#### *4.4.4.4 Model selection.*

Since the structure of the data set was altered by the act of aggregating testlet scores, the choice between 2LV-Corr and 2LV-Testlet cannot be informed by comparison of AIC, nor can it be informed by a likelihood ratio test. As a result, selection between the 2LV-Corr and 2LV-Testlet models must be based on conceptual adherence to the structure of the I-STUDIO instrument, and informed by absolute fit diagnostics that evaluate model fit without use of a "base model" since the 2LV-Corr and 2LV-Testlet models do not share a common base model.

Upon comparison of fit diagnostics, both models seem to fit the data quite well. However, the fit diagnostics associated with the 2LV-Testlet model (shown in Table 19)

are universally at least slightly better by comparison to the 2LV-Corr model (shown in Table 16). Moreover, the testlet structure is consistent with the I-STUDIO instrument design due to the use of vignettes and prompts with multiple tasks or scoring elements associated with each one. Consequently, the 2LV-Testlet model was chosen for item analysis using multidimensional item response theory. Figure 6 illustrates the 2LV-Testlet model.

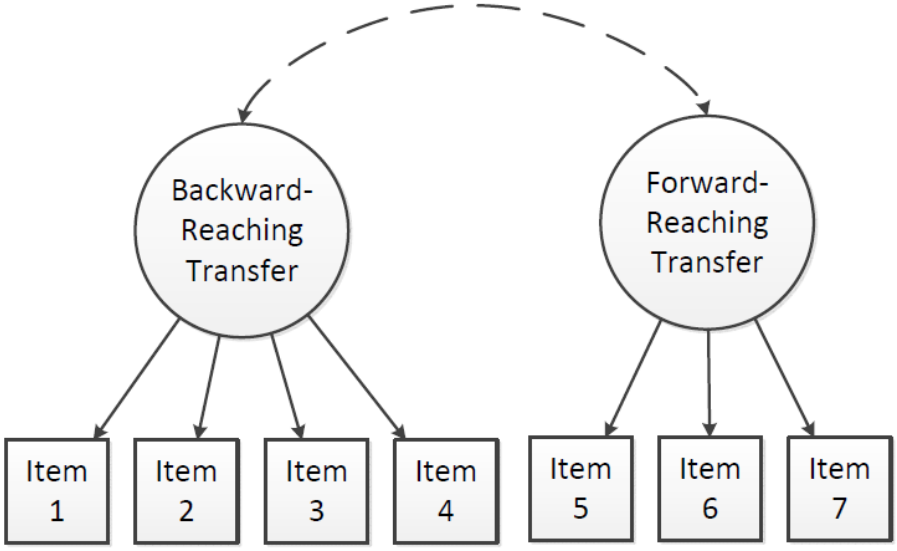


*Figure 6*. Confirmatory factor analysis model for testlet data on two dimensions.

### 4.4.5 Item analysis.

#### *4.4.5.1 Analysis of multidimensional item response data.*

Multidimensional item response theory (MIRT) was used to evaluate partial credit (PC) and graded response (GR) models of the testlet data. Comparison of AIC indicates that the MIRT-GR model (AIC = 4103.0) offers improvement over the MIRT-PC model (AIC = 4138.7) of the testlet data. Item fit diagnostics shown in Table 21 indicate that item 2 suggests marginal evidence of poor item fit ($p = 0.0401$) according to S-X2. While

S-X2 is robust to sample size when controlling type I error, it has low power to detect poor item fit for small sample sizes (Kang & Chen, 2008). A sample size of 178 is very small by IRT standards (Kang & Chen, 2008).

Table 21

*Item fit diagnostics associated with MIRT graded response model*

| Item (testlet) | S-X2 | d.f. | *P*-value |
|---|---|---|---|
| 1 | 32.38 | 50 | 0.9749 |
| 2 | 56.93 | 40 | 0.0401 |
| 3 | 39.02 | 30 | 0.1252 |
| 4 | 60.02 | 49 | 0.1344 |
| 5 | 46.08 | 44 | 0.3861 |
| 6 | 46.24 | 45 | 0.4208 |
| 7 | 40.84 | 53 | 0.8887 |

Factor loadings associated with the MIRT-GR model appear in Table 22 and the coefficient estimates are shown in Table 23. Note that $\alpha_{(\cdot)}$ represents item discernment and $\delta_{(\cdot)}$ estimates the median ability-level among students that earned the corresponding score (i.e. difficulty). The correlation between the Backward-Reaching dimension and the Forward-Reaching dimension is estimated as 0.811 by the MIRT-GR model. Overlaid test information curves associated with Backward-Reaching and Forward-Reaching transfer dimensions are shown in *Figure 7*. Overlaid item information curves appear in *Figure 8*. Option response functions (ORFs) associated with Backward-Reaching and Forward-Reaching transfer dimensions appear in *Figure 9* and *Figure 10*, respectively.

Table 22

*Factor loadings associated with MIRT graded response model*

| Item (testlet) | Backward-Reaching Dimension | Forward-Reaching Dimension |
|---|---|---|

| Item (testlet) | Backward-Reaching Dimension | Forward-Reaching Dimension |
|---|---|---|
| 1 | 0.60 | 0.00 |
| 2 | 0.55 | 0.00 |
| 3 | 0.40 | 0.00 |
| 4 | 0.58 | 0.00 |
| 5 | 0.00 | 0.73 |
| 6 | 0.00 | 0.76 |
| 7 | 0.00 | 0.59 |

Table 23

*I-STUDIO graded response model coefficient estimates*

| Item (testlet) | $\alpha_{(backward)}$ | $\alpha_{(forward)}$ | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_4$ | $\delta_5$ | $\delta_6$ | $\delta_7$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.261 | 0.000 | -3.39 | -2.47 | -0.75 | 0.29 | 1.58 | 2.77 | |
| 2 | 1.114 | 0.000 | -3.42 | -1.13 | -0.19 | 1.95 | | | |
| 3 | 0.738 | 0.000 | -2.42 | -1.54 | -0.40 | 0.05 | | | |
| 4 | 1.209 | 0.000 | -2.68 | -1.26 | -0.28 | 1.21 | | | |
| 5 | 0.000 | 1.813 | -4.00 | -2.40 | -1.72 | -0.03 | 0.62 | 1.54 | 2.07 |
| 6 | 0.000 | 1.964 | -3.02 | -2.09 | -0.32 | 1.10 | 2.92 | 3.57 | |
| 7 | 0.000 | 1.236 | -3.01 | -1.29 | -0.54 | -0.17 | 0.33 | 1.86 | |

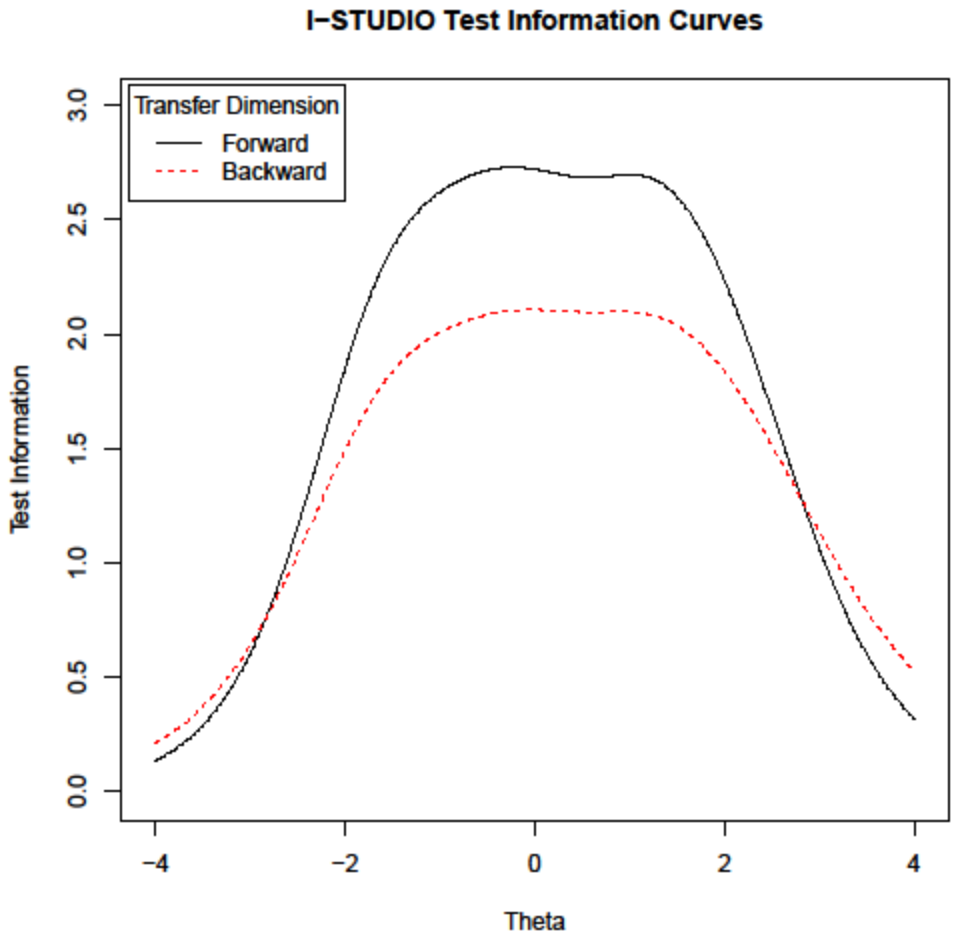


*Figure 7*. I-STUDIO Test information curves for Forward-Reaching and Backward-Reaching transfer dimensions.

The total test information curves show the precision with which I-STUDIO estimates ability according to each dimension. Inspection of *Figure 7* suggests that I-STUDIO estimates Forward-Reaching transfer with slightly better precision than Backward-Reaching transfer. Precision appears reasonably stable for ability estimates from about -2 to 2 on each dimension.

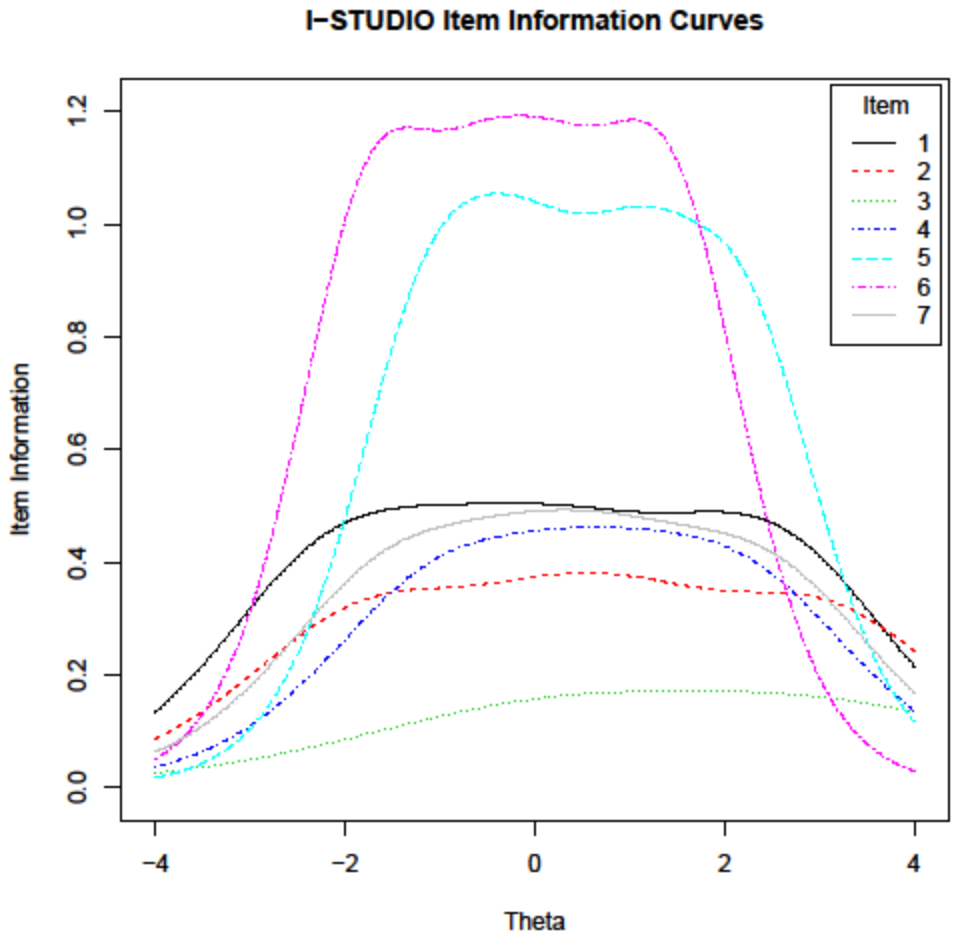


*Figure 8*. I-STUDIO item information curves.

