Designing an Evidence-based Assessment of Conceptual Understanding and Misunderstandings in Statistics

BY
NATALIE JORION
B.A., University of California, San Diego, La Jolla, CA, 2006
M.A., Northwestern University, Evanston, IL, 2009

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Learning Sciences
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Defense Committee:

James Pellegrino, UIC Learning Sciences and Psychology, Chair and Advisor
Alison Castro-Superfine, UIC Learning Sciences and MSCS
Mara Martinez, UIC Learning Sciences and MSCS
Yue Yin, UIC Educational Psychology
William Stout, University of Illinois at Urbana- Champaign

ACKNOWLEDGEMENTS

I would like to thank my advisor, Jim Pellegrino, for his support and guidance during my studies, as well as for giving me the opportunity to work on meaningful projects that inspired this research. Thanks to each of my other committee members, Yue Yin, Mara Martinez, Alison Castro-Superfine, and Bill Stout, for providing feedback on my initial drafts. Special thanks to Brian Gane and Lou DiBello for their mentorship and advice. Thanks to all of the graduate and post-doc students who provided comments about the preliminary draft of the assessment: Wenjuan Li, Mariya Yukhymenko, Carlos Salas, Hillary Rowe, and Gregory Bartoszek. A big thank you to my family for their encouragement throughout my long years of schooling. Finally, thanks to the hundreds of students I have had the privilege of teaching: through them, I have learned to look again at what I thought I knew with new eyes and new understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

LIST OF ABBREVIATIONS

AKs	Additional knowledge, skills, and abilities
AP	Advanced Placement
DCI	Dynamics Concept Inventory
DCM	Diagnostic classification modeling
CATS	Concept Assessment Tool for Statics
CAOS	Comprehensive Assessment of Outcomes in a first Statistics Course
CBAL	Cognitively Based Assessment of, for, and as Learning
CCSS	Common Core State Standards
CFA	Confirmatory factor analysis
CFs	Characteristic features
CI	Concept inventory
ECD	Evidence-centered design
EFA	Exploratory factor analysis
FK	Focal Knowledge
GAISE	Guidelines for Assessment and Instruction in Statistical Education
IRT	Item response theory
MTurk	Amazon Mechanical Turk
POs	Potential Observations
PWs	Potential Work Products
SCI	Statistics Concept Inventory
STEM	Science, technology, engineering, and math
VFs	Variable Features

SUMMARY

This study investigates the extent to which an assessment can help diagnose student conceptual understanding and misconceptions in the domain of statistics. A Statistics Concept Inventory (StatCI) was created using an evidence-centered design framework (Mislevy, Steinberg, & Almond, 2003). This assessment can serve as evidence for three claims about student performance regarding (1) overall understanding of statistical concepts, (2) understanding of specific concepts, and (3) propensity for misconceptions or errors. The researcher drew from a comprehensive literature review of student thinking in statistics to structure the conceptual domain of introductory statistics. This structure informed item creation; clusters of items correspond to a major concept, and each distractor maps onto a misconception.

The researcher conducted multiple studies to investigate the validity evidence for the assessment. For the first study, student think-alouds served to indicate how students interpreted items and what prior knowledge they leveraged to answer the items. Using this information, the researcher edited the assessment items to minimize misunderstandings and eliminate construct-irrelevant variance. For study 2A, the edited assessment was administered to 100 participants on Amazon Mechanical Turk. The researcher analyzed the response data, which helped to identify problematic items. Because the reliability for the scores was low (α=.58) and several items had higher alpha-if-item-deleted measures (α=.66 with seven items taken out), the researcher decided to conduct an additional study before the large-scale administration. Study 2B involved administering the modified assessment to 150 participants to further help validate the inventory. For the last study, the updated assessment was administered to 750 participants. Participant performance data was analyzed for response patterns demonstrating conceptual and errorful thinking. In particular, the researcher investigated the performance data by items, conceptual structure, distractors, and

demographical groups. These results serve as evidence that the assessment is measuring the targeted constructs and is able to identify learner misconceptions and errors. The outcomes of this program of research include: (1) a design pattern template that can be broadly applied to create other assessments in statistics; (2) a final assessment instrument that can be used in undergraduate first-year statistics courses; and (3) a methodology for applying ECD to concept inventory design.

I. INTRODUCTION

Statistical understanding has become an increasingly important skill in the 21st century. The influx of digital data has afforded richer and more varied information to interpret than ever before. This has created a need for analysts able to apply statistics adeptly, as well as consumers capable of critically interpreting statistics. A recent survey of university faculty found that professors see statistics as one of the most important areas of math for undergraduates to learn across disciplines (Conley, Drummond, Gonzalez, Rooseboom, Stout, 2011). High-quality statistics instruction has become essential not only for students seeking careers in statistics, but for everyone who uses data to support inferences and to make informed decisions.

However, helping students develop deep understanding of statistics is a challenging task. Students enter the classroom with misconceptions about variability and uncertainty, and often have spent their lifetimes using efficient but faulty heuristics (Tversky & Kahneman, 1974). Although instruction can help change students' statistical understanding, doing so does not necessarily reconcile these misconceptions. Even when teachers address learner misconceptions in the classroom, students may still entertain alternative mental models of phenomena (Shaughnessy, 2007).

Moreover, traditional instruments of classroom learning are generally insensitive to measuring student conceptual understanding. Students may attain high grades on exams and still entertain fundamental misconceptions about how to use and interpret statistics (Best, 1982). This discrepancy may in part be a result of instruction and assessments that stress the procedural aspects of statistics. Algorithmic understanding and rote memorization are relatively easy to assess with selected response items. In contrast, tapping into conceptual understanding may be difficult without using extended response items. In order to create selected response items that tap into conceptual

understanding, instructors must carefully consider plausible student errors. Because selected response items are faster to grade and provide more reliable measures, instructors teaching large classes may resort to assessing algorithmic thinking. Although this does not trivialize the importance of procedural knowledge, it also leaves out an essential part of conceptual understanding.

In addition, assessment feedback is comprised of total scores and incorrect answers, but generally does not show mastery of sub-domains or misconceptions. This situation is often true for both multiple choice and open response tests; even tests of conceptual understanding rarely have clear reporting systems showing patterns of student thinking (Jorion et al., 2014). Several professors have attempted to create statistics assessments of conceptual understanding. However, many of these assessments do not align with the developer's claims or do not report student misconceptions. There is therefore a need for reliable and valid statistics assessment tools that measure conceptual understanding and provide meaningful information on student errors.

This study describes the conceptualization and design of a statistics inventory based on the misconceptions literature. It details how the researcher will administer and analyze this assessment. The literature review is split up into three sections. The first section discusses the importance of aligning assessments with constructivist theories of learning. Perspectives on how learning takes place are summarized and the implications for assessment are provided. Three examples of assessments that align with constructivist theories are detailed. Amongst these three are concept inventories (CIs), assessments that are the focus of this study. CIs are assessments that seek to assess conceptual understanding and student misconceptions. A brief history of CIs is sketched out, along with issues of test validity.

The second section reviews statistics, the study's target domain, in terms of both content and pedagogy. This section problematizes statistical understanding and provides a comprehensive

domain review. Several existing assessments of statistical understanding are reviewed.

The third section lays the foundation for the study's theoretical framework. The study applies evidence-centered design (Mislevy, Steinberg, & Almond, 2003) to the creation of a new assessment in statistics.

The methods section follows this literature review and describes the participants and data collection process. The research consists of three studies. The first is a pilot study (n = 100) to identify items to retain and/or modify. The second (n=150) is an additional study to investigate item properties. The last study is an administration of the edited assessment on a larger sample (n = 750).

Overall, the goal of developing of this new inventory is to not only create a useful instrument appropriate to use within the classroom, but also to provide a method that can be generalized to the design of other instruments in the area of science, technology, engineering, and math (STEM) education.

II. LITERATURE REVIEW

2.1 Aligning Assessments with Constructivist Theories of Learning

Assessments are tools to help educators and learners make inferences about student understanding. Like any measurement instrument, these tools are useful the extent to which they accurately measure a clearly defined construct and convert this construct into interpretable information. Weighing scales, for example, are maximally useful when they determine the weight of an object and display this weight in calibrated units. Similarly, thermometers are maximally useful when they accurately detect a physical change with temperature and convert this change to an understandable numeric value. In both of these cases, weight and temperature are straightforward constructs with well-defined units of measurement that manifest physically with changes in the environment. Measuring learning, on the other hand, is more problematic than measuring weight or temperature because learning, the object of measurement, is a complicated construct. How do changes in learning manifest themselves? Does learning necessarily progress linearly and straightforwardly? In what ways can changes in learning be converted into actionable information? These are questions that concern learning scientists and educators, and they have a profound impact on how the measurement of learning should take place.

Historically, the role of classroom assessments has been limited to affirming whether students have attained knowledge in a specific domain. Generally, instructors use a classical test theory framework in assessment design, in which the educator reports the total number of correct answers. This number provides the educator and the learner an index of the degree to which the learner has attained subject mastery. Such information is akin to identifying the degree to which a patient is healthy or sick; the information tells little about the nature of the learner's problem. Total score on an assessment fails to capture many components of learning: it indicates nothing about what

the learner *does* know, what sub-concepts within the domain the learner has mastered, nor does it differentiate types of understanding. Such assessments treat learning as a process of acquiring information demonstrated through absent or present behaviors. It treats incorrect answers as a deficiency rather than as information.

Extensive research indicates that learning is a more nuanced process by which humans leverage and build upon their experiences. According to constructivist theory, learning is not a passive process focused on inputs and outputs, but rather an active process of constructing knowledge based on prior understandings (Bransford, Brown, Cocking, Donovan, & Pellegrino, 2000). Educators need to leverage students' preconceptions about a domain to help build on them. Although educators have successfully applied this paradigm to curriculum design, many are still using outdated models of learning for the design of assessments (Pellegrino, Chudowsky, & Glaser, 2001). This shortcoming is true not only of summative tests (tests *of* student learning), but also of formative tests (tests *for* student learning).

Like medical diagnoses, useful information from assessments about students' current understandings can only be derived when the symptoms of the learner are examined. Given that one of the fundamental purposes of education is to identify gaps in learner understanding to help students progress onto the path of mastery, it is important that tests should be designed to r fulfill this function. It is not just sufficient for educators to identify correct and incorrect student responses; they should also understand *why* a student might have answered incorrectly. In sum, assessments have the potential to provide much more information about student understanding than what they are currently doing.

To create more informative assessments requires several steps. First, learning, the construct of measurement, should be explicitly defined and grounded in empirical research—the model of

learner cognition. Second, since the instrument should be used in situations that elicit changes in learning, the tasks used for measurement should be built carefully to reflect the construct of interest and be instructionally sensitive. Third, the instrument should yield interpretable information relevant to the change in construct. Each step in this process should be refined and balanced as necessary in an iterative process.

2.1.1 The Assessment Triangle. An assessment is a tool designed to help make inferences about student understanding. For this tool to be effective in meeting its intended interpretive purpose, it must be coherent, which involves three interconnected elements: cognition, observation, and interpretation (Pellegrino et al., 2001). The *cognition* aspect defines the model of learning, such as the way in which a learner develops proficiency in a domain. The *observation* aspect defines specifications for the tasks that will be used to indicate learner competency. The *interpretation* aspect is the means for making sense of observations to make judgments about learner cognition. Each of these three components should be aligned with the other two. The application of this framework in the current study is explained in more detail below.

Cognition. In regards to the cognitive model, this study builds on a constructivist view that the path from novice to expert is one of knowledge reorganization built on prior understandings. Learning is not knowledge acquisition or replacement of one idea with another, but is instead the process of making sense of data using refined heuristics. Changes in conceptual understanding may not progress linearly. Learners may hold multiple contradicting ideas at once, which will be elicited by features of the situation.

Observation. Often in practice, students use different sets of skills and conceptual understandings than what assessment developers intend to measure. To create more reliable indicators of student conceptual understanding, developers should clearly define the construct to be

measured from the initial construction of the assessment. To connect the cognition aspects to the observation aspects explicitly, the study uses an evidence-centered design framework to create an assessment template (Mislevy & Haertel, 2006). This template defines the choices that must be made explicit in the assessment design process. It articulates the performance expectation for learners—that is, what kind of reasoning the students should be capable of and how they should apply this understanding. Such a template also facilitates item creation for future assessments of statistical understanding.

Interpretation. The models underpinning the cognition and observation elements in the assessment triangle framework have direct implications for the interpretation element. The interpretive element can help instructors make sense of student performance data from the assessment. Psychometric models can provide evidence regarding how well the student performance data supports the claims the assessment is supposed to satisfy. Four different approaches are used in the current study: classical test theory, item response theory, factor analysis (both exploratory and confirmatory), and diagnostic classification modeling. In addition, distractor analysis can support the formative use of assessments in the classroom.

2.1.2 Perspectives on learning and implications for assessment. Theories about learning have implications for assessment design. It is not enough to know whether students have attained subject mastery; educators also need to define what mastery means and know where students are on the path toward attainment. Where learners begin on this path toward subject mastery depends upon the learner's prior knowledge (Bransford et al., 2000). When the ideas based on prior knowledge do not agree with the formal targets of learning, it may be difficult to have learners reconcile their preconceptions with the normative disciplinary model. To help learners reconcile such differences, educators need to understand what prior experiences learners are leveraging. Assessment developers

need to consider what kind of intermediary steps students take on the path to developing subject mastery by providing contexts that elicit the use and misuse of common learner understanding. That is, items should be designed to provide opportunities for students to show what they do not yet understand. The importance of prior knowledge in learning is the basis of contemporary learning science theory (Bransford et al., 2000). Where learning scientists sometimes disagree is about how novice knowledge is organized and how conceptual change takes place.