Each item contributes to the total information curve. Inspection of *Figure 8* reveals that item 5 (Matched Pairs Study Design) and item 6 (Underlying Principle of Inference) contributed the most information. Item 3 (Display Screen Inspection) contributed somewhat less information than the other items.

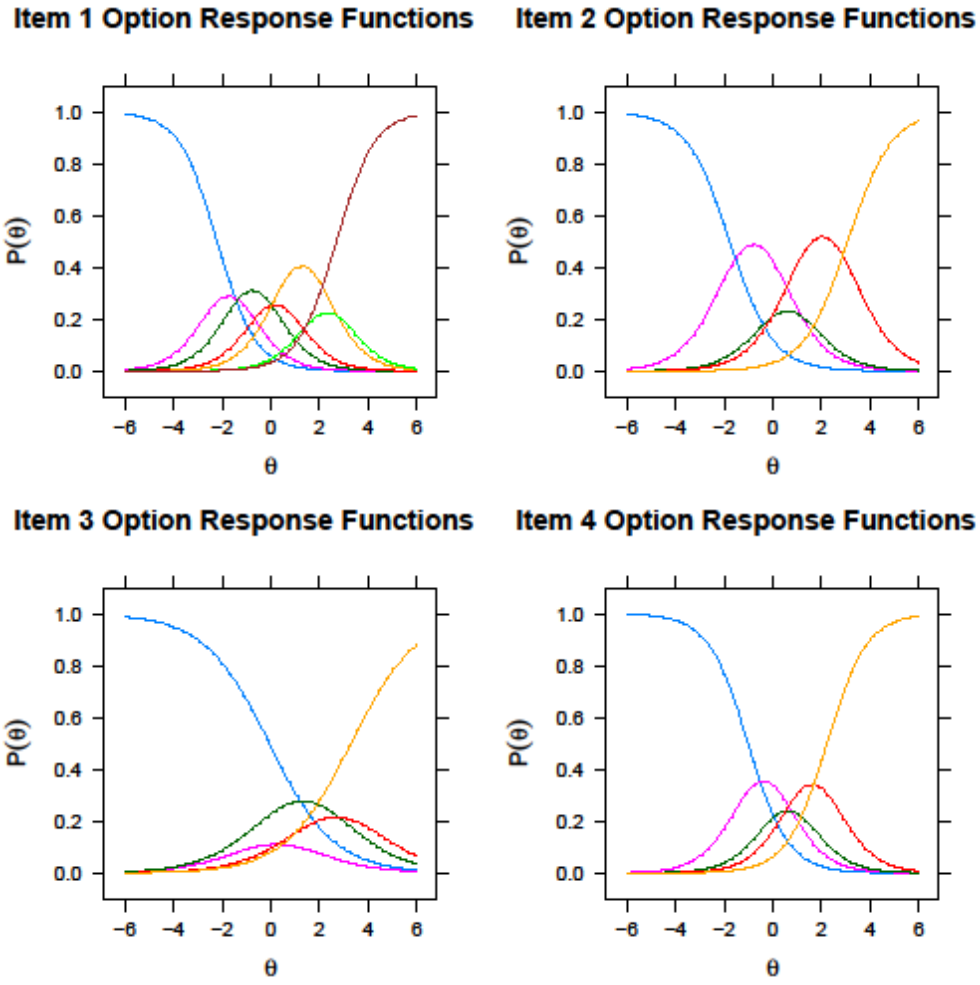


*Figure 9*. I-STUDIO option response functions for Backward-Reaching transfer testlets.

Constraints of the graded response model dictate that each successive ORF correspond to successive point values attained for the item. For example, item 3 (Display Screen Inspection) includes 4 distinct curves corresponding to possible outcomes of 0, 1, 2, 3, 4 points in succession. The ORFs for item 3 in *Figure 9* suggest that students across the entire ability spectrum were more likely to earn 0 or 2 points then 1 point. Several intermediate score outcomes were observed to be similarly unlikely for item 7 as shown in the ORF curves corresponding to Forward-Reaching transfer items on display in

*Figure 10*. As such, the ORF curves can be used to observe the score outcomes most likely to have been earned by students at a given ability level for a given item, or conversely to observe score outcomes that were never the most likely outcome for students at any ability for a given item.

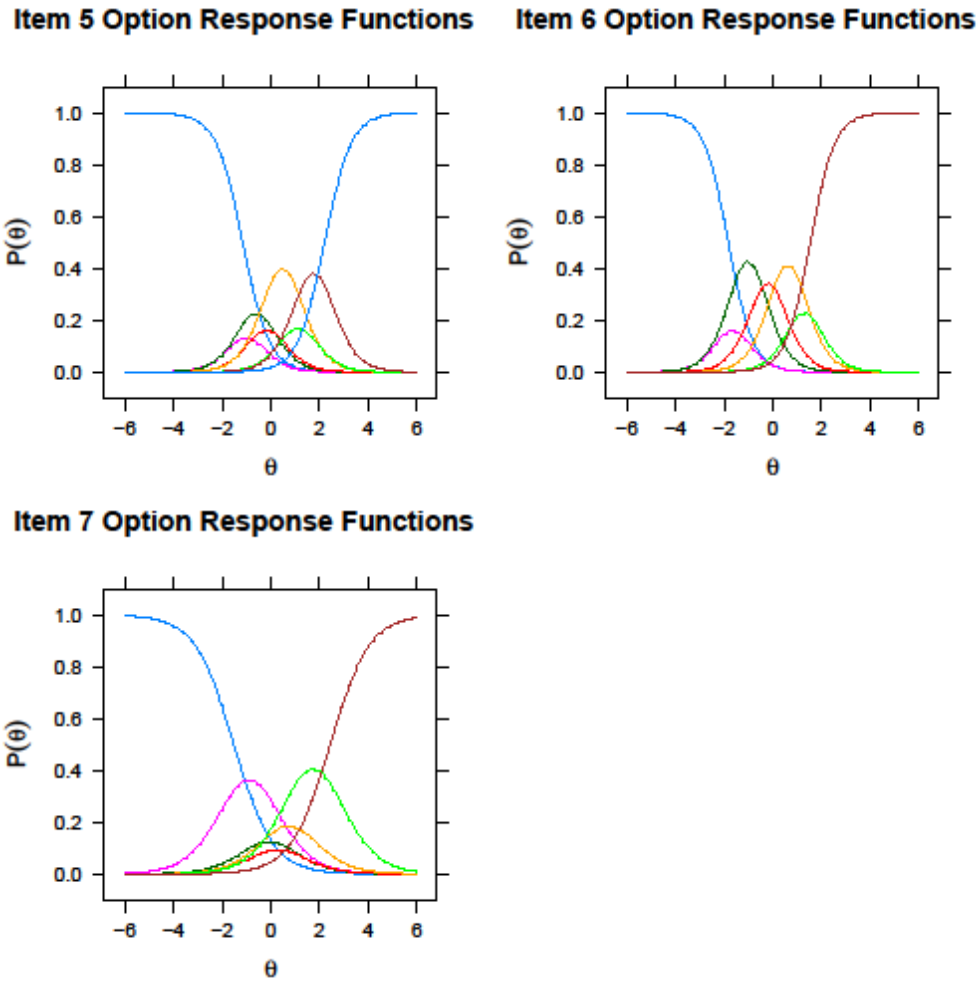


*Figure 10*. I-STUDIO option response functions for Forward-Reaching transfer testlets.

#### *4.4.5.2 Qualitative Analysis of student responses.*

While scoring student responses, qualitative evidence of unusual or unexpected response patterns were noted to accompany several items. Item references refer to the version of I-STUDIO presented to students for the field test found in Appendix G: I-STUDIO Version for Field Test. Flawed responses providing exemplars of themes observed among the data are tabulated in Table 24.

In item 1 (air traffic control), students were generally quite successful at posing viable research questions, but had great difficulty acting upon them and describing inferential analysis. Since each student was at liberty to construct their own research questions in 1a and 1b, and choose which research question to address in 1c, exemplars of universal themes were not observed.

Responses to item 2a (note identification) showed evidence of two noteworthy themes. The first theme indicates that a subset of students expressed a view that statistical inference only applies to quantitative data, and the second theme showed that some students seemed fixated on comparisons to a population of other people. These students did not recognize that we can generalize Carla's data to her process/probability of note identification. Flaws common among item 2b responses were based on use of only a point estimate to draw conclusions about Carla's note identification ability.

A common issue among responses to item 3 (display screen inspection) was the evidence of apparent conflict between a response to 3a advocating for use of statistical inference and then recommending a non-inferential solution in 3b such as inspection of all 150 screens and making a decision by comparing the observed proportion to the 5%

threshold. In item 4a (Walleye fishing), several students described that statistical inference should not be used because the scenario does not describe a designed experiment. Other students stated that statistical inference is not appropriate because there is no population of interest to which the two brothers would generalize, and still others contended that inference is not appropriate because one brother could simply get "lucky." For item 4b, several students stated that they would simply compare point estimates with no mention of inferential methods.

Among forward-reaching transfer tasks, a common flaw among student responses was simply to reiterate the stem without situating the response into a context of any kind. For example, a response to item 5 shown in Table 24 indicated that the student had some procedural knowledge for executing a paired t-test, but there is no indication that the student attempted to establish any kind of context beyond the generic examples provided in the item stem. It is also noteworthy that a nontrivial group of students chose not to attempt item 5—either skipped or stated that paired comparisons weren't "covered" in their class—but then continued to item 6.

Items 6 and 7 were parallel tasks such that item 6 prompted a scenario for which statistical inference is appropriate and item 7 prompted a scenario for which statistical inference was not appropriate. The most common issue was difficulty identifying the parameter. Students were frequently observed conflating the parameter with a population or a variable.

Table 24

*Flawed example responses (verbatim) to several I-STUDIO items*

| Item | Student Response (verbatim) |
|---|---|
| 2a | - "Statistical Inference should not be used to determine whether Carla has a good ear for music. Statistical inference should be used on things that can be measured, things that have a numerical value (i.e : # of eggs, # of gallons of milk). You can not measure using numbers whether or not she has a good ear for music."<br>- "Since statistical inferences measure population it would not be a good idea to use this in the case of carla because it is measuring the accuracy of her note identification skills, not measuring a population." |
| 2b | - "By considering how many notes they correctly guessed, and comparing it to the total number of notes played. If they correctly identify a majority of the notes, they have a good ear."<br>- "Conduct a simulation in which a student is asked to identify 10 random notes for one trial. Have the student complete multiple trials and count how many notes they correctly identified to get an accurate estimate of the proportion of times they were able to correctly identify the note. Record their results and make a graph of the distribution of the trials. Then compare this proportion with other people who are known to have a good ear for music." |
| 3 | - [Part a]"You should use statistical inference to determine whether the company should accept or reject the bulk order of display screens because the data gathered by the trained engineer must be analyzed to determine if there is more than 5% of display screens that are bad." [Part b] "In order for the company to reject the 150 display screens, more than 5 percent of the screens must be bad. That means that there has to be at least 8 screens out of 150 in order to meet the 5% rejection requirement. Anything less than 8 would not meet 5% requirement." |
| 4a | - "No, becasue this is not a random sample and is not a real experiment."<br>- "Statistical inference is not applicable in this case since the inference is not about a larger population, it is merely a comparison of two individuals"<br>- "No. Catching a fish is based largely on luck so you can't use statistics to see who the better fisherman is." |
| 4b | - "Sum all fish lenghts caught by Mark and divide them by N1. Sum all fish lenghts caught by Dank and divide them by N2. Compare the two mean lenghts." |
| 5 | - "With the participants of the matched pairs study being people and animals we are looking at the results of two treatments. We will create two lists of the results of the treatments, one being treatment1 and the other being treatment2. After doing so we will do a paired t |

| Item | Student Response (verbatim) |
|---|---|
| | test of t.test(treatment1, treatment2, paired=TRUE, mu=o, alternative= "two.sided") This will give us our p-value and if the p-value was less than 0.05 we would reject that there is no difference between the two treatment groups, but if the p-value was greater than 0.05 we cannot reject that there is no difference between the treatment groups." |
| 6 | - "The research question is if people prefer to run or bike as a form or cardio exercise. A group of 1,000 students would be randomly selected within a school campus as a sample from the whole campus. The population of this test would be the student body. The statistic is what students preferred as a form of exercise, whether to bike or run. The parameter is the results that would come from this study."<br>- "If you would like to figure out the average height of men aged from 20-35? / Population: Everyone in that age range / Sample: selections made from the population / Statistic: The height from the men / Parameter: The people who are getting tested " |
| 7 | - "Does alcohol contribute to worse G.P.A.? / parameter: college students in America / population: students at all colleges inAmerica / data: G.P.A., amount of times student drinks per week / / use data to see of there is a correlation between G.P.A. and summer of times student drinks per week"<br>- "1) The proportion of all undergraduate students that have a pet dog 2) Parameter of interest= dog owner 3) Population= all undergraduate students 4) Data= whether or not the students has a pet dog currently 5) You could use this data to support the already known population parameter value" |

# 5 Discussion

## 5.1 Study Summary

The I-STUDIO instrument was developed to explore the feasibility of creating an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. Data were collected and analyzed from a nationwide sample of students attending a wide variety of post-secondary institutions, and the I-STUDIO instrument was found to measure both forward-reaching and backward-reaching high road transfer outcomes with good psychometric properties.

The I-STUDIO instrument was developed according to a rigorous protocol of expert feedback and iterative piloting. The instrument was modeled after a test blueprint which was developed according to evidence in the literature describing characteristics of forward-reaching and backward-reaching high-road transfer (Salomon & Perkins, 1989). The blueprint was then scrutinized by a group of experts in statistics, education, measurement, and cognitive transfer and then modified prior to development of a draft instrument.

The preliminary assessment tool was created by organizing items borrowed and adapted from published assessment items in the literature (e.g. Chance, 2002; Garfield et al., 2012). The same group of expert reviewers then provided feedback for the instrument, and the I-STUDIO assessment was again refined prior to use with students. The first group of students to encounter the I-STUDIO assessment, completed the instrument during a think-aloud cognitive interview with the author in the room recording

audio and taking notes. The I-STUDIO instrument was again updated to mitigate issues that evoked confused or unintended responses from students.

Finally the I-STUDIO assessment was presented to a nationwide sample of nearly 2,000 students attending a wide variety of post-secondary institutions. One subset of 24 student responses was used to develop a scoring rubric that was refined by peer review and evaluated for inter-rater consistency. A random sample of 178 students was selected to represent all participating course, and their responses were evaluated to examine the reliability and validity of I-STUDIO as well as explore item response attributes.

## 5.2 Synthesis of Results

The goal of this study was to explore the feasibility of developing an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. Prior to this study, no published assessment had been designed to measure this specific outcome, and the literature suggested uncertainty about whether cognitive transfer can be achieved and measured following an introductory statistics course. Evidence supporting this central goal can be synthesized from general expert feedback, reliability metrics, validity evidence, and item analysis.

### 5.2.1 General comments from expert feedback.

Overall, the expert feedback for both the test blueprint and the draft instrument was generally quite positive. On several occasions, reviewers shared feedback critiquing example responses provided to accompany the draft instrument. These were intended to model minimally acceptable responses, and not necessarily a gold standard. The instructions to the expert reviewers were not clear to this effect, but any issues cited with

the example solutions were taken into careful consideration while crafting the scoring rubrics.

One reviewer recommended that it would be valuable to study whether students are able to transfer what they have learned new structures and frameworks that were not explicitly learned. For example, perhaps a class did not discuss methods for comparison of multiple group means (e.g. ANOVA), so. Specifically, the reviewer seems to be suggesting that the instrument emphasize greater distance of transfer. This comment is an important one and speaks to a key aspect of cognitive transfer (Bransford et al., 2000). Other poignant remarks provided among the expert feedback related to the type of items for which statistical inference is not appropriate. Another comment suggested that the instrument is missing "the clearly badly collected data situations and realize they shouldn't use inference when they don't have randomness." This does seem to be an important archetype for a data analyst to recognize, and such an item may warrant inclusion in a future version of I-STUDIO or a similar assessment tool. Now that this study has demonstrated evidence that high-road transfer outcomes can be reliably measured following the introductory statistics curriculum, modifications that increase the distance of transfer or touch on alternative archetypes are natural avenues for future research.

### 5.2.2 Evidence of quality of the I-STUDIO assessment tool.

#### *5.2.2.1 Reliability.*

The estimated reliability of I-STUDIO was quite strong given the context that the assessment tool aims to aid decisions at the curriculum (i.e. class) level rather than at the

level of an individual student. Along these lines, one way to characterize strength of instrument reliability is to consider likelihood of score differences being reversed upon repeated testing. Consider the event that the mean score for 25 students in class A is at the 75$^{th}$ percentile by comparison to some reference population, and the mean score for 25 students in class B is at the 50$^{th}$ percentile by comparison to the same reference population. Table 25, reproduced from Thorndike and Thorndike-Christ (2010), shows the probability that the mean score of class B would surpass the mean score of class A upon repeated testing is 0.001 for an instrument with estimated reliability of 0.70. This scenario falls well within the prediction interval produced by the Spearman-Brown split half reliability simulations, and is a close approximation to the lower bound for the transformed prediction interval projected to incorporate all 15 scoring elements. If the true reliability of I-STUDIO is closer to 0.80 the probability of a difference reversal in this scenario becomes trivial.

Table 25

*Probability of Difference Reversal with Repeated Testing for Classes of 25 Students*

| Test Reliability | Probability of Difference Reversal |
|---|---|
| 0.50 | 0.046 |
| 0.60 | 0.012 |
| 0.70 | 0.001 |
| 0.80 | <0.001 |

#### *5.2.2.2 Validity.*

Validity evidence supporting the I-STUDIO assessment was accrued through expert feedback while reviewing the test blueprint, expert feedback while reviewing the draft I-STUDIO assessment tool, scoring consistency among raters, estimated reliability, and

confirmatory factor analysis. Evidence of strong reliability metrics was discussed previously, so the balance of this section is focused on expert feedback, consistency of rubric use, and confirmatory factor analysis.

The expert feedback for the test blueprint and the draft version of the I-STUDIO assessment both provided favorable evidence that I-STUDIO was suitable for its intended use and likely to measure the intended outcomes. Test blueprint feedback helped to hone definitions of key terms and refine item concepts for use in the draft instrument. The feedback for the draft instrument included many useful suggestions to tune individual items to achieve their intended purposes.

One reviewer did remark that the I-STUDIO assessment seems to be "directing transfer, and not depending on spontaneous transfer." This is an important comment because spontaneous transfer is certainly at the core of the desired construct (Chance, 2002), but the operational details required for stimulating, observing, and measuring spontaneous transfer of inferential statistics knowledge greatly complicates things. The simple act of asking a statistics instructor to present students in a statistics class with a "test" that includes the term "statistics" in the title would logically compromise spontaneity. Having said that, the act of "directing transfer" as the reviewer stated provides an incremental step forward toward understanding how students transfer understanding to novel scenarios.

When studying reliability of rubric interpretation, the evidence suggests that inter-rater consistency (i.e. comparison of independent raters) was very high based on the proportion of score agreement and lack of serious discrepancy on any scoring element.

However, intra-rater consistency (analysis of single rater) results were suspiciously high, although the reasons for this seem quite clear. This scoring effort was conducted by one person over the course of several consecutive days. In order to mitigate drift of rubric interpretation, all 178 responses for a given item (e.g. item 4b) were scored within the space of a single day, often within a single sitting without interruption. As a result, responses that had been previously observed were easy to recognize. Furthermore, a protocol of instructions for rubric use including periodic review of the complete item rubric was followed as a second measure to prevent drift of rubric interpretation (Appendix J: I-STUDIO Scoring Rubric Use Instructions). As such, the estimated intra-rater consistency metric is almost certainly inflated, so perhaps more emphasis should be placed on the strength of the inter-rater reliability estimate. However, concerns for intra-rater consistency may be tempered somewhat given the operational steps taken to promote consistent interpretation of the rubric.

As a result of the confirmatory factor analysis, the item response data were most appropriately modeled according to the 2LV-Testlet structure. The data appear to fit the model very well according to fit diagnostics, and the testlet structure was well-suited to the design of the I-STUDIO assessment tool. However, one clear shortcoming of the 2LV-Testlet model is a loss of granularity to evaluate how well the tasks within each testlet function. For example, there may be some benefit to learning how well item 4a functions, but that information is confounded by item 4b since both parts were aggregated as a testlet score. Another shortcoming of the 2LV-Testlet model is that it would not allow a natural extension to incorporate the Discernment dimension if warranted by

future research. This seems a low risk since the Discernment dimension did not appear to meaningfully contribute, but this study had a relatively small sample size so it is possible that things may look differently with the benefit of additional data.

A potentially surprising outcome of the 2LV-Transfer model relates to the high correlation between the Backward-Reaching and Forward-Reaching dimensions. On the one hand, both dimensions are potentially related to a more abstract ability to achieve high-road transfer within a common domain of subject matter. On the other hand, it is surprising that the model produced such compelling evidence of multidimensionality when the dimensions were so highly correlated.