There are two divergent views on how learning takes place. The knowledge-as-theory view posits that novice understanding is coherent and consistent across contexts (Chi, 2005; Ioannides & Vosniadou, 2002). Change takes place similar to Kuhn's (1970) paradigm shifts, in which a revolution occurs within a scientific community when the dominant paradigm fails to explain phenomena. Similarly, novices have models of how the world works based on prior encounters with phenomena. When too much data begin to contradict this model, the novice must devise a new model with more explanatory power, or reject the contradictory data completely. The alternative view is the knowledge-as-elements perspective. This theory posits that novice knowledge is fragmented, loosely connected, and highly sensitive to context (Roschelle, 1994). DiSessa (1998) refers to these fragments of knowledge as "p-prims." Accordingly, novices may entertain several conflicting views at once that are elicited by superficial cues of the situation; they may change tactics when the problem is varied. Increased sophistication in understanding leads to knowledge reorganization and refinement, perhaps reflecting a more coherent, model-based knowledge structure.

These two perspectives have implications for the design of assessments that purport to measure student understanding. If novice mental models followed the knowledge-as-theory view, developers could expect to see consistent misconceptions applied independently of context. On the

other hand, if novice thinking followed the knowledge-as-elements view, developers would expect context to have a significant impact on emergent misconceptions. Thissen-Roe, Hunt, & Minstrell (2004) found that physics students reason similarly across abstract concepts but may have inconsistent theories when presented with different situations. Students might even entertain several conflicting misconceptions at once. Designers, in this case, should present learners with different tasks that assess the same concept in order to fully characterize the properties of learner knowledge (Smith, diSessa, & Roschelle, 1994).

Another implication of these views is how educators should conceptualize change. Does learning progress linearly in predictable "steps" or does learning happen intermittently with possible regressions? A knowledge-as-theory view assumes learners have a cohesive understanding and progress logically into more sophisticated models. This would support the learning trajectory literature that defines certain behaviors associated with a unidimensional latent trait. Wilson (2005), for example, uses Rasch modeling to identify learning progressions, a method that assumes unidimensionality.

A knowledge-as-elements view makes the assumption that learners may not travel on a straightforward path to understanding. Minstrell's (2001) facets-based approach to student thinking assumes that there are several disparate strategies, procedures, or heuristics that students might use to solve a problem. These facets are not ordered sequentially by degree of understanding. In Sadler's (1998) study using the Astronomy Concept Inventory, he found that in some cases, students have a higher probability of answering incorrectly after limited instruction. If it is possible for students to regress during the process of learning, it may be more difficult to identify the extent of learning or to compare among students.

Which of the two views assessment developers ascribe to would impact the corresponding measurement model for evaluating learning. Particularly, these assumptions about learning impact whether to use a univariate or multivariate model when analyzing student data. Measurement models that correspond to a knowledge-as-theory view include normalized gain score and Rasch modeling. Normalized gain scores of pre- and post- test measures are calculated by the change in scores divided by the greatest possible increase (Hake, 1998). This model assumes that student scores will increase in correspondence to an increase in their univariate latent trait. Another method could use Rasch modeling to create construct maps (Wilson, 2005). Such maps can help to identify when students have "progressed" conceptually based on what kind of answers the learner chooses.

In line with the knowledge-as-elements view, Bao and Redish (2001) propose that students may hold alternate belief states simultaneously, such as Newtonian, Galilean, and Aristotelian. The extent to which each of these is present depends on a student's beliefs states and the features of the task. To investigate knowledge states of learners, Huang (2003) applied the Andersen/Rasch multivariate item response theory model on two physics tests. He found that after one semester of physics, students tend to be in a mixed model state, supporting the knowledge-as-elements view.

Experimental studies on conceptual change have resulted in contradictory findings. Vosniadou and Brewer (1994) found that elementary school children explained the day and night cycles using logically consistent mental models. Similarly, Ioannides and Vosniadou (2002) also interviewed elementary school children about the concept of force and found that students used coherent explanatory frameworks. DiSessa, Gillespie, and Esterly (2004) did a quasi-replication study of Ioannides and Vosniadou's research and found that students changed answers based on features of the interview question, casting doubt on the consistency of learner knowledge. Southerland, Abrams, Cummins, and Anzelmo (2001) interviewed second to twelfth grade students

on biological phenomena and found that many students provided differing and even sometimes contradictory explanations. However, they did not discount the knowledge-as-theory perspective; they said that student explanations showed evidence of scientific conceptions alongside p-prims.

Özdemir & Clark (2007) posit that the extent to which these perspectives are applicable may depend on several variables. First, the age of the learner may be a factor: knowledge-as-theory perspective may be more pervasive in younger students, but less so in older students. DiSessa, in particular, developed the knowledge-as-elements perspective by studying undergraduate students while Vosniadou developed the knowledge-as-theory perspective by studying elementary school students. Another possible factor is the richness of the scientific domain. For example, diSessa was investigating student explanations in physics, a domain with which students have substantial first-hand experience. Vosniadou, on the other hand, was investigating student thinking in astronomy, a domain with which students have much less first-hand experience interacting with the key concepts and phenomena.

This has several implications for the current study. Measuring misconceptions in the domain of statistics will likely be a challenging task. It is not clear whether students will have persistent misconceptions or multiple conflicting belief states dependent on the context. On the one hand, a knowledge-as-elements view may reflect how students learn statistics. Students who learn statistics are older— they have generally finished the first year of college. Since older students are more likely to have knowledge-as-elements, learners in this study will likely demonstrate inconsistent knowledge structures. Regarding the impact of the richness of the domain on novice knowledge structure, it is unclear to what extent learners activate prior knowledge when solving problems in statistics. Students use heuristics to make judgments about variance and uncertainty in their

everyday lives (Tversky & Kahneman, 1974). Moreover, statistics and corresponding graphics are reported in the media, so students may leverage this knowledge.

On the other hand, a knowledge-as-theory view might better represent how students understand statistics. Some students enter the statistics classroom entirely unfamiliar with hypothesis testing. As a result, their understanding of this concept will depend on instruction. Hypothesis testing and other statistical concepts can be fairly complex to the novice learner. Students might not subscribe to an incorrect mental model of phenomena, but instead have crude conceptual understanding that must be refined. Since statistics is not a tangible phenomenon, the investigator hypothesizes that it is less rich in this regard to other domains. This may contribute to more consistency in misconceptions across problem contexts.

For the current study, the investigator assumes a knowledge-as-theory perspective of statistical learning. Once the items have undergone an initial screening for validation (to ensure that the items do not have construct-irrelevant variance, for example), the investigator can then examine what implication student response patterns have with respect to the alternative theories of learning. Items within a conceptual category will be isomorphic to allow context to vary while holding the conceptual knowledge needed to answer the item correctly constant.

When examining student errors on the exam, the researcher will investigate the extent to which learners show patterns of errorful thinking across items using chi-squared statistics by dividing the sample into high and low ability examinees. Individual students who choose three or more distractors mapped onto the same error will indicate evidence of a misconception. The researcher can also use Bock's (1972) nominal response model to further investigate how distractors may link to latent states and learning trajectories. These analyses can help to make inferences about individual student understanding.

However, it is possible that examinees will not show a consistent patterns of responses corresponding to mapping of misconceptions. Such an analysis would provide evidence for a knowledge-as-elements view of learning, where students hold alternative beliefs simultaneously. The researcher could then use an Andersen/Rasch multivariate item response theory model to investigate how features of an item elicit different knowledge states. It may also be that because students are leveraging different forms of prior knowledge for each conceptual category, and thus they are responding with systematic misconceptions for one knowledge category and at random for another category. 2.1.3 Example of assessments aligned with constructivist theories of learning. Several sets of assessments have extended beyond the classical test theory framework and have been designed to align with more complex models of student cognition. Four example assessment systems are provided here: (1) Cognitive Tutors, (2) Cognitively Based Assessment *of, for,* and *as* Learning (CBAL), (3) Diagnoser, and (4) Concept Inventories (CIs). The last of these four will be the focus of this research and discussed in more detail in the next section.

Cognitive Tutors is a computer-based program in which students learn Algebra I at their own pace and take assessments with real-time feedback and hints. The program provides information to students and teachers about individual processes or misconceptions that caused an error. Each module is based on a model derived from a cognitive task analysis of the domain knowledge, and evaluates student performance based on a standard of domain expertise (Anderson, Corbett, Koedinger, & Pelletier, 1995). By gathering data about the student and the student's interactions with the problems, the program is able to monitor the learning process, perform a diagnosis of the current versus the expected state, and select optimal tutoring strategies. The designers' assumptions about learner cognition are that competence depends on both declarative and procedural knowledge. Declarative knowledge by itself is inert, and procedural knowledge depends on knowing declarative

knowledge. However, the latter can only be acquired by *doing*. For this reason, the practice component of the assessment is essential, and the evaluation is relative to the learning objectives. An example algebra item is presented in Figure 1. Students are instructed to diagram the problem in the lower left box and create a corresponding equation in the lower right box. The tutor scaffolds students as they work through the example, providing hints as necessary. The idea is to help lead students to conceptual understanding by working through procedural problems.

One component that seems absent in the Cognitive Tutors literature is the measurement model and validity argument. This may be because the assessments are used as learning tools rather than as summative assessments. Cognitive Tutor developers use standardized assessments to evaluate student knowledge gains when comparing Cognitive Tutor to more commonplace curricula (Koedinger, 2002).

Another example is CBAL, a set of assessments in reading, writing, and math, designed to serve synergistically as summative, formative, and dynamic assessments of student understanding (Bennett, 2010). CBAL is based on learning progressions research which posits that there is a novice to expert trajectory of understanding. Examinee scores are not just based on correct answers but also students' solution steps (Fife, 2013). In this way, the formative CBAL component can indicate not just *whether* students have attained domain mastery, but also *where* they fall along this path to mastery.

Each CBAL mathematics assessment consists of extended tasks made up of both selected and constructed response sub-items. All of the CBAL assessments are administered on the computer,

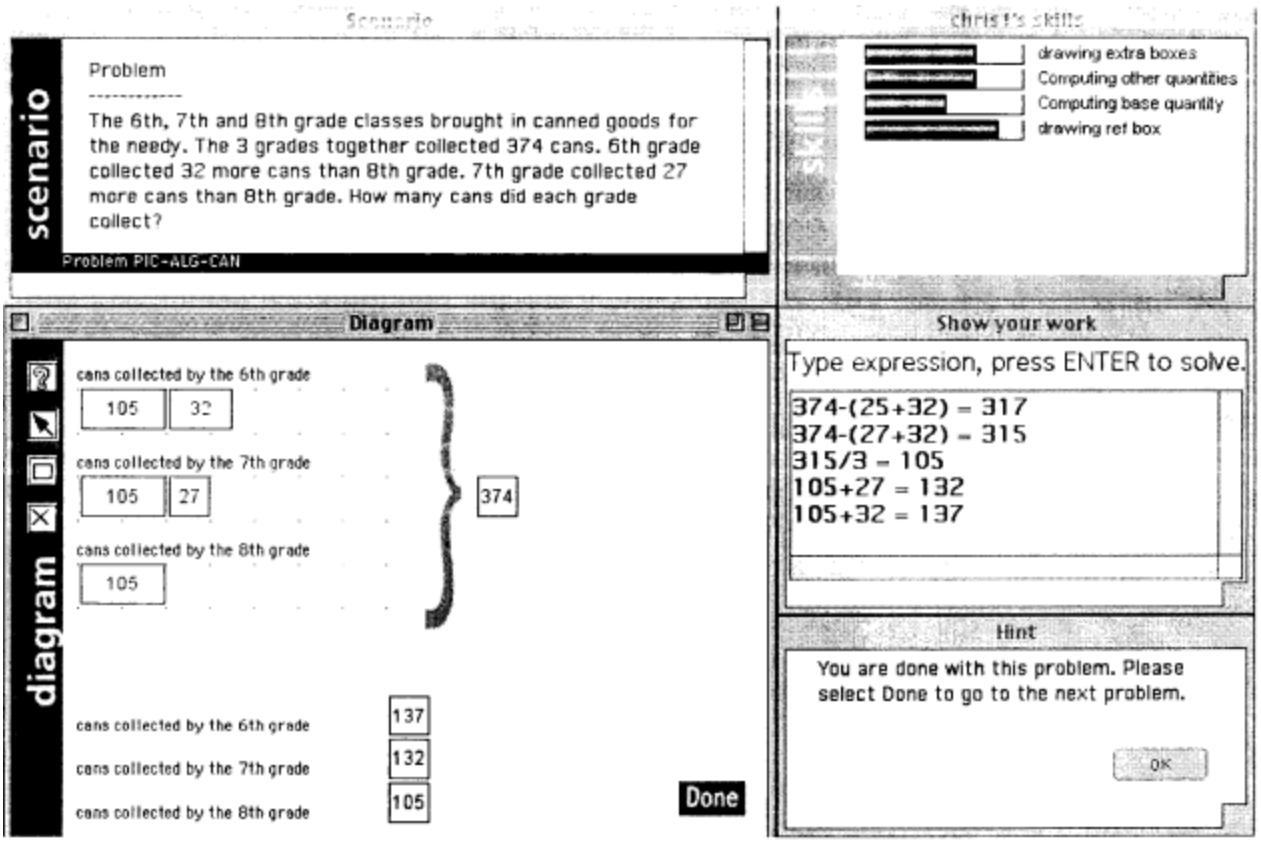


Figure 1. Example Cognitive Tutor Algebra item. Reprinted from *Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6* (p. 12), by K. Koedinger, 2002. Copyright 2002 by Kenneth Koedinger.

which makes it easier to collect detailed data about student responses, response time, and click patterns. Many items include a simulation based on a real-world scenario. Figure 2 contains such a simulation; students can mix different amounts of water and punch to see how proportions of each affect the sweetness of the punch. Constructed response items are scored based on a rubric, using both human and computer-based scoring. The developers of CBAL are currently researching ways to automate scoring, such as using M-rater for mathematical expressions and graphs (Fife, 2013).