Consequently, the CFA model results could be interpreted to offer somewhat mixed validity evidence. The results would seem to corroborate a theory of statistical thinking propagated by Wild and Pfannkuch (1999) that described the act of mental shuttling back and forth between the context domain and the schema for abstract modeling archetypes. The evidence of both forward-reaching and backward-reaching transfer as distinct dimensions suggests an ability to isolate and measure the dexterity with which students perform as they shuttle in each direction.

While affirming that more than one dimension was manifest in the scoring data, it was unexpected to learn that the Discernment dimension did not contribute further. In fact, while critiquing the draft I-STUDIO assessment tool, one reviewer expressed the opinion that the discernment dimension may even be more important than the other attributes tested by I-STUDIO as an indicator of student ability to apply statistics knowledge to novel contexts. However, two reasonable explanations come to mind. The

first explanation was foreshadowed in the test blueprint which asserted that forward-reaching high-road transfer by definition must include some measure of discernment. Therefore, the I-STUDIO assessment cannot include any forward-reaching high-road transfer item with no discernment required. As a result, it is conceivable that discernment ability was in part confounded with forward-reaching transfer ability as well as backward-reaching transfer ability.

Another possible explanation may relate to common practices for teaching and learning statistics. If backward-reaching transfer is approximated by tool selection, then in order to separate the discernment dimension from backward-reaching transfer a student would need to demonstrate an ability to recognize that a scenario may benefit from a statistical approach even in cases where they do not know what that approach should be. Statistical thinking of this nature may perhaps be expected from an advanced statistician, but is far more difficult for a novice in the introductory course (Lovett & Greenhouse, 2000).

#### *5.2.2.3 Item analysis.*

Item analysis consisted mainly of multidimensional item response theory (MIRT), and qualitative analysis of student responses to each item. Use of MIRT is not without a measure of caution given the relatively small sample size of this study. However, even if interpreted as a preliminary analysis, at a minimum this analysis would certainly invite further study. Furthermore, much can be learned through qualitative analysis of open-ended responses to the unique and demanding items that make up I-STUDIO.

Test information curves in *Figure 7* resulting from the MIRT analysis showed evidence of good coverage for abilities from roughly -2 to 2 on each dimension. Similarly, item information curves shown in *Figure 8* highlight the tasks that contribute most effectively across the range of ability levels. Item 3 (manufacturing lot inspection) stands out as the least informative. This is corroborated by coefficient estimates shown in Table 23 for both discrimination and difficulty, which were fairly low by comparison to the other items. Factor loadings were reasonably strong on both dimensions, though somewhat stronger on the Forward-Reaching dimension by comparison. The factor loadings associated with non-statistical data analysis (e.g. items 3 and 7) were slightly lower than the statistical items on each dimension.

Analysis of item fit using the S-X2 metric suggested marginal evidence that item 2 (note identification test) did not seem to function as well as expected. Possible resolutions to issues like this sometimes involve pursuing an isomorphic item, if perhaps the scenario is too unfamiliar for students to grasp. It is possible in this case that this item simply demands something different of students when compared with the other tasks in the instrument. For example, item 2 expects students to think about issues such as acceptance criteria and data collection differently than other items. Several students grappled with whether a better than chance (i.e. 1/7) result would really indicate that Carla has a "good ear for music," or should it be 80%? 90%? Some students suggested characterizing a whole population of students to establish a distribution for pitch recognition before we can declare what "good" might well look like. The item also required that students recognize the need for data collection—a key aspect of statistical thinking (Wild &

Pfannkuch, 1999)—but data collection was more overtly proposed in other Backward-Transfer items, perhaps leading this item to function a bit differently than expected.

Qualitative analysis of exceptional responses was focused primarily on noting patterns among flawed responses. Items on each dimension offer unique insights that may not be easily observed with a unidimensional or forced-response assessment tool. For example, a common mistake among the responses to questions 2a and 4a revealed a student misconception that a population for the purposes of inferential statistics must be a population of physical people or objects rather than a population of outcomes for some process. Consequently, corresponding solutions to 2b and 4b commonly imposed a population of music students to whom Carla could be compared and a population of fishermen to whom Mark and Dan could be compared.

Another noteworthy theme was the prevalence of contradictory responses among discernment tasks (2a, 3a, 4a) and strategy tasks (2b, 3b, 4b) among the Backward-Reaching Transfer items. Students frequently advocated for statistical inference in (a) and then described a non-inferential solution in (b), or rejected the need for inference in (a) and described an inferential strategy in (b). Perhaps the cause is as simple as unfamiliarity with the term “inference” but still an interesting result to observe from a group of students in the last weeks of a statistics course.

Finally, among the Forward-Reaching transfer items, there were a remarkable number of students that failed to properly identify the parameter of interest in a scenario of their own choosing. Responses were observed to conflate the parameter with almost every other detail of the scenario including statistics, populations, variables, and more. Again,

it’s possible that students simply struggle with the simple definition, but it would seem that the idea of the parameter is so near the core of inferential statistics that students could be expected to encounter the term somewhat regularly.

## 5.3 Study Limitations

During the process of designing and carrying out the study, several limitations deserve mention. First, and perhaps foremost, is the limited generalizability of the sample. Instructors participated on a voluntary basis, and all instructors that volunteered were included. Furthermore, instructors were only minimally constrained in their use of the instrument. It was requested that some incentive be offered to students in order to encourage legitimate effort, but the incentives were variable and some were more effective than others at stimulating the desired effort from students.

Moreover, the study aimed to produce an instrument robust to curriculum diversity, but the sample of participating courses apparently did not represent quite as much diversity as anticipated. All students that participated in the cognitive interviews had completed a course with at least half of its curriculum devoted to simulation-based methods, though this demographic was not well-represented during the field test. Only one participating class used a curriculum with substantial use of simulation-based methods; the total enrollment was 13 students and only 2 submitted useable responses. Another course in the study included nontrivial treatment of nonparametric methods in the curriculum. These responses were well-accommodated by the scoring rubric and provided preliminary evidence toward the aim of designing the assessment tool with robustness to curriculum approach, but more work is needed.

A few limitations among data collection are also apparent. For one, the test blueprint feedback questionnaire failed to explicitly request critique of task allocation. One reviewer did volunteer a remark in his feedback that he felt the item allocation seemed appropriate, but such feedback was not overtly solicited so other reviewers did not comment. Accompanying the field test data, it may have been nice to gather some basic demographics to assess differential item functioning, as well as final grade (or expected grade) and GPA for the purpose of corroborating scores as validity evidence. Similarly, the study does not include data necessary to assess whether transfer ability was measured distinctly from general intelligence, so the evidence cannot be used to inform either side of that debate (e.g. Detterman, 1993; Salomon & Perkins, 1989).

## 5.4 Implications for Teaching

If transfer outcomes are of value for the introductory statistics curriculum, then the I-STUDIO assessment tool provides an instrument with good psychometric properties that teachers can use for comparing outcomes of alternative curricula. Additionally, the I-STUDIO instrument can be used to measure the effect of curriculum changes designed to improve transfer outcomes. Again, the instrument and rubric are designed with intent to accommodate diverse curricula for the purpose of evaluating course outcomes.

## 5.5 Implications for Future Research

This study was scoped somewhat as a feasibility study. The results of the field study seemed to corroborate theoretical models for evoking backward-reaching and forward-reaching transfer outcomes, and data analysis presented strong reliability, rubric consistency, and validity. Consequently, one extension of value may be to simply score

and analyze a larger number of students from the present study. This work could help to refine estimates, advance qualitative themes observed, and test the capability of the scoring rubric to accommodate a wider variety of responses. Similarly, there may be value in targeted recruiting of introductory statistics curricula with unique approaches to further develop robustness of I-STUDIO to accommodate such diversity.

If the present study is interpreted to provide promising results that transfer outcomes of modest distance are measurable, then a natural extension may be to increase the distance of transfer. This could include incorporation of methods that push the students farther outside their experience (e.g. ANOVA or multiple regression for the introductory student). Alternatively, it may involve subjecting students to the assessment after nontrivial delay, such as at the beginning of a subsequent course or even after summer vacation.

Future research is also recommended to study discernment of whether statistical inference is appropriate for a problem setting. At this point it is not clear whether the discernment dimension could or should be expanded within the I-STUDIO instrument, or whether there would be value to creating a separate instrument for the purpose of measuring this outcome. It seems plausible that the discernment construct could have a place across the continuum of statistical literacy, reasoning, and thinking so further study to either isolate the outcome or understand its place within the larger paradigm of statistics education could be useful.

## 5.6 Conclusion

The I-STUDIO instrument was found to measure both forward-reaching and backward-reaching high road transfer outcomes with strong psychometric properties. Supporting evidence from national experts in the field suggests that I-STUDIO appropriately measures a construct of value to the introductory statistics curriculum. Data analysis included 178 student responses from a national sample 1935 responses contributed by 29 introductory statistics class sections across 12 courses at 11 different institutions.

Reliability evidence and inter-rater rubric consistency were both high, and the rubric was found robust to accommodate a variety of responses including nonparametric and simulation-based approaches. The I-STUDIO assessment tool has a strong battery of validity evidence including expert scrutiny and confirmatory factor analysis supporting its use as an instrument to measure cognitive transfer outcomes associated with the introductory statistics curriculum. The I-STUDIO instrument is well positioned to fill an important assessment role for the statistics education community to make the reliable comparisons of transfer outcomes that are needed to advance curriculum development and empower students to transfer statistical understanding to contexts beyond the introductory statistics course.

# 6 References


Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(5), 883-901.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Alexander, P. A., Murphy, P. K., & Kulikowich, J. M. (1998). What responses to domain-specific analogy problems reveal about emerging competence: A new perspective on an old acquaintance. *Journal of Educational Psychology, 90*(3), 397.

Alexander, P. A., & Murphy, P. K. (1999). Nurturing the seeds of transfer: A domain-specific perspective. *International Journal of Educational Research, 31*(7), 561-576. doi:10.1016/S0883-0355(99)00024-5

Atkinson, R. K., Catrambone, R., & Merrill, M. M. (2003). Aiding transfer in statistics: Examining the use of conceptually oriented equations and elaborations during subgoal learning. *Journal of Educational Psychology, 95*(4), 762.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York: Routledge.

Ben-Zvi, D., & Garfield, J. (2005). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15) Springer.

Binet, A. (1899). Attention et adaptation. *L'Année Psychologique, 6*(1), 248-404. doi:-10.3406/psy.1899.3114

Bloom, B. S. (1956). Taxonomy of educational objectives: Handbook I: Cognitive domain. *New York: David McKay*, *19*, 56.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academies.

Broers, N. J., Mur, M. C., & Bude, L. (2004). Directed self explanation in the study of statistics. In G. Burrill, & M. Camden (Eds.), *Curricular development in statistics education* (pp. 21-35). Voorburg, The Netherlands: International Statistical Institute.

Brogan, D., & Kutner, M. H. (1986). Graduate statistics service courses. *The American Statistician, 40*(3), 252-254.

Budé, L. (2006). Assessing students' understanding of statistics. Paper presented at the *Proceedings of the Seventh International Conference on Teaching of Statistics, CD-ROM, Salvador (Bahía), Brasil, International Association for Statistical Education.[Links]*,

Butterfield, E. C., & Nelson, G. D. (1991). Promoting positive transfer of different types. *Cognition and Instruction, 8*(1), 69-102.

CAUSE charter. (2006, April 26, 2014). Retrieved from https://www.causeweb.org/about/

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3)

Chi, M. T., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum. *Technology Innovations in Statistics Education, 1*(1), 1.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*(4), 347.