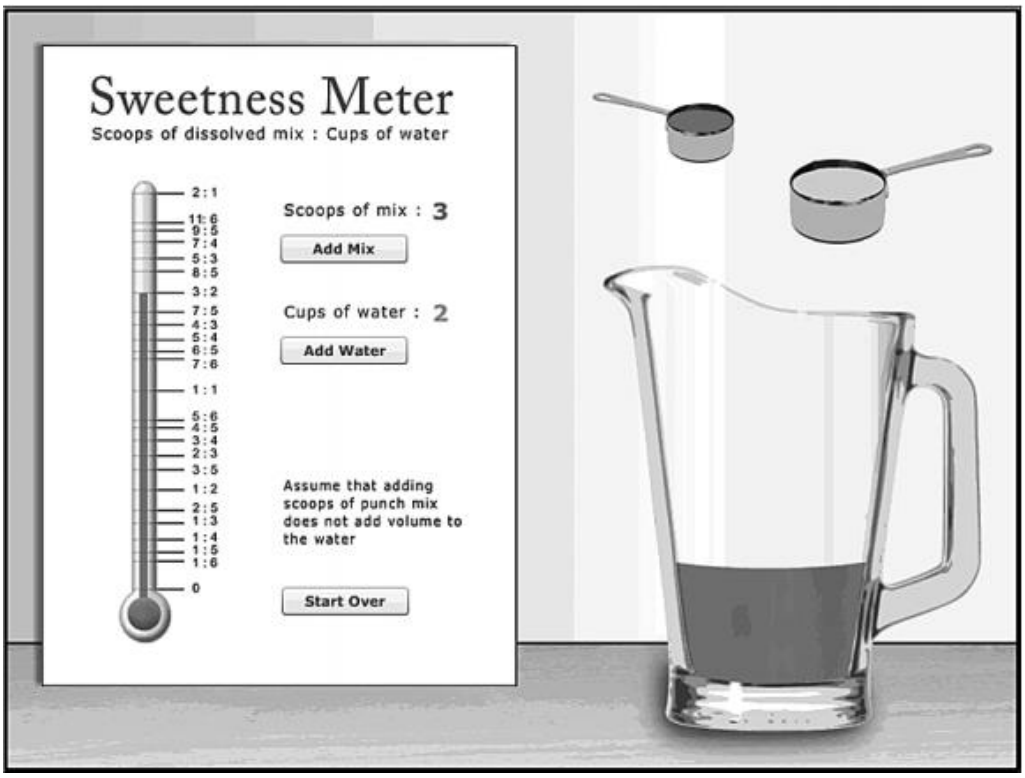


Figure 2. Example CBAL mathematics item. Reprinted from *Highlights from the Cognitively Based Assessment of, for, and as Learning (CBAL) Project in Mathematics* (p. 25), E. Graf, K. Harris, E. Marquez, J. Fife, & M. Redman, 2010, Princeton, NJ. Copyright © 2010 Educational Testing Service. www.ets.org

One finding is student interaction data can provide evidence on student understanding. Students have the option of using the simulation, such as the sweetness meter in Figure 2, in order to answer the item. For example, students who use the simulation but answer the item incorrectly may be at the lowest level of understanding; students who use the simulation and answer the item correctly may be at the next level of understanding; finally, students who do not use the simulation and answer correctly may be at the highest level of understanding (Arieli-Attali, 2013). One

challenge CBAL developers face in constructing relevant tasks based on real-world scenarios is minimizing construct-irrelevance, a major threat to assessment validity.

The third example is the Diagnoser system based on Minstrell's facets-based model (Hunt & Minstrell, 1994; Minstrell, 2000). Facets describe student thinking based on various types of reasoning, conceptual, and procedural difficulties. For example, they can be content specific, strategic, or logically generic. They are based on classroom observations and research on student thinking. Facet clusters are sets of these facets based around a conceptual idea. For example, one conceptual idea may be using forces to explain accelerated motion. The first facet in the cluster provides a conceptually appropriate example, and the second provides a more formulaic approach to the concept (e.g., facets 420 and 421 in

TABLE I). Other facets are problematic thinking that may arise from prior knowledge or formal instruction (e.g., facets 423 – 429 in

TABLE I). Each facet is an important intermediary step toward reaching the targeted standard, but they collectively do not necessarily represent an ordered group of checkpoints toward that goal. These facet clusters can help educators better identify where student understanding falls on one or more paths to conceptual understanding.

Diagnoser is a multiple-choice computerized assessment in which each answer is based on a particular facet. The items come in three parts. The first asks students to make a prediction about a situation. The second asks students to rate their confidence in their answer. The third asks to justify the first response. After each item, the program provides students with a "diagnosis" screen. Those

students who provide the correct answer receive encouragement; an incorrect but consistent answer prompts the Diagnoser to suggest other possibilities; and an incorrect and inconsistent answer

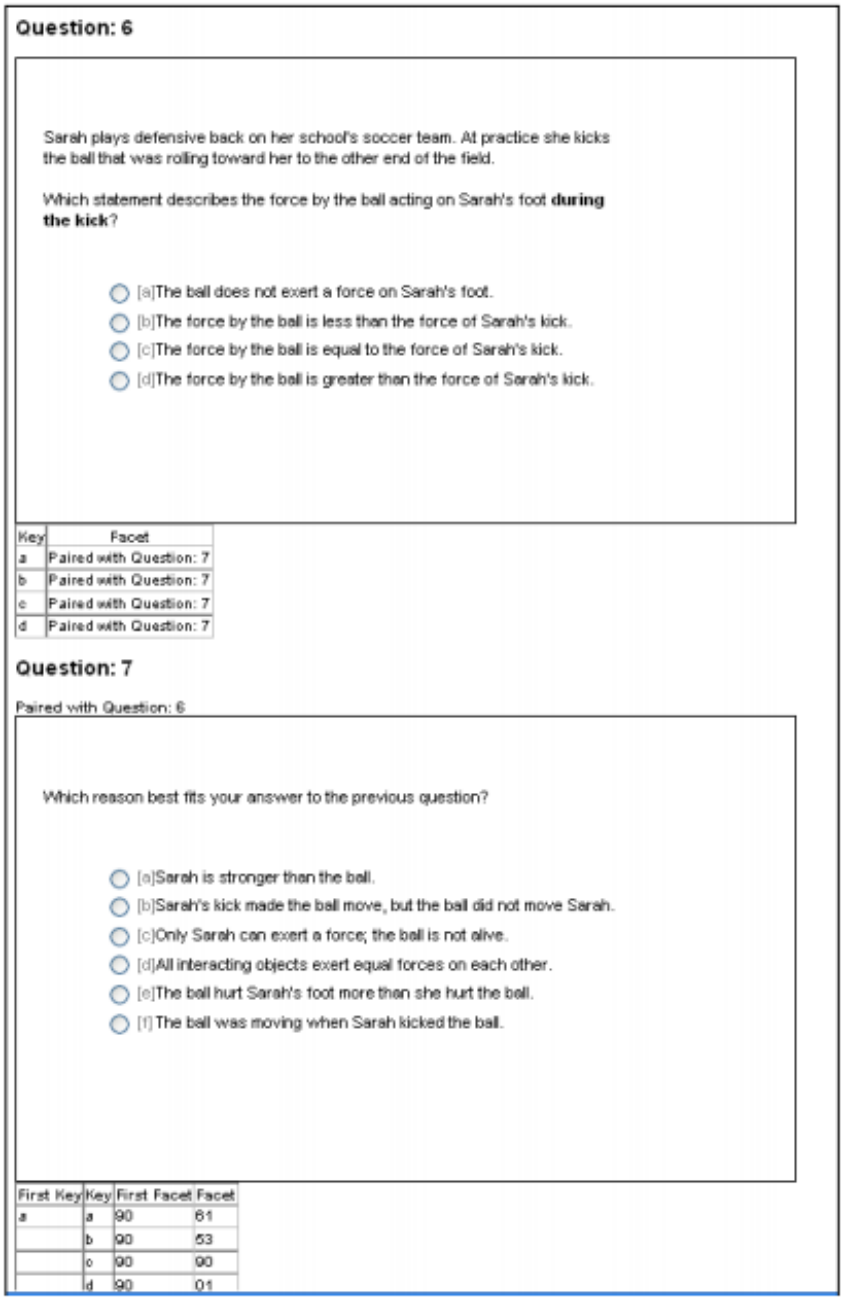
TABLE I

EXAMPLE FACET CLUSTER

Cluster 420: Forces to explain accelerated motion (Minstrell, 2001)	
*420	Acceleration is proportional to F_{net} and inversely proportional to mass.
*421	$A = F_{net} / m$
423	Acceleration is the result of any force applied to the object (not F_{net})
424	F_{net} depends upon potential (or on what's about to happen to it). If it is going to accelerate eventually, then it has a net force on it.
425	Objects do not have mass in space; they can accelerate without force.
426	Objects do not have "hold back" inertia in vertical situations, they just accelerate down.
428	Explaining any accelerated condition with an excess force proportional to the velocity acting on the object.
429	While the object is accelerating, it "has" force proportional to its velocity. Force is perceived as a property of the object.

will prompt the Diagnoser to point out the inconsistency. The program is embedded in instruction and used formatively to help students and instructors gauge student thinking. Figure 3 shows an example item pair from Diagnoser. Note that the second part of the item corresponds to option A. Diagnoser was not created with a certain measurement model in mind, but DeBarger et al. (2011) have used innovative psychometric methods to investigate aspects of the validity of this assessment system, including its measurement properties.

All three of these assessments have been designed with clearly defined constructs of student cognition. They also employ tasks that yield interpretable information about changes in learning and are instructionally sensitive. Despite their strengths, all three assessments have limitations given that they were created without explicit measurement models. This may be in part due to the fact that they were designed primarily for formative instructional uses. Thus, while they have schemes for



Question: 6

Sarah plays defensive back on her school's soccer team. At practice she kicks the ball that was rolling toward her to the other end of the field.

Which statement describes the force by the ball acting on Sarah's foot **during the kick**?

- [a]The ball does not exert a force on Sarah's foot.
- [b]The force by the ball is less than the force of Sarah's kick.
- [c]The force by the ball is equal to the force of Sarah's kick.
- [d]The force by the ball is greater than the force of Sarah's kick.

Key	Facet
a	Paired with Question: 7
b	Paired with Question: 7
c	Paired with Question: 7
d	Paired with Question: 7

Question: 7

Paired with Question: 6

Which reason best fits your answer to the previous question?

- [a]Sarah is stronger than the ball.
- [b]Sarah's kick made the ball move, but the ball did not move Sarah.
- [c]Only Sarah can exert a force; the ball is not alive.
- [d]All interacting objects exert equal forces on each other.
- [e]The ball hurt Sarah's foot more than she hurt the ball.
- [f]The ball was moving when Sarah kicked the ball.

First Key	Key	First Facet	Facet
a	a	90	61
	b	90	53
	c	90	90
	d	90	01

Figure 3. Example Diagnoser item pair. Reprinted from *Evaluating the Diagnostic Validity of a Facet-based Formative Assessment System* (p.7) by A. DeBarger, L. DiBello, J. Minstrell, M. Feng, W. Stout, J. Pellegrino, G. Haertel, C. Harris, & L. Ructinger, 2011, Washington, DC. Copyright 2011 by Jim Minstrell.

interpretation of student performance, they lack applications of measurement theory and methods to verify the adequacy of those interpretive schemes.

2.1.3 Concept Inventories: Past developments. One main issue facing STEM educators is that students often maintain erroneous beliefs even after formal instruction. Professors tend to stress procedural knowledge because it is easier to assess and assume their students have already mastered the necessary conceptual understandings. In order to facilitate conceptual change, professors need a tool to measure students' prior beliefs and understandings. Concept Inventories (CIs) were born out of this need. Generally, these are multiple-choice tests that purport to measure conceptual understanding, and often have distractors linked to particular student errors and misconceptions.

Instructors tend to use CIs as summative assessments or evaluative measures of educational interventions. However, CIs have the potential to play more formative roles in the classroom (Jorion et al., 2014). Teachers can use them to diagnose student understanding in order to provide targeted feedback and appropriate remedial instruction.

The Force Concept Inventory was the first CI to have a major impact in STEM education research. Developed by Hestenes and Halloun (1985), it and was designed to measure student conceptual understanding of Newtonian physics. The Force Concept Inventory consists of 30 multiple-choice items that do not require calculations to answer the items correctly. The items are based on 6 main conceptions and 30 misconceptions. An example conception is "kinematics," and a corresponding misconception for this category would be that position and velocity are the same. One of the main presuppositions the developers make about learner cognition is that all students have commonsense beliefs about the way the world works, which they leverage during formal learning of physics (Hestenes, Wells, & Swackhamer, 1992). Many of these beliefs are incompatible with Newtonian physics and resistant to change despite formal instruction. For example, one

commonsense understanding is that an inanimate intrinsic force keeps objects moving. A resulting misconception might be that force is supplied by a "hit," resulting in a loss of the original force. This contradicts Newton's first law of motion, which states that objects tend to rest or stay in motion with the same speed and direction unless acted upon by an external force. In addition, novices have difficulty differentiating concepts, causing confusion in word choice. For example, novices may refer to "velocity" and "acceleration" interchangeably.

2.1.4 Concept Inventories and issues of assessment validity. Since the creation of the Force Concept Inventory, more than 20 different CIs have been generated for use in STEM classrooms. The development of CIs is an important step in measuring more complex aspects of student understanding to help improve and inform teaching and learning. However, research has shown that CIs vary greatly in terms of their validity relative to intended uses (Jorion et al., 2014). It is not sufficient for developers to claim that their inventories function in a particular way; these claims need to be verified in terms of reliability and validity (Pellegrino et al., 2001).

Reliability is the degree to which the instrument is consistent. Ideally, an instrument should yield the same measurement under similar circumstances. One way to measure reliability is using Cronbach's alpha, which depends on the correlation among the items, the number of items, and the variance of scores on the entire test. This score can range from negative to 1, with negative to 0 values being not reliable and 1 being perfectly reliable. The better the exam can differentiate between examinees, the higher the alpha measure. Low inter-item correlation may indicate that students are guessing, using test-taking tricks, or using a different skill in general to answer the items. *Validity* is the degree to which the assessment aligns with the developers' claims about what the assessment should measure. This includes measuring the target construct and using the instrument in a particular context. Instrument validation is an argumentation process dependent on

the assessment's use, implementation, and interpretation; an instrument cannot be valid independent of its context of use (Kane, 2006; Kane, 2013).