Cox, B. D. (1997). The rediscovery of the active learner in adaptive contexts: A developmental-historical analysis. *Educational Psychologist, 32*(1), 41-55. doi:10.1207/s15326985ep3201_4

Deary, I. J. (2001). *Intelligence: A very short introduction* Oxford University Press.

delMas, R. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education, 10*(3) Retrieved from http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html

delMas, R. (2006). Defining and distinguishing statistical literacy,<br />Statistical Reasoning, and statistical thinking. Retrieved from https://apps3.cehd.umn.edu/artist/glossary.html

Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman R. J. Sternberg (Ed.), (pp. 1-24). Westport, CT, US: Ablex Publishing.

Efron, B. (2000). The bootstrap and modern statistics. *Journal of the American Statistical Association, 95*(452), 1293-1296.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19*(4), 676-685.

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics, 1*(4), 292-298.

Gaither, N., & Glorfeld, L. (1985). An evaluation of the use of tests of significance in organizational behavior research. *Academy of Management Review, 10*(4), 787-793.

Gal, I., Ahlgren, C., Burrill, G., Landwehr, J., Rich, W., & Begg, A. (1995). Working group: Assessment of interpretive skills. Paper presented at the *Writing Group Draft Summaries, Conference on Assessment Issues in Statistics Education*, 23-25.

Garfield, J. (1995). How students learn statistics. *International Statistical Review/Revue Internationale De Statistique, 63*(1), 25-34.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3)

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. Retrieved from http://link.springer.com.ezp1.lib.umn.edu/article/10.1007/s11858-012-0447-5/fulltext.html

Georghiades, P. (2000). Beyond conceptual change learning in science education: Focusing on transfer, durability and metacognition. *Educational Research, 42*(2), 119-139.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*(3), 306-355. doi:http://dx.doi.org.ezp1.lib.umn.edu/10.1016/0010-0285(80)90013-4

Giesbrecht, N., Sell, Y., Scialfa, C., Sandals, L., & Ehlers, P. (1997). Essential topics in introductory statistics and methodology courses. *Teaching of Psychology, 24*(4), 242-246.

Goska, R. E., & Ackerman, P. L. (1996). An aptitude–treatment interaction approach to transfer within training. *Journal of Educational Psychology, 88*(2), 249-259. doi:10.1037/0022-0663.88.2.249

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

Helfenstein, S. (2005). *Transfer: Review, reconstruction, and resolution* University of Jyväskylä.

ISOSTAT charter. (n.d., October 17, 2012). Retrieved from http://www.lawrence.edu/fast/ jordanj/isostat_charter.html/

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*, 391-406. doi:10.1111/j.1745-3984.2008.00071.x

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54*(3), 1-11.

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician, 64*(1), 52-58.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1), 1-20. Retrieved from http://www.amstat.org/publications/jse/v10n1/mills.html

Moore, D. S. (2007). Statistics among the liberal arts. *Journal of the American Statistical Association, 93*(444), 1253-1259.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429-434.

Perkins, D. N., & Salomon, G. (1988). Teaching for transfer. *Educational Leadership, 46*(1), 22-32.

Pfannkuch, M., & Wild, C. (2005). Towards an understanding of statistical thinking. *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17-46) Springer.

R Core Team. (2014). [R: A language and environment for statistical computing] (3.1.1 ed.). Vienna, Austria: R Foundation for Statistical Computing.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(1), 106-125.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*(5), 529-556.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*(1), 1-15.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Rossman, A. J., & Chance, B. L. (2001). *Workshop statistics: Discovery with data* (2nd ed.). Emeryville, CA: Key College Publishing.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569-600. doi:10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist, 24*(2), 113-142.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*

Schau, C., & Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *The American Statistician, 51*(2), 171-175.

Scheaffer, R. L. (1997). Discussion. *International Statistical Review, 65*(2), 156-158. doi:10.1111/j.1751-5823.1997.tb00396.x

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

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science, 26*, 127-140.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312.

Thorndike, E. L., & Woodworth, R. S. (1901a). The influence of improvement in one mental function upon the efficiency of other functions (I). *Psychological Review, 8*(3), 247.

Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in education and psychology* (8th ed.). Boston: Pearson.

Thorndike, E. L., & Woodworth, R. S. (1901b). The influence of improvement in one mental function upon the efficiency of other functions: III. functions involving attention, observation and discrimination. *Psychological Review, 8*(6), 553-564. doi:10.1037/h0071363

Thorndike, E. L., & Woodworth, R. S. (1901c). The influence of improvement in one mental function upon the efficiency of other functions: II. the estimation of magnitudes. *Psychological Review, 8*(4), 384-395. doi:10.1037/h0071280

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1) Retrieved from http://www.amstat.org/publications/jse/v19n1/tintle.pdf

Watson, J. M. (1997). Assessing statistical thinking using the media. *The Assessment Challenge in Statistics Education, ,* 107-121.

Wild, C. J., Pfannkuch, M., & Regan, M. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society, 174*(2), 247-295.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248. doi:10.1111/j.1751-5823.1999.tb00442.x

Wong, R. M., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction*, *12*(2), 233-262.

## Appendix A: Test Blueprint Prior to Expert Feedback

### Test Blueprint: Cognitive Transfer Outcomes for Introductory Statistics

### Introduction to Cognitive Transfer & Motivation for Developing an Instrument

Singley and Anderson (1989) defined cognitive transfer to concern "how knowledge acquired in one situation applies (or fails to apply) in other situations." Salomon and Perkins (1989) described processes that produce transfer. One type of transfer called **high road transfer** requires a deliberate consideration of abstract cognitive elements (i.e. skills, concepts, definitions) previously mastered (Cox, 1997; Salomon & Perkins, 1989). High-road transfer can be further divided into two types: **1) forward-reaching transfer** where students generalize abstract ideas for an undetermined future use **2) backward-reaching transfer** where abstraction consists of an intentional search of available schema for relevant cognitive elements that may be applied to a task at hand (Salomon & Perkins, 1989).

Based on a review of current literature, much can be done to promote and assess successful cognitive transfer outcomes for students of introductory statistics. However, no published assessment currently exists to measure this specific outcome, and the literature gives reason for uncertainty about whether cognitive transfer can be achieved and measured following an introductory statistics curriculum.

### Development of an Instrument

The goal of my dissertation research is to explore the feasibility of developing an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. The **primary outcomes** that the instrument will measure include:

(1) **discernment** of whether statistical inference is appropriate for a problem setting, and

(2) **demonstration of high-road transfer** in a novel problem setting.

The instrument will consist of 6-8 scenarios with one or more tasks worth 2-4 points each. Most tasks will be constructed response (open-ended), though some may be forced response (multiple choice) where appropriate. Table 1summarizes the distribution of assessment items among all combinations of *discernment* and *transfer* mechanism characteristics.

Table 1
*Example of items classified by assessment goals*

| Discernment Required? | Transfer Mechanism: Forward-Reaching | Transfer Mechanism: Backward-Reaching | Column Total |
|---|---|---|---|
| Yes, statistical inference appropriate | 2 | 2 | 4 |
| Yes, no statistical inference required | 1 | 1 | 2 |
| No | 0 | 1 | 1 |
| Row Total | 3 | 4 | 7 |

The item characteristics in the margins of Table 1 are further defined below followed by descriptions of the six possible item types corresponding to each cell of the table.

## 1. Item Characteristic Definitions

Definitions of the item characteristics identified as "Discernment Required?" and "Inferential Strategy" in the margins of Table 1follow. Example items corresponding to each type are shown in section 3.

**1.1 Forward-Reaching High Road Transfer (2-4 items):** Students are given abstract principles (e.g. concepts, methods, ideas), and then asked to invent a novel application.

**1.2 Backward-Reaching High Road Transfer (4-5 items):** Students are given a specific problem setting, and then asked to describe or demonstrate relevant abstract principles (e.g. concepts, methods, ideas).

**1.3 Discernment Required – Statistical Inference Appropriate (3 items):** Students must recognize applications that do benefit from statistical inference.

**1.4 Discernment Required – No Statistical Inference Required (2 items):** Students must recognize applications that do not benefit from statistical inference.

**1.5 No Discernment Required (2-3 items):** Some items in the instrument will not include a component intended to assess the discernment outcome; they contribute to measurement of high-road transfer only.

## 2. Description of the Six Item Types in Table 1

A second goal of the instrument is to measure the ability of students to demonstrate high-road transfer when faced with novel problem settings that warrant statistical inference. Expectations regarding assessment of this goal follow.

**2.1 Forward-reaching high-road transfer with discernment—statistical inference appropriate.** Example item shown in section 3.1.

- Item presents a set of abstract principles related to statistical inference
- Students propose a scenario consistent with the given abstract principles in which statistical inference is appropriate
- Students must explain how the abstract principles relate to their proposed application

**2.2 Forward-reaching high-road transfer with discernment—no statistical inference required.** Example item shown in section 3.2.

- Item presents abstract principles that preclude statistical inference
- Students propose a scenario consistent with the given abstract principles that does not require statistical inference
- Students must explain how the abstract principles relate to their proposed application

**2.3 Forward-reaching high-road transfer with no discernment required.**

By definition, forward-reaching high-road transfer requires that students are given one or more abstract principles, and then they are instructed to describe a novel application. When the student is asked to describe an application of statistical inference, then the student has exercised discernment in choosing an appropriate application. The same is true if the student is asked to describe an application that does not require statistical inference. Consequently, forward-reaching high-road transfer by definition must include some measure of discernment. Therefore, the assessment instrument will include no forward-reaching high-road transfer items with no discernment required.

**2.4 Backward-reaching high-road transfer with discernment—statistical inference appropriate.** Example item shown in section 3.3.

- Content for statistical items will be typical of the first course in statistics at the undergraduate level
- Students must determine that statistical inference should be applied in the described scenario, and explain why (see section 3.3, task A)
- Students propose a detailed strategy for conducting statistical inference appropriate for the given context, but need not actually conduct the analysis (see section 3.3, task B)

**2.5 Backward-reaching high-road transfer with discernment—no statistical inference required.** Example item shown in section 3.4.

- Content will include data-driven scenarios typical of the first course in statistics at the undergraduate level for which statistical inference is not required (e.g., known or knowable population parameter, deterministic outcome)
- Students must determine that statistical inference is not required in the described scenario, and explain why (see section 3.4, task A)
- Students propose a detailed strategy to evaluate the given context without statistical inference, but need not actually conduct the analysis (see section 3.3, task B)
- Tasks may also ask students to identify a modification to the problem setting that would warrant statistical inference

**2.6 Backward reaching high-road transfer with no discernment**. Example item shown in section 3.5.

- Students are given a problem setting and explicitly instructed to use statistical inference
- Students propose a detailed strategy for conducting statistical inference relevant to each research question, but need not actually conduct the analysis

## 3. Item Examples

### 3.1 Example 1: Forward-reaching high-road transfer with discernment—statistical inference appropriate

1. The underlying principle of all statistical inference is that one uses sample statistics to learn something about the unknown population parameters. Convince me that you understand this statement by writing one or two paragraphs describing a situation in which you might use a sample statistic to infer something about a population parameter. Clearly identify the sample, population, statistic, and parameter in your example. Be as specific as possible, and do not use any example we have discussed in class.

Figure 1: Example item described by Chance (2002).