There are several interrelated aspects of validity that may be applicable to assessment designed to support teaching and learning: cognitive, instructional, and inferential (Pellegrino, DiBello, Jorion, James, & Schroeder, 2013). *Cognitive validity* is the extent to which the test taps into disciplinary knowledge, which can be verified with expert domain analyses and student think-aloud studies. I*nstructional validity* is the extent to which the assessment supports teaching practice, demonstrated through teachers' interpretation and use of test scores in the classroom. Finally, *inferential validity* is the extent to which the assessment provides valid and reliable information about student performance. This can be investigated using psychometric analyses of student performance data. Factor analysis, for example, provides evidence regarding an assessment's construct validity.

Although CIs are designed by content experts, many fall short of adequately measuring student understanding (Pellegrino et al., 2013). Jorion et al. (2014) investigated the validity properties of three widely used inventories: the Statistics Concept Inventory (SCI; Allen, 2003), the Dynamics Concept Inventory (DCI; Gray et al., 2005), and the Concept Assessment Tool for Statics (CATS; Steif & Hansen, 2007).

The developers of these CIs made three main claims about how student scores on the inventories could provide evidence of (1) overall proficiency level, (2) understanding of sub-domains, and (3) propensity for misconceptions. Jorion et al. (2014) analyzed the degree to which the CI developers' claims about their assessment aligned with student performance data. In particular, they investigated overall reliability, individual item and sub-score functioning, and

distractor response patterns. Among the three CIs, the CATS was the only one to perform well on overall measures, and well as on individual item and sub-score functioning.

However, CI scores do not provide strong evidence that CIs can function as diagnostic tools of learner misconceptions. This may stem from lack of structural coherence. The DCI has too few items per category to make claims about student misconceptions. SCI analyses indicated poor construct validity for conceptual categories, resulting in weak evidence for misconceptions across items. The CATS was the only CI of the three to show evidence of construct validity. Further investigations of the CATS revealed that distractors within an item mapped onto the same misconception, making it difficult to differentiate student thinking across items. These analyses suggest that although CIs purport to measure misconceptions, even the CIs that demonstrate strong validity properties (overall and structurally) often fail to do so.

Learning theories may also explain why other inventories have been unsuccessful at measuring consistent mental models of phenomena. The Force Concept Inventory, for example, seems to assume a knowledge-as-theory view, even though the developers claim that commonsense knowledge can have situation-dependent meanings (Hestenes, Wells, Swackhamer, 1992). The developers present a taxonomy of misconceptions. Distractors for different items are mapped onto each respective misconception. This taxonomy implies that students will apply misconceptions consistently across items, even though the items differ in context. However, no published research has confirmed that the Force Concept Inventory can substantiate consistent learner misconceptions across items.

One inventory, however, has had success in measuring misconceptions. Sadler (1998) found that the Astronomy Concept Inventory could identify certain student misconceptions. Using Bock's (1972) nominal response model, he mapped distractors to student ability level. He included a "Don't

Know" category to provide students with an alternative to uninformed guessing. Most of the questions are relatively straightforward while tapping into important concepts. Example items include, "What causes night and day?" and "The main reason for it being hotter in summer than in winter is…". However, the inventory's success in identifying student misconceptions could be attributable to other reasons. First, students have relatively low "real life" interaction with astronomy as a domain; students learn astronomy through instruction and textbooks, and therefore their understanding might be more cohesive and less piece-meal, in line with a knowledge-as-theory perspective of learning. Second, the questions on the Astronomy Concept Inventory are posed on a fairly abstract level. The concepts, in these cases, are not applied across multiple situations or across different contexts. Claiming that students have robust misconceptions on a concept measured by only one item is problematic. Moreover, just because the data for one inventory indicates students hold robust misconceptions does not necessarily mean that it is possible to do so in all other domains.

Given that CIs are created to align with a constructivist view of learning, these tests should indicate prior student beliefs that influence formal learning. Why might the current CIs be failing to accomplish this task? First, to measure student understanding, items within a conceptual category should show evidence of clustering together in structural analyses; i.e., there should be evidence that students use similar skills on items within the same conceptual category. Since many CIs do not show strong evidence for measuring the constructs that developers claim (Jorion et al., 2014), it is difficult to conclude that incorrect answer choices among items are derived from similar misconceptions. Second, an instrument needs to identify clearly how specific distractors differ in terms of beliefs; otherwise, there is no way to assert with certainty that students picking different distractors hold the same misconception. Third, sample size may become an issue when conducting

analyses focused on distractor choices. Generally, the majority of students pick the correct answer and those who entertain a faulty belief do so amongst three or four different distractor choices. The resulting smaller sample size decreases the power of analyses focused on distractor selection. Fourth, it may be that the construct of learning is built upon a faulty assumption of learner cognition. If student understanding within a domain is dependent upon context, then it would be difficult to measure consistent conceptual understandings across different contexts. Although it may not be possible to account for each of these issues in the assessment design process, a carefully constructed assessment may avoid the majority of these issues.

To provide evidence that a student entertains a specific belief, an assessment should have distractors linked to a particular belief across multiple items. Choosing a distractor once does not yield enough evidence that a student entertains a particular belief. Students who do not understand the concepts but choose correctly are guessing, whereas those who understand but choose incorrectly may have a slip. In addition, contextual factors may elicit particular beliefs. Therefore, it is important that developers research and map out the cognitive landscape of the target domain to ensure that assessment items will be instructionally sensitive.

2.2 Conceptual Understanding in Statistics:

Application of a Principled Approach to Assessment Design

Although CIs are a step towards developing tests of important STEM learning outcomes, there is still work to be done in creating assessment instruments that are both reliable and valid indicators of student understanding in many STEM education areas including the very important domain of statistics. We are now living in an information age when data is ubiquitous and endless. More than ever, it has become essential to know how to make sense of data while understanding the concept of randomness. According to one survey of university professors, statistics is one of the most important subjects for undergraduates to understand across disciplines (Conley et al., 2011). Nonetheless, statistics remain a challenge to teach. Students struggle with statistical misconceptions resulting from the use of efficient but faulty heuristics for everyday reasoning (Tversky & Kahneman, 1974). Feelings of anxiety are commonly associated with learning statistics (Onwuegbuzie & Wilson, 2003). And too often, professors teach statistics as a set of recipes and procedures rather than interrelated concepts. Some of the difficulty may derive from the challenges of teaching and assessing conceptual understanding. There is therefore a need for additional research on conceptual understanding in statistics. This study sets itself apart from previous attempts at creating an assessment of statistical understanding in several ways.

First, this study uses evidence-centered design to ground the conceptual framework of the assessment (Mislevy et al., 2003). Each category for this new statistics inventory is based on a big idea and enduring understanding in statistics. Several items have been designed to correspond to each category. This correspondence is mapped out in a Q-matrix, a table specifying the item and concept relationships. These concepts are also mapped onto the Common Core Standards to ensure that they are aligned with the learning objectives of recent national standards. Second, the

distractors for each item are based on misconceptions documented in previous statistics education research. Several distractors across questions have been created that tap into the same misconception, which provides evidence of robust student misunderstandings. Such an assessment has the potential to play a variety of formative roles in the classroom, from helping to provide learners and teachers diagnostic feedback on student and classroom-level performance, to helping evaluate the effectiveness of a new curriculum.

In the next section the principled approach to assessment design summarized above is explained in considerable detail. The explication starts with a general description of the major components of an Evidence-Centered Design (ECD) approach. Each of the major components of ECD is then elaborated as it has been applied to the domain of statistics. The section ends with a detailed description of the inventory that has been constructed, which serves as the focus for the further empirical studies proposed as part of this dissertation project.

2.2.1 Components of Evidence-Centered Design. ECD is a systematic approach that focuses on the evidence of competence as a basis for constructing assessment tasks (Mislevy, et al., 2003; Mislevy & Haertel, 2006). The ECD framework is composed of a five-stage design: the Domain Analysis, Domain Modeling, the Conceptual Assessment Framework, the Assessment Implementation, and the Assessment Delivery. The first step in the design is the Domain Analysis, which involves gathering evidence about the target domain for the sake of assessment design. This includes reviewing research on the subject matter, such as concepts, terminology, and representations; the pedagogy; and examples of assessments currently used for this purpose.

The second step is Domain Modeling, which involves organizing the information from the Domain Analysis to make claims about those aspects of student proficiency that will be the target for assessment. In particular, this step specifies the focal knowledge, skills, and abilities that constitute

evidence of proficiency; what tasks and potential observations would make up this evidence; and what characteristic features would be expected as part of the design of such tasks.

The next stage is the Conceptual Assessment Framework, which lays out the blueprint for implementing the assessment. There are three components of the Conceptual Assessment Framework: the student model, the task model, and the evidence model. The student model shows what the assessment developer is trying measure in terms of student proficiencies. This could be a single measure of proficiency for a pass/fail summative assessment, or a multi-dimensional model of proficiency for a diagnostic assessment purported to provide detailed feedback on student understanding. The task model specifies the relevant features of the task and the variable features of the assessment items—that is, the student work products to provide evidence for the proficiencies. The evidence model bridges these other two models by providing information regarding what kind of student responses and observations would show evidence for these targeted competencies. The evidence model consists of evidence rules and a statistical model. The evidence rules translate the student work product into evaluative summaries of these products, such as one or more scores. The statistical model specifies the relationship between the evidence rule and the competency model. This allows for the competency model to be updated accordingly.

After the Conceptual Assessment Framework comes Assessment Implementation, in which the developer constructs the assessment tasks and scoring rules based on the Conceptual Assessment Framework blueprint. Assessment Delivery is the final step where students interact with the task. In addition, their performances are evaluated, and they receive feedback or test scores.

There are several advantages to using such a systematic design approach. For one, it supports the assessment's validity argument by clearly identifying the goals of learning, suitable observations to identify these goals, and tasks that would provide the grounds for these observations.

This allows for transparency when linking performance tasks to constructs, which is useful for accountability purposes. Additionally, a byproduct of the ECD process is the creation of an assessment blueprint that can be reused in different contexts.

It is worthwhile mentioning that there are alternative approaches to instrument design. One such method is construct mapping (Wilson, 2005). Such a map requires a coherent definition of the construct to be measured, as well as the various levels of proficiency from novice to expert learner. From this, the researcher would identify what would constitute more or less evidence of this proficiency. One assumption in this method is that there is a single underlying continuum of proficiency per construct. This is not necessarily the case for statistical reasoning, especially if we consider the knowledge-in-pieces theory of learning seriously. ECD allows for more flexibility in assumptions regarding intermediary states of knowledge. Therefore, the ECD approach was chosen as a more appropriate design framework to guide this research.

2.2.2 Domain Analysis. This section presents a domain analysis of six important and difficult concepts in statistics based on an extensive literature review. Each of these concepts is mapped onto the Common Core State Standards (CCSS). Prevalent learner misconceptions and other types of errorful thinking are integrated into each conceptual cluster. This domain analysis establishes the foundation for application of the other components of the ECD framework to produce the Statistics Concept Inventory that is the focus of this current study. Undergraduate statistics does not have a standardized curriculum, and therefore statistics courses vary across schools and professors in respect to topics, sequence, and pedagogical strategies. Recommendations for what to teach in statistics have also changed over time. The availability of statistical software now renders hand calculations of statistics unnecessary, making understanding and interpreting the statistics more important (Franklin et al., 2007). However, there is some overlap across courses. Three different

organizations have created a list of important topics in statistics: the Common Core State Standards (CCSS), the College Board for Advanced Placement (AP) statistics, and the American Statistical Association's Guidelines for Assessment and Instruction in Statistical Education (GAISE). All three sets of recommendations are for kindergarten to 12th grade students. In addition, GAISE has a set of recommendations for an introductory college course.

The CCSS details what students should know and be able to do in English language arts and math (Common Core State Standards Initiative, 2010). It seeks to establish consistent standards for those U.S. States that have adopted the standards. These standards are supposed to be more reflective of real-world skills and college requirements than previous curricula. In the statistics and probability section of the CCSS, there are four categories: (1) interpreting categorical and quantitative data, (2) making inferences and justifying conclusions, (3) conditional probability and the rules of probability, and (4) using probability to make decisions.

AP Statistics is a college-level course for high school students that covers four topics: (1) exploratory analysis of data using graphical and numerical techniques, (2) planning and conducting a study, (3) probability, and (4) statistical inference (College Board, 2010). Many of these topics correspond to those found in the CCSS.

The GAISE emphasize statistical literacy over statistical techniques, and therefore recommendations are not ordered by topics (Franklin et al., 2007). As per the suggestion of Moore (1997), relatively new topics such as hands-on data analysis are incorporated in these objectives at the expense of other topics, such as formal probability. Following are some of these recommendations ordered by concepts: (1) Understand that variability is quantifiable; (2) Know how sampling distributions applies to making statistical inferences based on sample data; (3) Understand the concept of statistical significance and *p*-values and how to make appropriate uses of statistical

inferences; (4) Know how to interpret numerical summaries and graphical displays of data; (5) Know how to interpret data on two variables and understand the difference between correlation and causation.

Converting these recommendations to key conceptual categories is not a straightforward task. For one, not all of these standards address deep conceptual understanding. Some of them recommend that students know how to use tools to derive answers such as, "Use calculators, spreadsheets, and tables to estimate areas under the normal curve" (1.4) or "Compute (using technology)… the correlation coefficient of a linear fit" (1.8) (CCSSI, 2010). As a result, not all the recommendations can be mapped onto relevant concepts for this Statistics Concept Inventory. In addition, the CCSS include probability as a core knowledge component, while the GAISE argues against its inclusion. Ultimately, the choice of topics is up to the discretion of the particular instructor and college program.

During the analysis of an existing Statistics Concept Inventory (Allen, 2006), a group of researchers proposed an alternative approach based on a different, conceptually-driven domain analysis. Following Backwards Design (Wiggins & McTighe, 2005) procedures that had been successfully used to generate a domain analysis in other areas of STEM, William Stout, a distinguished statistics professor from the University of Illinois at Urbana-Champaign, created a list of "big ideas and enduring understandings" in statistics (Stout, 2013). Seven researchers, all of whom were knowledgeable in statistics, helped to revise the Stout domain analysis document. The resultant domain analysis outlined eight important conceptual categories for a first-year statistics course (the unpublished domain analysis document is included as Appendix A):

1. Elementary probability, which involves reasoning probabilistically by applying conditional probability and independence.

2. Hypothesis testing, which involves knowing which test to use, setting up the test, and interpreting results.
3. Large sample theory results, including understanding the central limit theorem and the law of large numbers as they impact the large sample behavior of the sample average and how sample size relates to large sample theories.
4. Univariate data graphing, which involves interpreting graphs of various types.
5. Important statistical indices, such as understanding measures of central tendency and dispersion, and demonstrating the effect of outliers on these indices.
6. Confidence intervals, including knowing the probability modeling assumptions, sampling procedures, and sample size concerns that make each of the standard Confidence Interval procedures appropriate.
7. Elementary regression/correlation, which includes carrying out analyses and interpreting them, as well as graphing of bivariate data.
8. Sampling, which includes sampling requirements to produce data acceptable for effective statistical analysis.

The domain analysis document in Appendix A describes what students should know and be able to do to demonstrate proficiency in each category.

The preliminary domain analysis list for the StatCI inventory consisted of these eight categories. These categories were modified in several ways to be more relevant to first year undergraduate students. First, four professors at the University of Illinois at Chicago were asked to examine these categories and consider the extent to which they aligned with the typical curriculum for the first year statistics course in psychology. They unanimously indicated that confidence intervals were beyond the scope for first year statistics. This category was therefore taken out of the

concepts for the domain analysis. Second, the author judged graphing to be a secondary skill rather than a primary point of conceptual understanding, so this was not included among the list of targeted conceptual categories to guide design of the instrument.

The resulting conceptual categories for the current StatCI are (1) interpreting univariate data, (2) interpreting bivariate data, (3) applying probability, (4) understanding how sampling applies to statistical inferences, (5) understanding how large sample theory applies to statistical inferences, and (6) understanding and correctly interpreting statistical significance. Each category is briefly described below with respect to critical aspects of student knowledge and understanding, including typical misunderstandings that can be found in the research and instructional literature.

Univariate data. Univariate statistics are an essential first step in statistical analyses; they are methods for characterizing a large set of quantitative information for a single variable using measures of central tendency, dispersion, and graphs. Students should understand how each of these measures indicate different things about data and how to recognize each based on representations of data.

Students often have several misconceptions about how to best represent univariate data. Students may learn how to calculate the mean, median, and mode, and know the definition of each, but they may not know how to select the appropriate measure of central tendency given the data (Garfield & Chance, 2000). They may fail to consider outliers when computing the mean. In other words, students may know procedural differences in obtaining measures of central tendency, but they may not know conceptual differences in applying these measures to real data for purposes of representing and interpreting those data.

Students may also find the idea of variability conceptually challenging. They may think of the common definition of variability as “not consistent or not having a fixed pattern” rather than as

the spread of a distribution. Therefore, they may believe that a narrow, bumpy graph has more variability than a spread out, smooth graph. Students will often have experience calculating measures of central tendency in elementary and high school, but may still not understand conceptual differences between them as estimators of a dataset.

Bivariate data. Bivariate analysis involves interpreting the relationship between two variables, using measures such as correlation, regression, t-tests, or chi-square test. Students should be able to distinguish between correlation and causality, and understand that correlation is not causation (Garfield & Chance, 2000). They should know how to interpret scatterplots and two-way tables. They should also know that when they are comparing multiple groups, they should not focus exclusively on the average of each group to make a judgment about the difference in means. One common misconception is that a negative correlation implies no correlation (Batanero, Estepa, Godino, & Green, 1996). Another is that if XY and YZ are correlated, then XZ must be correlated (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2009). Students might also have difficulty reading representations of bivariate data and making appropriate judgments about how strongly two variables are related. Students will have had less experience dealing with bivariate data compared to univariate data, since the former is not always a topic covered in elementary and high schools. However, they will likely have had exposure to relational studies from the media, which may contribute to their confusion about correlation and causality.

Probability. Probability theory provides the tools to model random experiments and to study sampling distributions of statistical indices. In probability, the composition of the population is known and we predict the likelihood of future events. In statistics, the composition of the population is unknown, but, given the statistical indices of the sample, we can make probabilistic inferences about the composition of the population. However, we can never be certain about the composition

of the population.

Students should be able to apply probability theory to statistical problems. They should be able to reason about randomness and uncertainty to make judgments about the chance of certain events. In dealing with probability problems, students will leverage their prior experiences dealing with events based on likelihoods. However, humans often develop efficient but faulty heuristics when using probability to predict the likelihood of events (Tversky & Kahneman, 1974). Although there are many misconceptions revolving around how to apply probability correctly, three will be mentioned here.

First is the representativeness misconception, which is the belief that the likelihood of a sample is a function of how closely it resembles the population (Kahneman, Slovic, & Tversky, 1982). When considering results of tossing a fair coin, people often believe that HTHT is more likely than HHHH. Another version of this is the Gambler's Fallacy, which views chance as a self-correcting process. For example, if a series of four coin tosses resulted in HHHH, some people may think that tails would be *more* likely to land on the fifth toss. Similarly, the conjunction fallacy is when people think that the events A and B are more likely to happen than just A. For example, Tversky and Kahneman (1986) provided participants with the follow example:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable?
>
> (a) Linda is a bank teller. (b) Linda is a bank teller and is active in the feminist movement.

In the study, most participants chose the second option, even though mathematically it is less likely to occur than the first option. Again, this shows that many people base probabilities on

representativeness rather than on statistical properties and relationships such as multiplying the base probabilities of independent events.

Second, the equiprobability bias is the view that different outcomes will be equally likely in an experiment, regardless of the likelihood of each outcome (Tversky & Kahneman, 1986). In two dice rolls, some people may think rolling two sixes is as likely as rolling one six and one five. Morsanyi, Primi, Chiesi, and Handley (2009) found that this bias increased with statistics education. One explanation for this is that students misunderstand the concept of randomness. In addition, predilection for this bias increased when the item concerned the decisions of people, likely because people are seen as unpredictable, autonomous agents (Callaert, 2004).

Third, the outcome orientation is the heuristic to calculate the probability of a series of events, the individual events should be treated as yes/no decisions (Konold 1989). For example, if there is a 75% of rain on four days, people with an outcome orientation would say it should rain on all four days instead of three of the four days.

In applying probability strategies, students are likely to draw from their everyday experiences with randomness and chance. Context can also play an important role in how students apply heuristics and misconceptions.

Sampling. Inferences about the population are made from samples, and representative sampling can help to ensure that generalizations from the sample are warranted. Averaging or summing of quantities tends to produce normal distributions (although the original distribution may not necessarily be a normal distribution). Students should know how samples are related to populations and how sample size and representativeness affects the strength of statistical analyses. Sampling is related to the concept of probability in that sampling uses random selection from the population.

Students may have many misunderstandings about sampling and particularly have difficulty applying ideas of sampling to statistical inference. Often they confuse the original sample with the result of the sampling process as well as sample and population distributions (Chance, delMas, & Garfield, 2004). This may be a result of incorrectly employing the representativeness heuristic (Tversky & Kahneman, 1986). Similarly, they may believe that random sampling is a self-correcting process. They may not understand how sample size is related to sampling. They may think that as the sample size increases, the closer the distribution will look like the normal distribution. Students may leverage their ideas of probability and representativeness when answering conceptual items on sampling.

Large sample theory. Large sample theory provides a mechanism for generalizing how sample size relates to large sample theoretic results and implications for inferential statistics. When observations are drawn randomly from the population, the population mean is the limit of the sample mean as the sample size increases. However, this does not necessarily mean that as a sample size gets larger, the mean will necessarily regress to the mean.

Students should understand the central limit theorem and how the law of large numbers impacts the large sample behavior of the sample average and of inferential statistics. That is, they should understand the relationship between sample size, sampling error, and predictive accuracy of a statistical test. A prerequisite skill may be to understand the role that variability plays in random events.

Students tend to have several misunderstandings involving large sample theory. Tversky and Kahneman (1986) showed that students often are unable to make judgments about how sample size affects the characteristics of the sample. One question they asked students is the following:

A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day. In the smaller one, about 15 babies are born each day. Although the overall proportion of boys is about 50%, the actual proportion at either hospital may be greater or less on any day. At the end of a year, which hospital will have the greater number of days on which more than 60% of the babies born were boys?

A. The large hospital

B. The smaller hospital

C. Neither – the number of these days will be about the same

The correct answer is B, since the smaller hospital is more likely to have a variable proportion of boys. However, most students in the study chose C, indicating that they did not understand the impact of sample size on variability.

In some cases, students do use sample size to make inferences about variability, but these are not always for the right reasons. In Bar-Hillel's study (1982), most students (80%, n=72) answered the following question correctly:

Two pollsters are conducting a survey to estimate the proportion of voters who intend to vote YES on a certain referendum. Firm A is surveying a sample of 400 individuals. Firm B is surveying a sample of 1000 individuals. Whose estimate would you be more confident in accepting?

Firm A's____ Firm B's____ About the same____

Well, Pollatsek, and Boyce (1990) ran a series of experiments and found that students tend to use information about sample size appropriately when asked about the *accuracy* of the sample means, but tended answer items about the tails of the sampling distribution incorrectly. They reasoned that students tended to answer the pollster question correctly because it was about the accuracy of the

sample means; the fact that students tended not to answer the tails questions correctly indicated that they did not have a deep understanding of how variability impacts sample averages.

These experiments demonstrate that students may not understand how sample size affects the variance of the mean (Chance et al., 2004). Students may think that a large enough sample can represent the characteristics of the population without doing any hypothesis testing. Some students may even believe that any sample, regardless of size, is representative of the population. Such thinking results in undue confidence in the reliability of small samples, and may underestimate the size of confidence intervals and overestimate replication results in future experiments.

Statistical significance. A p-value is "the conditional probability of at least as extreme outcomes and it could not be attached to the null hypothesis" (Falk & Greenbaum, 1995, p.94). Failing to prove a difference does not prove that the null hypothesis is true. Differences in p-values can be explained by sample size, standard deviation, study design, and chance. Students should be able to interpret hypothesis testing, considering levels of significance, p-value, and power. They should understand the meaning of p-values, alpha testing, and significance testing.

Students find it challenging to understand how to conduct significance tests. Part of this is due to the foreign vocabulary of hypothesis testing. They may have confusion about the meaning of p-values, alpha, null hypothesis, alternative hypothesis, test statistic, and the critical value (Vallecillos & Batanero, 1996). For example, they may think that the p-value is the probability of the null hypothesis, the probability of obtaining the same data or making an error, or the strength of the treatment (Sotos et al., 2009). Some might have difficulty interpreting the numerical value of the p-value. Students may think that lower p-values have stronger treatment effects (Gliner, Leech, & Morgan, 2002). Another misconception is that alpha is the probability that one of the hypotheses are true or that it is the probability of rejection when the null is incorrect. They might also see the

significance level as the probability of the null hypothesis being true given it has been rejected. Statistical significance is an abstract concept mired in a new language, making it more difficult for students to conceptualize. Students may try to leverage their understanding about decision making and probability to solve problems dealing with hypothesis testing.

Interpreting statistical significance is one skill and understanding the implications of hypothesis testing is another conceptually difficult task. Statistics helps to identify true patterns from random chance using hypothesis testing, which often require assumptions about the sampling procedure used. Given statistical significance, students should be able to distinguish true from false patterns. They should be able to recognize when each of the standard hypothesis tests is appropriate to use, based on the probability modeling assumptions made and sampling procedure used (such as the observations being independent, the sample size being large, and the population being normal).

The most prevalent mistake students make in statistical inference is viewing the results of a hypothesis test as a logical proof (Vallecillos, 1995); seeing hypothesis testing as a probabilistic proof (Falk & Greenbaum, 1995); or not understanding the relationship between hypothesis testing and the decision process. The logical proof misconception may derive from students' experiences in mathematics, in which proving a formula means deterministically showing that it is true or false. Students might also misunderstand the evaluation of statistical significance, in that they might not know the difference between practical and statistical significance. Researchers need to know about the sample size and the design of the experiment in order to make a judgment regarding practical significance. This misconception might stem from the everyday meaning of "significant" meaning important, whereas in statistics it has a different meaning.

2.2.3 Domain Modeling. This next layer of the ECD framework organizes the information from the Domain Analysis into a narrative about the assessment argument (Mislevy & Riconscente,

2005). A design pattern template serves as the mechanism for formally representing this argument (see Appendix B). The design pattern template in Appendix B contains multiple elements and what follows is an unpacking of its key elements.

The first section of the design pattern template provides a rationale for the domain model. As mentioned in the domain analysis, conceptual understanding is difficult to assess, and students often maintain erroneous beliefs even after formal instruction. Professors tend to stress procedural knowledge and assume their students have already mastered the necessary conceptual understandings. In order to facilitate conceptual change, professors need a tool to measure students' prior beliefs and understandings. Students' pattern of responses can serve as evidence for conceptual understandings and misconceptions. Consistent choice of correct answers within a conceptual category can serve as evidence of student mastery. Similarly, students who choose more than one distractor mapped to a particular misconception may entertain an alternate understanding of the target concept.

The construct labels section of the design pattern template indicates all labels for the 11 types of Focal Knowledge (FKs) within the 6 categories designated in the Domain Analysis. The Focal Knowledge, Skills, and Abilities section describes each of the learning objectives for each of the six categories. This includes the high level claims about what introductory statistics students should know and understand for each concept. Under each of these categories, FKs of what students should understand about the topic were described. The FKs were sub-divided when the conceptual knowledge pieces could be differentiated. The following provides brief descriptions of each of the FKs:

- **FK1a- Measures of central tendency**: Ability to summarize data using estimates of central tendency, i.e., mean, median, mode. Knowing how outliers impact each measure.

- **FK1b- Measures of dispersion**: Ability to summarize data using measures of dispersion, i.e., variation and standard deviation.
- **FK2- Correlation**: Ability to interpret studies using correlations and recognize limitations. Knowing when it is justified to make a claim about causation from results of a statistical analysis.
- **FK3a- Probability theory**: Ability to apply probability theory to statistical problems accurately.
- **FK3b- Events in Probability**: Ability to interpret the relationship between events. In particular, determine if a relationship is dependent or independent. Ability to interpret the outcomes of multiple events.
- **FK4a- Sampling**: Ability to identify how "good" samples are created. Ability to differentiate between sample drawn from a population and sampling. Recognize properties of a normal distribution.
- **FK4b- Sample size and probability**: Ability to recognize the nature of the relationship between sample size and probability.
- **FK4c- Data gathering**: Ability to identify qualities of "good" data gathering. Requires sampling with some randomized mechanism and independent replication.
- **FK5- Large sample theory**: Ability to state the implications of large sample theory on probability and recognize importance in practice of statistical research.
- **FK6a- P-values**: Understanding the practical significance of p-values and alpha. Understanding that differences in p-values can be explained by sample size, standard deviation, study design, and chance.
- **FK6b- Hypothesis testing**: Understanding the practical significance of hypothesis testing.