Assessment Item Characteristics

- Instrument objectives assessed
  - Discernment—statistical inference appropriate
  - Forward-reaching high-road transfer
- Scoring considerations
  - Open-ended task suitable for objective scoring (e.g., compare to checklist)
  - Full credit awarded for a response that
    - Proposes a context for which statistical inference is appropriate (discernment)
    - identifies the sample, population, statistic, and parameter in the context of the proposed context (transfer).
  - Partial credit awarded for responses that
    - conflate or misidentify the sample, population, statistic, or parameter
    - properly identify the above concepts yet fail to describe them within the context they have proposed

### 3.2 Example 2: Forward-reaching high-road transfer with discernment—no statistical inference required

2. The underlying principle of all statistical inference is that one uses sample statistics to learn something about the unknown population parameters. However, statistical inference (e.g., confidence intervals, hypothesis tests, etc.) is not required when the value of the population parameter is known (e.g., data represent the entire intended population).

Write a short paragraph describing a situation and accompanying research question for which you might collect data to address the research question, yet statistical inference is

not required. Describe how you would analyze the data to address the research question. Be as specific as possible, and do not use any example we have discussed in class.

Figure 2: Example forward-reaching transfer item that does not require statistical inference.

Assessment Item Characteristics

- Instrument objectives assessed
  - Discernment—No statistical inference required
  - Forward-reaching high-road transfer
- Scoring considerations
  - Open-ended task scored against a rubric
  - Full credit awarded for responses that describe:
    - An appropriate research question (discernment)
    - Data-based context for which the population parameter of interest can be known (transfer)
  - Partial credit awarded for responses that include references to
    - Randomness/Sampling variability
    - Generalizability

### 3.3 Example 3: Backward-reaching high-road transfer with discernment—statistical inference appropriate

3. Some people who have a good ear for music can identify the notes they hear when music is played. One note identification test consists of a music teacher choosing one of seven notes (A, B, C, D, E, F, G) at random and playing it on the piano. The student is asked to name which note was played while standing in the room facing away from the piano so that he cannot see which note the teacher plays on the piano.

Suppose you want to determine whether the student has a "good ear for music" using this note identification test.

A.) Recall that the underlying principle of all statistical inference is that a sample statistic is used to learn something about the unknown population parameter. Could statistical inference be used to determine whether the student has a "good ear for music"? Explain why you could or could not use statistical inference in this scenario.

B.) Explain how you would decide whether the student has a good ear for music using the note identification test. (*Be sure to give enough detail that a classmate could easily follow your method.)*

Figure 3: Example item adapted from a MOST instrument described by Garfield et al. (2012).

Assessment Item Characteristics

- Instrument objectives assessed
  - Task 3A: Discernment—statistical inference appropriate
  - Task 3B: Backward-reaching high-road transfer
- Scoring considerations
  - Task 3A
    - Open-ended task scored against a rubric
    - Satisfactory responses should:
      - Recommend use of statistical methods
      - Acknowledge the role of chance and randomness in determining whether a student has a good ear for music
    - A response that does not acknowledge randomness would be unsatisfactory
  - Task 3B
    - Open-ended task suitable for objective scoring (e.g., compare to checklist)
    - Full credit awarded for responses that describe:
      - an acceptable chance or null model
      - accommodation for sampling variability
      - appropriate test statistic
      - a method to generate a *p*-value and/or confidence interval
      - significance level and/or confidence level
    - response requirements are intended to have sufficient generality to accommodate a simulation-based or non-simulation-based approach without penalty

### 3.4 Example 4: Backward-reaching high-road transfer with discernment—no statistical inference required

4. Micron Technologies manufactures customized laptop computers for its customers by assembling various parts such as circuit boards, processors, and display screens purchased in bulk from other companies. Micron Technologies has placed a bulk order of 50 display screens from ScreenPro Manufacturing. Based on the contract between the two companies, Micron Technologies may choose to either accept the entire bulk order of 50 display screens, or reject the entire bulk order of 50 display screens for a refund.

It is a simple task for a trained engineer to determine whether an individual display screen is good or bad, and the contract agreement permits Micron Technologies to inspect each individual display screen before deciding whether to accept or reject the whole order.

A.) Recall that the underlying principle of all statistical inference is that one uses sample statistics to learn something about the unknown population parameters. Could statistical inference be used to determine whether Micron Technologies should accept or reject the order of display screens? Explain why you could or could not use statistical inference in this scenario.

B.) In either case, explain how you would decide whether Micron Technologies should accept or reject the order of display screens. (*Be sure to give enough detail that a classmate could easily follow your method.)*

Figure 4: Example item that does not require statistical inference.

Assessment Item Characteristics

- Instrument objectives assessed
  - Task 4A: Discernment—No statistical inference required
  - Task 4B: Backward-reaching high-road transfer
- Scoring considerations
  - Task 4A
    - Open-ended task suitable for objective scoring using a checklist
    - Full credit awarded for a response that concludes that statistical inference is not recommended because the population parameter of interest is (or can be) known
    - Partial credit awarded for responses that recommend statistical inference, yet justifies why the engineer should not inspect all 50 display screens
  - Task 4B
    - Open-ended task suitable for subjective scoring (e.g., compare to rubric)
    - If the student has <u>not</u> advocated for statistical inference in 4A
      - Full credit awarded for a response that requires the engineer to inspect all 50 display screens and reject the order if the proportion of bad display screens is too high (e.g., a criterion set by Micron Technologies)
      - Partial credit awarded for a response that describes statistical inference, or fails to reference acceptance criteria
    - If the student <u>has</u> advocated for statistical inference in 4A
      - Full credit for 4B will be awarded for describing
        - an acceptable chance or null model
        - accommodation for sampling variability
        - appropriate test statistic
        - a method to generate a *p*-value and/or confidence interval

- significance level and/or confidence level

### 3.5 Example 5: Backward reaching high-road transfer with no discernment

Students are given a data set accompanied by a short explanation of the data and how they were collected. Students are then asked to propose one viable research question that could be investigated using the provided data as well as a strategy that you have learned in class to address your question using statistical inference. NOTE TO STUDENT: Just explain how to conduct the statistical analysis; you do not need to do the analysis.

Assessment Item Characteristics

- Instrument objectives assessed
  - Discernment objective is not assessed by this item
  - Backward-reaching high-road transfer
- Scoring considerations
  - Open-ended task scored against a rubric
  - Full credit awarded for responses that
    - identify a viable research question relevant to the problem setting
    - propose a corresponding strategy to conduct a statistical analysis in each case (students are not required to actually perform the analysis)
  - Partial credit awarded for responses that
    - Fail to identify two appropriate research questions
    - Identify appropriate research questions, but fail to propose a corresponding strategy for statistical analysis in each case

## Appendix B: Expert Feedback Questionnaire Accompanying Test Blueprint

Dear [Reviewer],

I am truly grateful that you agreed to review my test blueprint and assist my research into the feasibility of developing an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. The test blueprint has been provided separately in both MS Word and PDF for your convenience. The first three pages of the test blueprint introduce the learning outcomes and describe the item types. The remainder of the document presents several examples corresponding to each item type described. Please review the test blueprint and complete the feedback questionnaire by providing your responses following each question. Specific instructions for recording your responses are provided for set of items.

When you are finished, please email the completed questionnaire as a MS Word or PDF document by **October 31, 2014**. Your review is very important to me, so if you aren't able to send me your feedback by that date, please let me know when you think you would be able to provide your feedback.

Thank you again for the generosity of your participation, and I look forward to your feedback! Please don't hesitate to contact me if you have any questions.

Sincerely,

Matthew D. Beckman
University of Minnesota
beckm109@umn.edu
612-655-5235

**Directions for Questions 1 & 2:** Please select the response that best reflects your opinion for each question, and then explain your answer typing your feedback in the space provided. In order to mark a checkbox, double-click on the chosen box (☐) and select "checked" (☒).

1. After completing an introductory statistics course, how important or unimportant is it that students be able to discern when a problem setting outside of class would or would not benefit from a statistical approach?

☐ Not Important ☐ Somewhat Unimportant ☐ Somewhat Important ☐ Important

Please explain your answer.

2. After completing an introductory statistics course, how important or unimportant is it that students be able to apply the statistical knowledge they have learned to novel problem settings outside of class?

☐ Not Important ☐ Somewhat Unimportant ☐ Somewhat Important ☐ Important

Please explain your answer.

**Item Characteristic Definitions (Questions 3-7)**

**Directions:** After reviewing the *Item Characteristic Definitions* in section 1 of the test blueprint, please indicate whether or not each of the definitions is clear. Please comment on unclear definitions by typing your feedback in the space provided following each prompt. Use as much space as you like when providing feedback.

3. Forward-Reaching High Road Transfer (Section 1.1). Is the definition clear?

   ☐ Yes

   ☐ No

   If not, how could the definition be improved?

4. Backward-Reaching High Road Transfer (Section 1.2). Is the definition clear?

   ☐ Yes

   ☐ No

   If not, please explain.

5. Discernment Required – Statistical Inference Appropriate (Section 1.3). Is the definition clear?

   ☐ Yes

   ☐ No

   If not, please explain.

6. Discernment Required – No Statistical Inference Required (Section 1.4). Is the definition clear?

   ☐ Yes

   ☐ No

   If not, please explain.

7. No Discernment Required (Section 1.5). Is the definition clear?

   ☐ Yes

   ☐ No

   If not, please explain.

**Item Type Descriptions (Questions 8-13)**

**Directions.** After reviewing the *Descriptions of the Six Item Types* in Section 2 of the test blueprint, please reflect on the following two questions and type your feedback in the space following each prompt. Use as much space as you like to respond to each prompt.

- Is the description clear?
- Does this item type seem important?

8. Forward-reaching high-road transfer with discernment—statistical inference appropriate. (Section 2.1)

   Is the description clear? ☐ Yes ☐ No
   If not, please explain.

   Is this item type important? ☐ Yes ☐ No
   If not, please explain.

9. Forward-reaching high-road transfer with discernment—no statistical inference required. (Section 2.2)

   Is the description clear? ☐ Yes ☐ No
   If not, please explain.

   Is this item type important? ☐ Yes ☐ No
   If not, please explain.

10. Forward-reaching high-road transfer with no discernment required. (Section 2.3)

    Is the description clear? ☐ Yes ☐ No
    If not, please explain.

11. Backward-reaching high-road transfer with discernment—statistical inference appropriate. (Section 2.4)

    Is the description clear? ☐ Yes ☐ No
    If not, please explain.

    Is this item type important? ☐ Yes ☐ No
    If not, please explain.

12. Backward-reaching high-road transfer with discernment—no statistical inference required. (Section 2.5)

    Is the description clear? ☐ Yes ☐ No
    If not, please explain.

    Is this item type important? ☐ Yes ☐ No
    If not, please explain.

13. Backward reaching high-road transfer with no discernment. (Section 2.6)

    Is the description clear? ☐ Yes ☐ No
    If not, please explain.

    Is this item type important? ☐ Yes ☐ No
    If not, please explain.

**General Feedback (Questions 14-16)**

**Directions.** Please type your feedback in the space provided following each prompt. Use as much space as you like to respond to each prompt.

14. Do the item examples generally seem to align well with the definitions, descriptions, and intended learning outcomes? Please explain by referencing specific examples.

15. Do you feel that anything is incomplete or missing from the test blueprint?

16. Please share your overall evaluation of the test blueprint as well as any general comments that you have about this project.

## Appendix C: Final Test Blueprint

### Test Blueprint: Cognitive Transfer Outcomes for Introductory Statistics

### Introduction to Cognitive Transfer & Motivation for Developing an Instrument

Singley and Anderson (1989) defined cognitive transfer to concern "how knowledge acquired in one situation applies (or fails to apply) in other situations." Salomon and Perkins (1989) described processes that produce transfer. One type of transfer called **high road transfer** requires a deliberate consideration of abstract cognitive elements (i.e. skills, concepts, definitions) previously mastered (Cox, 1997; Salomon & Perkins, 1989). High-road transfer can be further divided into two types: **1) forward-reaching transfer** where students generalize abstract ideas for an undetermined future use **2) backward-reaching transfer** where abstraction consists of an intentional search of available schema for relevant cognitive elements that may be applied to a task at hand (Salomon & Perkins, 1989).

Based on a review of current literature, much can be done to promote and assess successful cognitive transfer outcomes for students of introductory statistics. However, no published assessment currently exists to measure this specific outcome, and the literature gives reason for uncertainty about whether cognitive transfer can be achieved and measured following an introductory statistics curriculum.

### Development of an Instrument

The goal of my dissertation research is to explore the feasibility of developing an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. The **primary outcomes** that the instrument will measure include:

(1) **discernment** of whether statistical inference is appropriate for a problem setting, and

(2) **demonstration of high-road transfer** in a novel problem setting.

The instrument will consist of 6-8 scenarios with one or more tasks worth 2-4 points each. Most tasks will be constructed response (open-ended), though some may be forced response (multiple choice) where appropriate. Table 1summarizes the distribution of assessment items among all combinations of *discernment* and *transfer* mechanism characteristics.

Table 1

*Example of items classified by assessment goals*

| | Transfer Mechanism | | |
|---|---|---|---|
| Discernment Required? | Forward-Reaching | Backward-Reaching | Column Total |
| Yes, statistical inference appropriate | 2 | 2 | 4 |
| Yes, no statistical inference required | 1 | 1 | 2 |
| No | 0 | 1 | 1 |
| Row Total | 3 | 4 | 7 |

The item characteristics in the margins of Table 1 are further defined below followed by descriptions of the six possible item types corresponding to each cell of the table.

## 3. Item Characteristic Definitions

Definitions of the item characteristics identified as "Discernment Required?" and "Inferential Strategy" in the margins of Table 1follow. Example items corresponding to each type are shown in section 3.

**1.1 Forward-Reaching High Road Transfer (2-4 items):** Students are given abstract principles (e.g. concepts, methods, ideas), and then asked to invent a novel application.

**1.2 Backward-Reaching High Road Transfer (4-5 items):** Students are given a specific problem setting, and then asked to describe or demonstrate relevant abstract principles (e.g. concepts, methods, ideas).

**1.3 Discernment Required – Statistical Inference Appropriate (3 items):** Students must recognize applications that do benefit from statistical inference.

**1.4 Discernment Required –Statistical Inference not Appropriate (2 items):** Students must recognize applications that do not benefit from statistical inference.

**1.5 No Discernment Required (2-3 items):** Students must demonstrate high-road transfer, but the student does not need to recognize whether or not statistical inference is appropriate because the problem makes it clear whether or not the answer should include statistical inference.

## 4. Description of the Six Item Types in Table 1

A second goal of the instrument is to measure the ability of students to demonstrate high-road transfer when faced with novel problem settings that warrant statistical inference. Expectations regarding assessment of this goal follow.

**2.1 Backward-reaching high-road transfer with discernment—statistical inference appropriate.** Example item shown in section 3.3.

- Content for statistical items will be typical of the first course in statistics at the undergraduate level
- Students must determine that statistical inference should be applied in the described scenario, and explain why (see section 3.3, task A)
- Students propose a detailed strategy for conducting statistical inference appropriate for the given context, but need not actually conduct the analysis (see section 3.3, task B)

**2.2 Backward-reaching high-road transfer with discernment—statistical inference <u>not</u> appropriate.** Example item shown in section 3.4.

- Content will include data-driven scenarios typical of the first course in statistics at the undergraduate level for which statistical inference <u>is not</u> required (e.g., known or knowable population parameter, deterministic outcome)
- Students must determine that statistical inference is <u>not</u> required in the described scenario, and explain why (see section 3.4, task A)
- Students propose a detailed strategy to evaluate the given context without statistical inference, but need not actually conduct the analysis (see section 3.3, task B)
- Tasks may also ask students to identify a modification to the problem setting that would warrant statistical inference

**2.3 Backward reaching high-road transfer with no discernment**. Example item shown in section 3.5.

- Students are given a problem setting and explicitly instructed to use statistical inference
- Students propose a detailed strategy for conducting statistical inference relevant to each research question, but need not actually conduct the analysis

**2.4 Forward-reaching high-road transfer with discernment—statistical inference appropriate.** Example item shown in section 3.1.

- Item presents a set of abstract principles related to statistical inference
- Students choose or propose a scenario consistent with the given abstract principles in which statistical inference is appropriate
- Students must explain how the abstract principles relate to their proposed application

**2.5 Forward-reaching high-road transfer with discernment— statistical inference <u>not</u> required.** Example item shown in section 3.2.

- Item presents abstract principles that preclude statistical inference
- Students choose or propose a scenario consistent with the given abstract principles that does not require statistical inference
- Students must explain how the abstract principles relate to their proposed application

**2.6 Forward-reaching high-road transfer with no discernment required.**

By definition, forward-reaching high-road transfer requires that students are given one or more abstract principles, and then they are instructed to describe a novel application. When the student is asked to describe an application of statistical inference, then the student has exercised discernment in choosing an appropriate application. The same is true if the student is asked to describe an application that does not require statistical inference. Consequently, forward-reaching high-road transfer by definition must include some measure of discernment. Therefore, the assessment instrument will include no forward-reaching high-road transfer items with no discernment required.

## 3. Item Examples

### 3.1 Example 1: Forward-reaching high-road transfer with discernment—statistical inference appropriate

1. The underlying principle of all statistical inference is that one uses sample statistics to learn something about the unknown population parameters. Convince me that you understand this statement by writing one or two paragraphs describing a situation in which you might use a sample statistic to infer something about a population parameter. Clearly identify the sample, population, statistic, and parameter in your example. Be as specific as possible, and do not use any example we have discussed in class.

Figure 1: Example item described by Chance (2002).

Assessment Item Characteristics

- Instrument objectives assessed
  - Discernment—statistical inference appropriate
  - Forward-reaching high-road transfer
- Scoring considerations
  - Open-ended task suitable for objective scoring (e.g., compare to checklist)
  - Full credit awarded for a response that
    - Proposes a context for which statistical inference is appropriate (discernment)
    - identifies the sample, population, statistic, and parameter in the context of the proposed context (transfer).
  - Partial credit awarded for responses that
    - conflate or misidentify the sample, population, statistic, or parameter
    - properly identify the above concepts yet fail to describe them within the context they have proposed

### 3.2 Example 2: Forward-reaching high-road transfer with discernment—statistical inference <u>not</u> appropriate

2. The underlying principle of all statistical inference is that one uses sample statistics to learn something about the unknown population parameters. However, statistical inference (e.g., confidence intervals, hypothesis tests, etc.) <u>is not</u> required when the value of the population parameter is known (e.g., data represent the entire intended population).

Write a short paragraph describing a situation and accompanying research question for which you might collect data to address the research question, yet statistical inference <u>is</u>

not required. Describe how you would analyze the data to address the research question. Be as specific as possible, and do not use any example we have discussed in class.

Figure 2: Example forward-reaching transfer item that does not require statistical inference.

Assessment Item Characteristics

- Instrument objectives assessed
  - Discernment—No statistical inference required
  - Forward-reaching high-road transfer
- Scoring considerations
  - Open-ended task scored against a rubric
  - Full credit awarded for responses that describe:
    - An appropriate research question (discernment)
    - Data-based context for which the population parameter of interest can be known (transfer)
  - Partial credit awarded for responses that include references to
    - Randomness/Sampling variability
    - Generalizability

### 3.3 Example 3: Backward-reaching high-road transfer with discernment—statistical inference appropriate

3. Some people who have a good ear for music can identify the notes they hear when music is played. One note identification test consists of a music teacher choosing one of seven notes (A, B, C, D, E, F, G) at random and playing it on the piano. The student is asked to name which note was played while standing in the room facing away from the piano so that he cannot see which note the teacher plays on the piano.

Suppose you want to determine whether the student has a "good ear for music" using this note identification test.

A.) Recall that the underlying principle of all statistical inference is that a sample statistic is used to learn something about the unknown population parameter. Could statistical inference be used to determine whether the student has a "good ear for music"? Explain why you could or could not use statistical inference in this scenario.

B.) In either case, explain how you would decide whether the student has a good ear for music using the note identification test. (*Be sure to give enough detail that a classmate could easily follow your method.)*

Figure 3: Example item adapted from a MOST instrument described by Garfield et al. (2012).

Assessment Item Characteristics

- Instrument objectives assessed
  - Task 3A: Discernment—statistical inference appropriate
  - Task 3B: Backward-reaching high-road transfer
- Scoring considerations
  - Task 3A
    - Open-ended task scored against a rubric
    - Satisfactory responses should:
      - Recommend use of statistical methods
      - Acknowledge the role of chance and randomness in determining whether a student has a good ear for music
    - A response that does not acknowledge randomness would be unsatisfactory
  - Task 3B
    - Open-ended task suitable for objective scoring (e.g., compare to checklist)
    - Full credit awarded for responses that describe:

        - an acceptable chance or null model
        - accommodation for sampling variability
        - appropriate test statistic
        - a method to generate a *p*-value and/or confidence interval
        - significance level and/or confidence level
    - response requirements are intended to have sufficient generality to accommodate a simulation-based or non-simulation-based approach without penalty

### 3.4 Example 4: Backward-reaching high-road transfer with discernment—statistical inference <u>not</u> appropriate

4. Micron Technologies manufactures customized laptop computers for its customers by assembling various parts such as circuit boards, processors, and display screens purchased in bulk from other companies. Micron Technologies has placed a bulk order of 50 display screens from ScreenPro Manufacturing. Based on the contract between the two companies, Micron Technologies may choose to either accept the entire bulk order of 50 display screens, or reject the entire bulk order of 50 display screens for a refund.

It is a simple task for a trained engineer to determine whether an individual display screen is good or bad, and the contract agreement permits Micron Technologies to inspect each individual display screen before deciding whether to accept or reject the whole order.

A.) Recall that the underlying principle of all statistical inference is that one uses sample statistics to learn something about the unknown population parameters. Could statistical inference be used to determine whether Micron Technologies should accept or reject the order of display screens? Explain why you could or could not use statistical inference in this scenario.

B.) In either case, explain how you would decide whether Micron Technologies should accept or reject the order of display screens. (*Be sure to give enough detail that a classmate could easily follow your method.)*

Figure 4: Example item that does not require statistical inference.