In order to answer StatCI items correctly, students will have to use additional knowledge, skills, and abilities (AKs). While these skills are not the direct targets of assessment, they are unavoidable because of the medium of the assessment and/or the context of the questions. It is important that an assessment developer keep these in mind when designing the assessment to avoid having the AKs unduly influence how students respond to the items. AKs can be the cause of unexpected poor response patterns by introducing construct-irrelevant variance (Messick, 1989). The AKs in the design pattern template include the ability to read and interpret graphs, knowledge of variables and functions, literacy skills, number sense, and familiarity with real-world situations.

Potential Observations (PO) in the design pattern template are the possible types of responses students could provide to show evidence for the FKs. For selected response items, the accuracy of the answer chosen could be indicative of the specified FK. Another possibility includes the degree of certainty a student demonstrates in selecting the answer (such as if as student switches between responses). Students that do not switch between answers are more likely to be confident about their knowledge of the construct. Time taken by students to respond to the task may also indicate the degree to which the student finds the item cognitively demanding. Finally, if the task is an open response or talk aloud, the quality of the rationale the student provides for the answer would provide evidence for the degree of FK understanding.

Potential Work Products (PW) in the design pattern template are methods by which students can demonstrate evidence for the specified FKs. One PW is answers to selected response items. These answers must indicate as much as possible that students leveraged conceptual understanding to answer the item correctly. Examples of learning objectives may include: reason probabilistically to solve statistical problems, interpret statistical output, predict what will happen in a given situation and justify the corresponding selection with reasoning. Another PW may be recordings or

transcripts of students trying to work through problems. The last example PWs are computer records of students' interaction with online assessment (for example, time it took for students to complete answer, whether students switched answers, whether students used any supplementary tools). Each method has its own affordances. For example, transcripts may provide richer information on student thinking than selected response data, but will also be more time intensive and difficult to assess. The final PW will dictate how the actual assessment is conducted.

Characteristic features (CF) in the design pattern template are aspects of tasks that will elicit the desired evidence for the target FKs. The primary CF indicates that the tasks should ask students to evaluate interpretations or predict outcomes by presenting a situation that requires understanding the targeted focal knowledge, skills, or abilities.

Variable Features (VF) in the design pattern template are aspects of the tasks that can be modified to vary task difficulty or focus. Students may know how to respond to tasks with which they are familiar; answering correctly may not demonstrate conceptual understanding but rote memorization. Applying concepts to new contexts can indicate more robust transfer of knowledge. Therefore, it is important that students are provided a variety of contexts in which to apply their conceptual understanding. One possible VF is the type of distribution to focus on. Students are generally more familiar with the normal distribution, and may mistakenly assume the mean is appropriate measure of central tendency across all contexts. Another VF is the way in which data are represented. For example, data could be presented through graphs, text, tables, or diagrams. Students may be familiar with applying certain methods on one type of representation, but may have to leverage conceptual understanding when dealing with a representation with which they are not familiar. The third VF is the statistical test for applying reasoning about statistical significance. The last VF is the real-life context for the item.

To ensure that each of the topics was relevant and covered in the Common Core curriculum, the categories and learning objectives of the design pattern template were mapped onto the standards and this mapping is shown in the last section of the design pattern template. Appendix C lists all of the standards for statistics in one column and the corresponding FKs. Each FK is linked to a standard. However, not all of the standards are linked to FKs; some of the standards elicit procedural understanding, which is outside the scope of this assessment.

The Domain Modeling layer serves as the platform for the next layer in the ECD framework, one that coordinates the variables, tasks, and scoring mechanisms. Before turning to a description of that design layer as it pertains to the current StatCI instrument, it is important to consider how the information contained in both the Domain Analysis and Domain Model aligns with the focal knowledge and skills assessed in inventories that have been previously developed for introductory statistics. The next section discusses various properties of four such inventories and evaluates their adequacy for assessing conceptual understanding in this instructional domain. After reviewing the strengths and limitations of these existing inventories, discussion returns to application of the other components of ECD to design of the current StatCI instrument.

2.2.4 Existing Assessments of Statistical Conceptual Understanding. Currently, there are four publicly available statistics assessments purportedly focused on conceptual understanding for an introductory statistics course.[1] In addition, there is a Diagnostic Statistics Assessment (Masters and Famularo, 2015) designed to measure three misconceptions held by middle school students.[2] This section will discuss each of these four assessments in terms of how each corresponds to the

[1] Research on the teaching of statistics appear in multiple literatures—such as in math education—but issues about student learning are largely in the psychological research literature.

[2] Since the scope of the Diagnostic Statistics Assessment is narrowly focused, it has been excluded from this analysis. However, the challenges the developers faced in creating this assessment are relevant to all conceptual assessments in the domain of statistics and will be discussed in the following section.

preceding domain analysis, as well as their format, content, and validity evidence. TABLE II maps the domain model created in this study to those concepts covered on each of the four assessments.

The first is the Statistical Reasoning Assessment (Garfield, 1998). TABLE III indicates the eight correct reasoning skills and eight misconceptions tested on the Statistical Reasoning Assessment. The Statistical Reasoning Assessment's reasoning skills overlaps with those in the domain analysis, in regards to probabilities, univariate data (understanding how to select an appropriate average), sampling, large sample theory, and bivariate data (correlation vs. causality). In addition, this test includes computation of probabilities, which the domain analysis intentionally leaves out (procedural understanding is out of the scope of this domain analysis). The Statistical Reasoning Assessment does not cover measures of spread and interpretation of statistical significance, which the domain analysis has shown are important concepts in first year statistics. The Statistical Reasoning Assessment is a multiple-choice test with 20 items. Responses generally consist of an answer and a statement of reasoning to support the choice. The instrument assesses eight correct reasoning skills and eight misconceptions (see TABLE III). Two to four items map onto each skill. The score report provides information on student performance for each of the two major subscales: correct reasoning and misconceptions. Liu (1998) reports that the subscales were moderately reliable, with a test-retest reliability of .70 for the correct reasoning component and .75 for the misconceptions component. Tempelaar (2004) found low and sometimes negative inter-item correlations. The eight correct reasoning scales had an alpha of .24 and the eight misconception scales had an alpha of .06. This supports the conclusion that the instrument may not do an adequate job of reliably tapping into conceptual understanding. One possible reason for these negative findings is that there are too many skills being tested on too few items.

The Statistical Reasoning Assessment was later adapted into the Quantitative Reasoning

Quotient, with modified item format, scoring, score reports, and additional misconception categories (Sundre & Thelk, 2003). The Quantitative Reasoning Quotient is a 40 item test designed to assess 11 skills and 15 misconceptions in the domain of undergraduate statistics. Since the Quantitative Reasoning Quotient has more conceptual categories than the Statistical Reasoning Assessment, it

TABLE II

PRE-EXISTING STATISTICAL ASSESSMENTS MAPPED BY CONCEPT

Assessment	Statistical Reasoning Assessment	Quantitative Reasoning Quotient	Comprehensive Assessment of Outcomes in a first Statistics Course	Statistics Concept Inventory
FK1a- Measures of central tendency	C2	C2/C3	C2 (descriptive)	C2 (descriptive)
FK1b- Measures of dispersion			C2 (descriptive)	C2 (descriptive)
FK2- Correlation	C6	C7	C6 (bivariate)	
FK3a- Probability theory	C1	C1	C7	C1
FK3b- Events in Probability	C4	C5		
FK4a- Sampling	C5	C6	C8	
FK4b- Sample size and probability	C8			
FK4c- Data gathering		C10/C11	C1	
FK5- Large sample theory	C8	C9		
FK6a- P-values				
FK6b- Hypothesis testing			C10	C4 (inferential)
Not Included	C3. Correctly computes probability, C7. Correctly interprets two-way tables	C4. Correctly computes probability, C8. Correctly interprets two-way tables	C3. Graphical representation, C4. Box plots, C5. Normal distributions, C9. Confidence intervals	C2. Graphing

TABLE III

THE STATISTICAL REASONING ASSESSMENT SKILLS *(Garfield and Chance, 2000)*	
Correct reasoning skills	1. Correctly interprets probabilities 2. Understands how to select an appropriate average 3. Correctly computes probability a. Understands probabilities as ratios b. Uses combinatorial reasoning 4. Understands independence 5. Understands sampling variability 6. Distinguishes between correlation and causality 7. Correctly interprets two-way tables 8. Understands importance of large samples
Misconceptions	1. Misconceptions involving averages a. Average is the most common number b. Fails to take outliers into consideration when computing the mean c. Compares groups based on their averages d. Confuses mean with median 2. Outcome orientation misconception 3. Good samples have to represent a high percentage of the population 4. Law of small numbers 5. Representativeness misconception 6. Correlation implies causation 7. Equiprobability bias 8. Groups can only be compared if they are the same size

shares similar overlap with the domain analysis (see Table IV). In addition, the Quantitative Reasoning Quotient has several categories that the domain analysis does not: differentiating between measures of central tendency and selecting the appropriate average, understanding sources of bias and error, and recognizing good experiment design. This version still did not include measures of

TABLE IV

THE QUANTITATIVE REASONING QUOTIENT SKILLS
(Sundre, 2003)

Correct reasoning skills	1. Correctly interprets probabilities 2. Correctly interprets measures of central tendency 3. Understand how to select appropriate average 4. Correctly computes probability 5. Understand independence 6. Understands sampling variability 7. Distinguishes between correlation and causation 8. Correctly interprets two-way tables 9. Understands the importance of large samples 10. Understands sources of bias and error 11. Recognizes features of good experimental design
Misconceptions	1. Misconceptions involving averages 2. Outcome orientation misconception 3. Good samples have to represent a high percentage of the population 4. Law of small numbers 5. Representativeness misconception 6. Correlation implies causation 7. Equiprobability bias 8. Groups can only be compared if they are the same size 9. Failure to distinguish the difference between a sample and a population 10. Failure to consider and evaluate all of the data 11. Inability to create and evaluate fractions or percents 12. Only large effects can be considered meaningful 13. Failure to recognize potential sources of bias and error 14. Assumes more decimal places indicate greater accuracy 15. Inability to interpret probabilities

spread or hypothesis testing. Like the Statistical Reasoning Assessment, two to four items map onto each skill. Each item is worth up to 2 points, with possible partial credit. Sundre and Thelk's analysis of the instrument suggested it is not a reliable measure of student learning. Cronbach's alpha on the overall instrument was .55 (n = 1083) and students often performed better on the instrument before statistics instruction than after completion of a course. The test developers also had difficulties measuring misconceptions on this test; they could not develop reliable indicators of individual misconceptions sufficient to warrant the inferences desired about these constructs (Sundre, D., personal communication, October 15, 2014).

Given that reliability is proportional to the number of items on the assessment and this has twice the number of items than the Statistical Reasoning Assessment, the instrument's poor reliability measure is somewhat surprising. One possible reason for the low reliability is that several items have lengthy text stems with several items based on this text. This means students have to retain textual details while trying to reason conceptually, which might lead to extraneous cognitive load (Bransford et al., 2000). Long text items might also introduce construct-irrelevant variance, since a needed skill to answering the item correctly is reading comprehension and background knowledge (Haladyna & Downing, 2004). Another possible reason for low reliability is that several items have unique formats that also might make it more challenging for students to respond. For example, item 6 has two sets of distractors labeled a, b, c, and d (8 distractors in all). Which of the two sets the student chooses from depends on the previous answer.

The third assessment is the Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS), a 40 item multiple-choice item test (DelMas, Garfield, Ooms, & Chance, 2006). CAOS overlaps with the domain analysis, except that it has several additional topics: graphical representation, box plots, normal distributions, and confidence intervals. Even though this test is

comprehensive as compared to the domain analysis, there is only one published article about its validity properties. This article includes a reliability analyses, comparisons of pre- and post-test measures, and analyses of distractors; there are no analyses of the assessment's structural properties.

This test underwent four iterations and reviews by statistics experts. The items were designed so that students would have to reason and not compute or recall definitions. The test is divided into 10 topics: data collection and design, descriptive statistics, graphical representation, box-plots, normal distributions, bivariate data, probability, sampling variability, confidence intervals, and tests of significance. Its most recent validation study resulted in a Cronbach's alpha of .77. The overall pre-test to post-test scores indicated only a marginal increase. The percentage of correct responses even decreased on certain items. There is no additional published information about the assessment's validity properties, indicating that overall this instrument lacks evidence to support the claims the developers are trying to assert about its interpretation and use.

The fourth assessment is the Statistics Concept Inventory (SCI), a 38 item multiple-choice test (Allen, 2006). The SCI covers three of the topics on the domain analysis (probability, descriptive statistics, and inferential statistics). It also has graphing as a category that the domain analysis does not cover. The scope of the SCI categories is much broader that the ones proposed on the domain analysis, which, as previously mentioned, can be problematic for diagnostic purposes.

An analysis of this inventory also indicated poor reliability and structural properties (Jorion et al., 2014). The Cronbach's alpha score was .64 (n = 402), a modest reliability for an assessment of this length. An exploratory factor analysis revealed the items did not load on the developers' pre-defined categories and overall accounted for less than 20% of the variance in scores. For comparison, another concept inventory, the Concept Assessment Tool for Statics (CATS), had a reliability measure of .84 (n = 1372) and a factor analysis that explained 54% of the variance in

scores. Jorion et al. Posited that this difference could be attributed to the construction of the respective tests. The CATS had well-defined conceptual categories; in contrast, the SCI had broadly defined, chapter-book topics (probability, graphing, descriptive statistics, inferential statistics).