Assessment Item Characteristics

- Instrument objectives assessed
  - Task 4A: Discernment—No statistical inference required
  - Task 4B: Backward-reaching high-road transfer
- Scoring considerations
  - Task 4A
    - Open-ended task suitable for objective scoring using a checklist
    - Full credit awarded for a response that concludes that statistical inference is not recommended because the population parameter of interest is (or can be) known
    - Partial credit awarded for responses that recommend statistical inference, yet justifies why the engineer should not inspect all 50 display screens
  - Task 4B
    - Open-ended task suitable for subjective scoring (e.g., compare to rubric)

- If the student has not advocated for statistical inference in 4A
  - Full credit awarded for a response that requires the engineer to inspect all 50 display screens and reject the order if the proportion of bad display screens is too high (e.g., a criterion set by Micron Technologies)
  - Partial credit awarded for a response that describes statistical inference, or fails to reference acceptance criteria
- If the student has advocated for statistical inference in 4A
  - Full credit for 4B will be awarded for describing
    - an acceptable chance or null model
    - accommodation for sampling variability
    - appropriate test statistic
    - a method to generate a *p*-value and/or confidence interval
    - significance level and/or confidence level

### 3.5 Example 5: Backward reaching high-road transfer with no discernment

Students are given a data set accompanied by a short explanation of the data and how they were collected. Students are then asked to propose one viable research question that could be investigated using the provided data as well as a strategy that you have learned in class to address your question using statistical inference. NOTE TO STUDENT: Just explain how to conduct the statistical analysis; you do not need to do the analysis.

Assessment Item Characteristics

- Instrument objectives assessed
    - Discernment objective is not assessed by this item
    - Backward-reaching high-road transfer
- Scoring considerations
    - Open-ended task scored against a rubric
    - Full credit awarded for responses that
        - identify a viable research question relevant to the problem setting
        - propose a corresponding strategy to conduct a statistical analysis in each case (students are not required to actually perform the analysis)
    - Partial credit awarded for responses that
        - Fail to identify appropriate research question
        - Identify appropriate research questions, but fail to propose a corresponding strategy for statistical analysis in each case

## Appendix D: Draft I-STUDIO Version Prior to Expert Feedback

The I-STUDIO assessment tool and associated scoring rubric are available by request from the author or advisors.

## Appendix E: Expert Feedback Questionnaire Accompanying Draft I-STUDIO Assessment Tool

Dear Dr. [Reviewer],

I have taken the feedback that I received on my test blueprint and have used the results to develop an assessment tool. The tool is called the Introductory Statistics Transfer of Understanding and Discernment Outcomes (I-STUDIO) assessment tool. I am truly grateful that you agreed to review these materials and assist my research into the feasibility of developing an assessment tool for the purpose of quantifying cognitive transfer outcomes for introductory statistics students. The I-STUDIO assessment has been provided separately in both MS Word and PDF along with a copy of the revised test blueprint for your convenience. The actual assessment will be delivered to students electronically. It begins with an IRB-approved consent form, followed by directions to students, and then the items appear on subsequent pages.

Please review the I-STUDIO assessment and complete the feedback questionnaire by providing your responses following each question. Specific instructions for recording your responses are provided for each set of items. When you are finished, please email the completed questionnaire as a MS Word or PDF document by **December 12, 2014**. Your review is very important to me, so if you aren't able to send me your feedback by that date, please let me know when you think you would be able to provide your feedback.

Thank you again for the generosity of your participation, and I look forward to your feedback! Please don't hesitate to contact me if you have any questions.

Sincerely,

Matthew D. Beckman
University of Minnesota
beckm109@umn.edu
612-655-5235

**Assessment Front-Matter**

**Directions:** Please comment by typing your feedback in the space provided following each prompt. Use as much space as you like when providing feedback.

1. Are the **directions** on page 3 clear?

   ☐ Yes

   ☐ No

   How could the directions be improved?

**Item Feedback**

**Directions:** After reviewing each item in the I-STUDIO assessment, please indicate whether or not the item aligns with the intended characteristic(s). You may want to refer to the test blueprint while responding, in particular, the item characteristic definitions (section 1) and description of item types (section 2). Please comment on each item by typing your feedback in the space provided following each prompt. Use as much space as you like when providing feedback.

2. Does **item 1** (Walleye Fishermen) align with the characteristics described in the test blueprint for Backward-reaching high-road transfer with discernment—statistical inference appropriate?

   ☐ Yes

   ☐ No

   How could the item be improved?

3. Does **item 2** (Note Identification) align with the characteristics described in the test blueprint for Backward-reaching high-road transfer with discernment—statistical inference appropriate?

   ☐ Yes

   ☐ No

   How could the item be improved?

4. Does **item 3** (Micron Technologies) align with the characteristics described in the test blueprint for Backward-reaching high-road transfer with discernment—statistical inference **NOT** appropriate?

   ☐ Yes

   ☐ No

   How could the item be improved?

5. Does **item 4** (Air Traffic Control) align with the characteristics described in the test blueprint for Backward reaching high-road transfer with no discernment?

   ☐ Yes

   ☐ No

   How could the item be improved?

6. Does **item 5** (Underlying Principle of Inference) align with the characteristics described in the test blueprint for Forward-reaching high-road transfer with discernment—statistical inference appropriate?

   ☐ Yes

   ☐ No

   How could the item be improved?

7. Does **item 6** (Inference Not Appropriate) align with the characteristics described in the test blueprint for Forward-reaching high-road transfer with discernment—statistical inference **NOT** appropriate?

   ☐ Yes

   ☐ No

   How could the item be improved?

8. Does **item 7** (Matched Pairs Design) align with the characteristics described in the test blueprint for Forward-reaching high-road transfer with discernment—statistical inference appropriate?

   ☐ Yes

   ☐ No

   How could the item be improved?

**General Feedback.**

**Directions:** Please select the response that best reflects your opinion for each question, and then explain your answer typing your feedback in the space provided. In order to mark a checkbox, double-click on the chosen box (☐) and select "checked" (☒). Please type open-ended remarks in the area provided following each prompt. Use as much space as you like to respond to each prompt.

The goals of the I-STUDIO assessment are to measure:

- Discernment of whether statistical inference is appropriate for a problem setting, and

- Demonstration of high-road transfer in novel problem settings.

9. Think about a student that has completed an introductory course in statistical methods, to what extent do you agree or disagree that the I-STUDIO assessment measures whether students would be able to discern whether statistical inference is appropriate for problem settings outside of class?

☐ Agree ☐ Somewhat Agree ☐ Somewhat Disagree ☐ Disagree

Please explain your answer.

10. Think about a student that has completed an introductory course in statistical methods, to what extent do you agree or disagree that the I-STUDIO assessment measures whether students would be able to demonstrate high-road transfer in novel problem settings outside of class?

☐ Agree ☐ Somewhat Agree ☐ Somewhat Disagree ☐ Disagree

Please explain your answer.

11. Please share anything you feel is missing from the I-STUDIO assessment.

12. Please share any general comments that you have about this project.

## Appendix F: I-STUDIO Version for Cognitive Interviews

The I-STUDIO assessment tool and associated scoring rubric are available by request from the author or advisors.

## Appendix G: I-STUDIO Version for Field Test

The I-STUDIO assessment tool and associated scoring rubric are available by request from the author or advisors.

## Appendix H: I-STUDIO Draft Scoring Rubric

***Introductory Statistics Transfer of Understanding and Discernment Outcomes (I-STUDIO) Assessment Rubric***

The I-STUDIO assessment tool and associated scoring rubric are available by request from the author or advisors.

## Appendix I: I-STUDIO Final Scoring Rubric for Field Test

***Introductory Statistics Transfer of Understanding and Discernment Outcomes (I-STUDIO) Assessment Rubric***

The I-STUDIO assessment tool and associated scoring rubric are available by request from the author or advisors.

## Appendix J: I-STUDIO Scoring Rubric Use Instructions

# I-STUDIO Scoring Rubric Use Instructions

**Recommendations for consistent application of scoring rubric**

It is recommended to complete the entire sequence of instructions for the intended scenario prior to scoring any student response under the following conditions;

- Upon beginning a new scoring session
- After a break from scoring that lasts longer than 20 minutes
- At regular intervals within a scoring session for a given task
    - Every $10^{th}$ student for the first 20 students
    - Every $20^{th}$ student thereafter
- Upon switching to begin scoring a new task within a scoring session

Upon opening scoring spreadsheet, “hide” columns A through E in order to obscure course information prior to scoring any responses.

**Scoring instructions by scenario**

- Scenario 1 (ATC Preparation): Prompt A & Prompt B
    - Read entire scenario and prompts A & B
    - Study rubrics and accompanying sample responses for
        - **q1a_rubric**
        - **q1b_rubric**
        - **q1b_redundancy**
    - For each student, enter <u>all</u> of the following in the scoring spreadsheet before moving on to the next student
        - **q1a_score**
        - **q1b_score**
        - **q1b_redundancy_score**

- Scenario 1 (ATC Preparation): Prompt C
    - Read entire scenario and prompts A, B, & C
    - Study rubric and accompanying sample responses for **q1c_rubric**
    - Enter **q1c_score** in the scoring spreadsheet

- Scenario 2 (Note Identification): Prompt A
    - Read entire scenario and prompt A
    - Study rubric and accompanying sample responses for **q2a_rubric**
    - Enter **q2a_score** in the scoring spreadsheet

- Scenario 2 (Note Identification): Prompt B
    - Read entire scenario and prompts A & B
    - Study rubric and accompanying sample responses for **q2b_rubric**
    - Enter **q2b_score** in the scoring spreadsheet

- Scenario 3 (Bulk Electronics): Prompt A
    - Read entire scenario and prompt A
    - Study rubric and accompanying sample responses for **q3a_rubric**
    - Enter **q3a_score** in the scoring spreadsheet

- Scenario 3 (Bulk Electronics): Prompt B
    - Read entire scenario and prompts A & B
    - Study rubric and accompanying sample responses for **q3b_rubric**
    - Enter **q3b_score** in the scoring spreadsheet

- Scenario 4 (Walleye Fishing): Prompt A
    - Read entire scenario and prompt A
    - Study rubric and accompanying sample responses for **q4a_rubric**
    - Enter **q4a_score** in the scoring spreadsheet

- Scenario 4 (Walleye Fishing): Prompt B
    - Read entire scenario and prompts A & B
    - Study rubric and accompanying sample responses for **q4b_rubric**
    - Enter **q4b_score** in the scoring spreadsheet

- Scenario 5 (Matched Pairs Study)
    - Read entire scenario and associated prompts
    - Study rubric and accompanying sample responses for **q5_rubric**
    - For each student, enter <u>all</u> of the following in the scoring spreadsheet before moving on to the next student
        - **q5_scenario_score_01**
        - **q5_participants_score_01**
        - **q5_treatments_score_01**
        - **q5_response_score_01**
        - **q5_pairing_score_01**
        - **q5_analysis_score_01**
        - **q5_interpretation_score_01**
        - **q5_lacksReplication_score_01**

- Scenario 6 (Underlying Principle of Statistical Inference)
    - Read entire scenario and associated prompts
    - Study rubric and accompanying sample responses for **q6_rubric**
    - For each student, enter all of the following in the scoring spreadsheet before moving on to the next student
        - **q6_scenario_score_01**
        - **q6_question_score_01**
        - **q6_population_score_01**
        - **q6_sample_score_01**
        - **q6_statistic_score_01**
        - **q6_parameter_score_01**
        - **q6_biasedSample_score_01**

- Scenario 7 (Statistical Inference NOT Required)
    - Read entire scenario and associated prompts
    - Study rubric and accompanying sample responses for **q7_rubric**
    - For each student, enter all of the following in the scoring spreadsheet before moving on to the next student
        - **q7_scenario_score_01**
        - **q7_question_score_01**
        - **q7_parameter_score_01**
        - **q7_population_score_01**
        - **q7_data_score_01**
        - **q7_analysis_score_01**