The instruments discussed above individually and collectively lack various forms of validity evidence. Thus, in the domain of statistics there is ongoing need for an instrument with strong validity evidence that includes adequate construct representation and desirable measurement properties. Development and validation of such an instrument requires a conceptually principled design process (Jorion et al., 2014). Such a design process should clearly define three models: (1) the targeted cognitive models and the resulting conceptual categories that encapsulate conceptual thinking, (2) the corresponding tasks that can elicit this thinking, and (3) the measurement models that can identify and differentiate among different types of thinking. These three components serve as the basis for obtaining evidence in support of the validity argument one would want to make about student understanding in the domain of statistics. Without such a framework, it is problematic to make well-founded, empirically supported assertions about what students know and can do within a domain. Development of the current StatCI instrument is consistent with such a principled design and validation approach and in the next section we describe the next major element in application of ECD to design of the current assessment of conceptual understanding in statistics.

2.2.5 Challenges in Designing Conceptual Assessments of Statistics. Test design influences an assessment's reliability and the degree to which it is valid. Some of the subpar performance of the existing assessments of statistical concepts could be attributed to the way in which the tests were constructed. The Statistics Reasoning Assessment has too many concepts with too few items per concept; the Quantitative Reasoning Quotient has lengthy text stems and unusual item formats; the Statistics Concept Inventory uses an overly broad set of conceptual categories.

Another contributing factor may be that statistics as a domain has some inherent challenges unique regarding measurement of conceptual understanding. Jessica Masters, who created a reliable diagnostic assessment of conceptual understanding in geometry (Masters, 2013), noted four difficulties she encountered when creating the Diagnostic Statistics Assessment (Masters and Famularo, 2015). Although her assessment is geared towards middle school students, the challenges that she describes are applicable to any assessment of conceptual understanding in statistics. The first challenge is choosing the appropriate level of context for an item. Context is particularly important in statistics; figuring out the appropriate statistical measure is dependent upon the context of the situation. However, context also increases demands related to cognitive load and may introduce construct-irrelevant variance, resulting in items that are less reliable. Related to the first difficulty is finding meaningful and realistic contexts. Any context grounded in experience may present scenarios with which certain demographic groups may be unfamiliar, introducing bias and sensitivity issues.

Third, analyzing data often requires making sense of visual displays. But having examinees interpret visual information adds an extra barrier to demonstrating statistical understanding. On the SCI, the developers included a "graphical" category, which had the lowest subscale reliability of the assessment's four categories (α=.27; Jorion et al., 2014). This may be because the "graphical" category is not a precisely defined concept. However, more refined definitions of this concept, such as "Interpret and make comparisons between different graphical representations" (Stone, 2006), are not concepts in themselves; the graph is the medium in which researchers communicate information, not a concept in itself.

Additionally, misunderstandings of basic statistical language can also present a barrier to item comprehension. Words like "average" and "significant" have different meanings in quotidian

versus statistical language. Developers should use accessible statistics vocabulary in item writing, especially on assessments that are purported to be used as pre-tests; however, identifying statistical words that are as unequivocal as possible is a challenge unto itself. In sum, statistics assessment developers have to be mindful of challenges particular to tapping into statistical understanding, such as item context, visual displays, and language use—and these are in addition to general difficulties in test design that any item developer may face.

2.2.6 Conceptual Assessment Framework. The Conceptual Assessment Framework leverages the Domain Analysis and the Domain Modeling to create the assessment blueprint (Riconscente, Mislevy, & Hamel, 2005). The Conceptual Assessment Framework is composed of the Student Model, the Task Model, and the Evidence Model. The Student Model specifies what should be assessed. The Task Model specifies the tasks that should elicit the behaviors and the environment in which these tasks should take place. The Evidence Model indicates the behaviors that should reveal these constructs, which consists of evaluation rules and the measurement model. All three of these are specified in the task template (Appendix D), which describes the technical details for building the assessment. The design pattern from the Domain Modeling informed the specifications for the task template. The task template first details the Student Model, which is (1) one overall measure of proficiency, (2) subscores of performance on 6 specified concepts, and (3) patterns of common errors and misconceptions. For the overall model of student proficiency, a univariate, continuous measurement model would be required, whereas for the subscores, a multivariate model is appropriate. Both of these will utilize dichotomous scoring responses of right and wrong answers. Reporting patterns of erroneous thinking will require polytomous data and a multivariate measurement model. The evaluation procedures uses automated scoring with an answer key. The Task Model variable involves use of selected response items presented via computer using

the Qualtrics software platform. Respondents may be asked to predict the outcome of a situation and justify their reasoning, or to interpret different types of data, and provide appropriate inferences. Each of the items was mapped onto an FK designated in the design pattern and this can be represented in summary form via a Q-matrix (see Appendix E). The Q-matrix shows the alignment of items with FKs and in most cases there is a one-to-one mapping of an item to an FK. There are, however, some cases where an item is presumed to tap multiple FKs. Each distractor for every item is linked to a unique misconception or common student error. The template-level task model variables include the content area (that is, the specific construct being assessed), the complexity of the construct, and the familiarity of the students with the content. The activity summary states that students fill out the test online and once they are finished, they can review their results.

The Evidence Model is specified in the "participants and data collection" section. This section details how students' responses to the multiple-choice items (the observed variables) will be used to make inferences about student understanding (the unobserved variables) using measurement models. The evaluation rules will serve to detail how to extract the salient observed variables from the student work products. The measurement model will help to make inferences about these extracted observed variables about the designated latent constructs students possess. This model is how developers can create evidence to validate the pre-specified claims that the inventory is supposed to make about student understanding.

To illustrate how some of the key elements of the Conceptual Assessment Framework are instantiated, we can take one of the items from the current inventory and map it against the features described above. The following is an example item from the inventory for FK3a, probability theory:

> Q15. Nora is a 36 year-old woman who loves cats. In 2014, her town had 25 cat breeders and 500 post office employees. Which of the following is the most likely? Assume being a

cat breeder and working at the post office are independent of one another.

A. She works as a cat breeder.

B. She works at the post office.

C. She is a cat breeder who supports her passion by working at the post office.

D. It is equally likely that she works at the post office or is a cat breeder.

This item requires students to apply knowledge about probability to determine the likelihood of certain events. Statistically speaking, since there are more post office employees than cat breeders in the town, it is more likely that Nora will work at the post office. Each of the distractor choices are linked to a common faulty heuristic. Choice **A** is linked to the base rate fallacy (M3a.5), which is when people ignore statistical information and instead rely on irrelevant information. Some people may think that since Nora loves cats, it is more likely that she is a cat breeder, regardless of the relative probability. Choice **C** is linked to the conjunctive fallacy (M3a.7), which is when people assume that the intersection of two events is more probable than one of the events. Choice **D** is linked to the equiprobability bias (M3a.2), which is when people believe that the probability of two outcomes is equally likely, even though the probabilities of the individual events are different.

Before administering the pilot version, the assessment was reviewed by three content experts who have pedagogical content knowledge in the domain (one professor, two statistics teaching assistants). Pedagogical content knowledge requires familiarity with idiosyncratic student difficulties, misconceptions, prior conceptions, and representations particular to the domain (Shulman, 1986). In creating educational materials, it is necessary but not sufficient to have content expertise, partially due to the expert blind spot (Nathan & Koedinger, 2000). This is why it was necessary to have experts with pedagogical content knowledge review the items. All three experts assessed each item for clarity and correctness. The researcher asked each expert: (1) whether the

item was correct in terms of content, (2) whether students might have confusion about certain features of the item, (3) whether the distractors might capture common student responses, (4) the estimated difficulty of the item, and (5) any other suggested changes to the item. In turn, a decision was made about whether to modify or delete the item.

The pilot instrument tests six important concepts in statistics (a link to the full instrument is provided in Appendix F.) The six concepts are as follows: univariate data, bivariate data (correlation), probability, sampling, large sample theory, and statistical significance. Item context was varied within a conceptual cluster to test if student knowledge was dependent on contextual features. Each item has three distractors linked to a different misconception. The same misconceptions were assessed across items within a concept. For an example of how item distractors were mapped onto errorful thinking within the probability category, see Table V.

To summarize, the researcher performed three main steps to design the assessment using evidence-centered design. The first was to conduct a Domain Analysis, which consisted of identifying and organizing important topics in statistics derived from sources such as the Common Core Standards for Mathematics, the College Board, and the American Statistical Association. A statistics professor created a list of big ideas and enduring understandings, and the researcher mapped these the preceding three resources. This Domain Analysis served as a foundation for the second step, the Domain Model, which defines the skills and abilities that would constitute evidence for proficiency of each specified topic. Building on this second step, the third step was the creation of a Conceptual Assessment Framework, which identified the evidence linking student proficiencies to the relevant tasks. This design process is a principled approach of specifying important constructs and connecting these to tasks that can serve to provide evidence for student statistical understanding.

TABLE V
EXAMPLE OF MAPPED DISTRACTORS FOR PROBABILITY

Distractor	FM3a.2	FM3a.4	FM3a.5	FM3a.6	FM3b.7
	Equiprobability Bias	Time Axis	Base Rate Fallacy	Outcome orientation	Conjunction Fallacy
Q10.a	0	0	0	1	0
Q10.b	0	1	0	0	0
Q11.a	0	0	0	1	0
Q11.b	0	1	0	0	0
Q11.c	0	0	0	0	1
Q12.a	0	0	1	0	0
Q12.c	0	0	0	0	1
Q12.d	1	0	0	0	0
Q13.a	0	0	1	0	0
Q13.c	0	0	0	0	1
Q13.d	1	0	0	0	0
Q14.c	1	0	0	0	0

III. METHOD AND ANALYSIS

3.1 Participants and Data Collection

The current research project consisted of four studies (1, 2A, 2B, and 3), each of which served to support claims about the Statistics Concept Inventory (StatCI). The first study was a pilot think aloud with four undergraduates. The other studies used participants from Amazon Mechanical Turk (MTurk) (See Appendix H: Demographics of Participants for justification of using MTurk participants and possible implications for the study). Study 2A was an initial analysis of response data from 100 participants. The researcher added study 2B because the results from study 2A suggested the need for additional revisions to the assessment items. Study 3 assessed 750 participants with a revised version of the StatCI. For each of the studies, participants were screened to (1) be over 18 years old and (2) have taken at least one undergraduate statistics course. All participants were compensated for their time.

For the MTurk studies, participants were asked to complete a demographics questionnaire (see Appendix G). Participants were asked their gender, highest level of education completed, and ethnicity. These data were used to determine if measurement bias occurred on particular questions in the test. If participants with the same latent trait have different probabilities of answering the item correctly based on group membership, this indicates that there may be something peculiar with an item. This would call for a careful investigation and revision of the items as necessary. Finally, participants were also asked how many statistics classes they have taken and how many years it has been since they have taken a statistics course. These results will be used to inform eligibility criteria for future iterations of this assessment.

3.2 Analysis

The analysis serves to substantiate three claims about the Statistics Concept Inventory's (StatCI) intended purposes:

Claim 1: Students' CI scores can be used to indicate their overall understanding of all concepts identified in the CI. This claim is about a student's overall proficiency in the conceptual domain represented by the set of items in the inventory. This is reported as a single number on a unidimensional scale. Each individual item contributes to this measure coherently, so it is important that items are functioning well individually. There are four ways of evaluating this claim. First, the researcher should ensure that each item is contributing to the overall proficiency measure. This includes calculating alpha-if-item-deleted, item discrimination, item difficulty, and item response theory (IRT) model-fit (Crocker & Algina, 2006). Items with multiple measures falling outside of the recommended ranges may not be functioning as the developer intended. Second, the researcher can use Cronbach's alpha, a measure based on the number of items and inter-item correlations, to evaluate overall reliability. Third, the standard error of measurement indicates the confidence with which particular scores can be differentiated. Fourth, participants with more instructional experience in statistics should score higher on the inventory than those with less experience. Each analysis plays a complementary role in evaluating the extent to which the CI measures overall domain mastery.

Claim 2: Students' CI scores can be used to indicate their understanding of specific concepts. These claims are about a student's proficiency on each of the concepts identified by developers as constituting the domain of the inventory. This claim asserts that the CI has differentiable subgroups that make up the conceptual domain; it is the degree to which

performance on the items align with the developer's hypothesized constructs. This performance measure is reported as several numbers on a multidimensional scale. Methods used to evaluate this claim include subscale alphas, exploratory factor analysis, confirmatory factor analysis, and diagnostic classification modeling.

Claim 3: Students' CI scores can be used to indicate their propensity for misconceptions or student errors. These claims are about student misconceptions or common errors in thinking. This claim asserts that student distractor response patterns can indicate the extent to which students subscribe to common errors or misconceptions. The value is reported as several numbers corresponding to the particular error. There are two main ways to investigate this measure. First, the researcher can compare high and low performing students' response patterns. Low performing students are more likely to have misconceptions about the domain, so the answers they choose can be indicative of misconceptions. Second, Bock's (1972) nominal response model, a polytomous IRT scoring response model, can be used to map distractors to a numerical value for latent ability. Each answer choice can be associated with a unique level of difficulty, which can indicate knowledge progressions (Sadler, 2009).

These claims follow the Evidentiary Validity Framework as outlined in Jorion et al. (2015), which describes the process and application of this framework in more detail. TABLE VI provides a summary of each analytic approach linked to the claim about what the CI intends to measure.

TABLE VI

ANALYTIC FRAMEWORK FOR ANALYSES OF CI DATA

Reprinted from "An Analytic Framework for Evaluating the Validity of Concept Inventory Claims," by N. Jorion, B. Gane, K. James, L., Schroeder, L. DiBello, and J. W. Pellegrino, 2015, 104(4), p.458-459, Journal of Engineering Education. Adapted with permission.

Analyses	Claim 1 – Overall understanding	Claim 2 – Understanding of specific concepts
	Classical test theory	
Calculate item difficulty	Provides the total percentage of students who answer an item correctly; the higher the score, the easier the item. The most effective items overall have mid-ranges of difficulty.	Provides item functioning information for each of the items within a concept category.
Calculate item discrimination	Provides the correlation between item right and wrong score (represented by 1 and 0) and total score. Items are most effective when they discriminate well between students with high or low total scores.	Provides item discriminability information for each item within a concept category.
Calculate total score Cronbach's alpha	Provides an estimate of total score reliability, where reliability means that a given student's total score would be nearly the same if we were able to administer the test multiple times to the same student. Alpha is an internal consistency index among item responses, used as a reliability estimate.	Provides, for each concept, the total score across all items within that concept category. It can be reported as subscale scores that reflect a student's proficiency with each concept. Its reliability can be estimated by computing Cronbach's alpha of the subscale score.
Calculate alpha-with-item-deleted	Provides a measure of how well a given item coheres with the remaining items and is consistent with the total score on the remaining items.	Provides, for each concept, a measure of how well a given item coheres with the remaining items in that concept group and is consistent with total score on the remaining items in that concept group.

Item response theory (IRT)		
Select a unidimensional IRT model: Estimate model parameters Evaluate model: fit; standard errors of estimates; and the reasonability and usefulness of estimated item parameter values	Provides a student's estimated overall proficiency as a number (theta) on a continuous scale.	Multidimensional IRT models provide a profile vector of proficiencies on individual concepts, often as subscores on separate subtests thought to be conceptually distinct.
Determine *where* the test is functioning more and less well: Plot a Wright Map to compare item locations and student proficiencies on a common scale Plot the item and test information functions as functions of theta.	Provides model parameters for individual items that represent aspects of item performance. For example, IRT models place item difficulty on the same scale as student ability, and allows comparing degree of match between items and students. The item and test information functions show for which students the test provides good information.	For subtests of sufficient length, provides multi-dimensional model parameter estimates for items and students to support reporting of student proficiencies on separate concepts.

	Structural analysis	
Perform an exploratory factor analysis (EFA) Determine the number of factors in an EFA solution, the factor loadings of individual items, and the proportion of variance explained by the solution.	Provides a structure of one or more latent factors that can explain the item inter-correlations, if such exist. A "good" factor solution has strong factor loadings for each item, and explains a high proportion of the inter-item variability. Such a factor solution would support the interpretation of total score as an overall measure of student knowledge.	Provides a latent factor structure that is capable of explaining the inter-item correlations in the data, if such a structure exists. Such a structure, if it exists, is derived directly from the data and can be compared to the developers' concept structure to test whether the data support the developers' hypothesized concepts as tested by the inventory?.
Perform a confirmatory factor analysis (CFA) Define a CFA structure that reflects the developers' hypothesized concept structure. Determine whether the posited CFA structure can be estimated from the data. Evaluate the model.		Provides an evaluation of how well a specified latent structure can explain the inter-item correlations in the data. This differs from EFA because the researcher can use the developers' hypothesized concepts to posit a latent structure, and then directly test that posited structure
	Diagnostic classification modeling (DCM)	
Define a DCM that is consistent with (or close to) the developers' concept structure.		Provides an evaluation of whether reliable reporting of student profiles of individual concept proficiencies is possible, and if so provides model parameter estimates for such a model.

3.2.1 Study 1. The first study was a talk-aloud protocol of four undergraduate students, in which students answered items and explained their thinking. The aim of this study was to determine the extent to which student thinking aligned with the expected statistical thinking. Protocols can serve as evidence of construct-irrelevant variance. Ideally, students should be answering items correctly because they are leveraging conceptual understanding that corresponds to the item. However, students may fixate on features of the item in a way that was not intended by the developer. This may lead to two undesirable situations. First, students may find hints embedded in the item, causing them to answering correctly even though they do not have the corresponding conceptual knowledge. Second, students may be led astray by inconsequential details, causing them to answer incorrectly even though they have the appropriate conceptual knowledge. Student protocol studies are one way to identify incongruence between student thinking and responses. In particular, it can help highlight features of items that should be revised. Results from this analysis informed revisions in wording, format, and distractors of the items.

3.2.2 Study 2A and 2B. Study 2Awas a pilot administration of the StatCI. The goal of this study was to conduct a preliminary check of the StatCI's construct validity regarding the items, the conceptual groupings, and the distractors. The findings served as the basis for edits for the second version of the assessment. The sample consisted of 100 participants.

Item functioning. Item functioning was investigated using classical test theory and the 1PL item response theory model. Classical test theory and 1PL item response theory are appropriate methods of investigating dichotomous data, given the sample size (Morizot, Ainsworth, & Reise, 2007). Classical test theory is a total score measure. It assumes each person has a true score, which is the observed score plus an error term (Crocker & Algina, 2006). Measures within the classical test theory framework that can provide evidence about claims of overall ability across concepts include item difficulty, item discrimination, and Cronbach's alpha-if-item-deleted. Subscale alphas

can provide evidence about coherence of conceptual categories. Cronbach's alpha can provide evidence about the overall reliability of the instrument.

Item response theory is another framework in which the person's true score is defined as a unidimensional latent trait. Compared to classical test theory, item response theory allows for more robust claims of item parameters, such as invariance across all populations. 1PL item response theory can provide evidence for item difficulty.

For the first administration, those items that have classical test theory difficulty levels below .2 and above .8 will be scrutinized. Items with low discrimination values will have to be modified or eliminated. It is possible such items have features that provide hints or obscure what the item is asking, especially if similar results are produced in the 1PL analysis. Misfitting items may be a threat to construct validity, so such items may have to be edited or deleted entirely.

Conceptual group functioning. The relationship between the items will be investigated using tetrachoric correlations, which are correlations adjusted for dichotomous data (Bonett, & Price, 2005). Items within the same conceptual group should have higher correlations. Items with negative tetrachoric correlations could indicate that the items are measuring separate constructs. More advanced approaches for analyzing data grouping, such as exploratory factor analysis or structural equation modeling, would not be appropriate given the sample size. The general rule of thumb for calculating appropriate minimum sample size for structural equation modeling is 10 subjects per parameter (Kline, 2005). The structural equation model for this assessment would consist of 30 loadings and 30 error terms (one for each item). If the six groups are allowed to covary, there would be an additional 15 correlations. This would yield a total of 75 parameters, for a recommended sample size of 750.

Distractor analyses. Distractor analyses can provide evidence about student progression towards understanding and propensity for particular errors and/or misconceptions. Correct answers

on the assessment can indicate the degree to which students have mastered concepts and distractors can indicate where students have misconceptions. Cross-tabulations of responses can be useful to this end. However, just because a developer claims that an assessment's item distractors map onto particular misconceptions does not necessarily mean that learner who pick the distractors necessarily subscribe to that misconception. Picking a distractor linked to a misconception might be caused by the sustained presence of that misconception, contextual factors that may elicit particular misconceptions, or just a lapse in judgment. Without more than one occurrence of such a misconception, it is difficult to make such evaluations of student thinking with any degree of confidence. Infrequently chosen distractors were edited or dropped. Items were scrutinized according to these analyses and modified as necessary.

The researcher conducted an additional study, Study 2B, when the results from Study 2A indicated poor reliability likely related to the sampling procedure. The results from Study 2A served to inform modifications to the StatCI. Another 100 participants were sampled for this study using the revised version. The purpose of this study was to ensure that the items were functioning as intended and would help for the next iteration of the assessment. This study was less in-depth than study 2A, and therefore did not include an item response theory analysis.

3.2.3 Study 3. The third study used the beta version of the StatCI. The goal of this study was a more comprehensive study of the StatCI's construct validity. Seven hundred and fifty participants on MTurk took a revised version of the assessment. Although there is no gold standard for minimum sample size for these analyses, a larger number of participants are recommended in order to obtain adequate parameter estimation results (MacCallum, Widaman, Zhang, & Hong, 1999). Moreover, 750 is an adequate sample size estimate for the hypothesized structural equation model. All analyses run on the previous version of the assessment will be run on the new data. Three additional analyses will be conducted for item functioning, conceptual group functioning, and

distractor analyses.

Item functioning. A 2PL item response theory model was used to investigate item functioning. In addition to calculating item difficulty like the 1PL model, the 2PL model also calculated item discrimination. Because the 2PL model has more parameters than the 1PL model, researchers suggest that the minimal sample size to conduct this analyses is 200 (Morizot, Ainsworth, & Reise, 2007). In addition, item bias was also investigated using differential item functioning. This analysis flags items that have different probabilities of being answered correctly by people with the same latent ability but membership in different groups. The presence of differential item functioning on an assessment can be a threat to its construct validity (Ackerman, 1992). Items that demonstrated the presence of differential item functioning were scrutinized for construct-irrelevant variance.

Conceptual group functioning. Factor analysis was used to investigate the assessment's structural properties based upon inter-item correlations. Results can substantiate claims that the hypothesized constructs underlying an assessment's design align with the latent factors resulting from the factor analysis. In exploratory factor analysis, no model is specified a priori. Resulting factor loadings can be compared to the developer's hypothesized structure. When factor loadings do not align with the developer's constructs, it may indicate that the items are measuring a different construct than intended. In confirmatory factor analysis, the researcher can verify the fit of a specified model and compare among several models. Such an analysis can provide evidence that the assessment is measuring the targeted constructs.

Distractor analyses. Distractors were investigated using chi-squared tabulations and Bock's (1972) nominal response model. The nominal response model, a polytomous item response theory model, can be used to substantiate claims about item distractors. Each distractor can be mapped onto a latent state, so that the distractors can be quantified in order of difficulty. This can be useful for

tracking how learners progress on the path of conceptual understanding. Sadler (1998) used this model to investigate student misconceptions in astronomy and how instruction might foster misconceptions. For an item asking about the cause of day and night, students of moderate ability had a lower probability of choosing the correct answer compared to those with a lower ability level. Consistency of responses across distractors (when participants chose distractors mapped onto particular misconceptions for several items) can serve as evidence for the context sensitivity of certain misconceptions. However, it is also possible that statistical conceptual understanding is dependent on superficial features of the items; certain features may elicit certain misconceptions more than others. In this case, it would be unlikely than any distractor analyses will result in reliable patterns in response data.

IV. RESULTS

4.1 Study 1

Student talk-aloud protocols were conducted in October 2015. The participants were recruited through the psychology listserv. To be eligible for the study, students needed (1) to have completed the statistical methods course, (2) to be fluent in English, and (3) to be over 18. Selected students were compensated for their participation. Four psychology students participated in the interviews. All of the participants were undergraduate females. Each student had completed a course in statistical methods within the past year and expressed high self-efficacy in statistics. Two of the students stated that their previous statistics course was "easy," and that they received high marks in the class without studying. Each student finished a talk-aloud of the 30-item exam within an hour. The researcher asked the students to read the items, talk through their thinking and provide any observations about the item or choice, and then indicate their final answer. When the participants chose an answer directly without providing an explanation, the interviewer asked, "Why did you choose that answer?" The interviews were audio-recorded and transcribed later for analysis.

The researcher considered three questions for the analysis of the student talk-aloud data:

(1) To what extent does student reasoning align with student responses?

(2) Are students who answer confidently more likely to answer correctly?

(3) Did students miss the item because it was difficult or because of construct-irrelevant variance?

4.1.1 To what extent does student reasoning align with student responses? Ideally, students who use appropriate reasoning should choose the correct answer, and students who use incorrect reasoning should choose a distractor. In creating strong items, the aim is to minimize false positives (correct answers but incorrect thinking) and false negatives (incorrect answers but correct thinking). To this end, each student response was labeled with the following codes: *No idea, No*

reasoning, *False reasoning*, *Mixed reasoning*, and *Correct reasoning*.

1. *No idea* is when students expressed that they did not know how to answer the question and were guessing. An example is the following: "I think the answer is C. I'm not sure why, though."
2. *No reasoning* is when the students did not clearly articulate a reason for their answer, either because they had no idea or it was a missed opportunity on the part of the interviewer to probe the student's understanding. In either case, it indicates a lack of understanding on the part of the student. An example would be, "I thought the second part was B at first but C is better."
3. *False reasoning* is when students use incorrect reasoning to decide on the answer. For example, one student said, "There's more variation in graph A because it's not normally distributed." This student evaluates the variability of a distribution by comparing it to a normal distribution—a heuristic which may work in some cases, but is not always the correct method for determining the spread of a distribution. *False reasoning* could also be when students answer based on irrelevant features of the item, such the context or wording. For instance, Q2 asked which of two athletes played more consistently based on a density plot. One of the students interpreted "consistently" as a positive attribute, and falsely picked the distractor choice based on the player with the larger mean rather than the narrower distribution.
4. *Mixed reasoning* is when students used incorrect and correct reasoning in their explanations. For example, Q13 asked which sequence was more likely, {Boy, Boy, Boy, Boy, Boy} or {Boy, Boy, Boy, Girl, Girl}. One participant chose the correct answer—that either of the sequences were likely. However, the participant added, "If they are only choosing a boy or girl four times, [they are equally likely], but even then,

you can't base it off of just a couple of times. So if this were to continue for much more, I feel like either of them could because you don't know what's going next. But it's only 5 times. If it were much more, I don't think it [referring to the all boy sequence] would be likely." In this case, the student correctly stated that both sequences would be as likely, but only because there were five cases. If there were much more than five cases, then the all-boy sequence would *not* be as likely. The reasoning is correct for the item's context, but would not be correct given a longer sequence.

5. *Correct reasoning* is when students answer an item with the correct reasoning. In response to the previous sequence problem, one of the participants responded, "This one is straight forward. It could be either of them because there are the same number [of items] in each sequence. So probability would be the same in each case."

Tallies of each code were tabulated according to whether the student answered the item correctly. These tallies are given in Table VII. Mixed answers for the two part questions were not included (when students answered the first part correctly and the second part incorrectly, or vice versa). This is because the responses were not completely "correct" or "incorrect." There were ten such responses omitted. 110 cases appear in the table, for a total of 120 responses (4 participants x 30 items = 120 total responses). Overall, this table indicates that most participants picked the correct answer using correct reasoning, and most participants chose the incorrect answer using incorrect reasoning. There was one case in which a participant answered an item correctly using incorrect reasoning. For item 19, the first distractors were as follows: (A) *Increase the sample size so that the distribution looks closer to a normal distribution.* (B) *Survey at least 30 employees to ensure that the sample will be normally distributed.* Both of these are linked to particular misconceptions: corresponding to distractor B, some students believe that a sample size of 30 is a magical number when it comes to creating a normal distribution.

TABLE VII
CROSS TABULATIONS OF STUDENT REASONING AND CORRECT ANSWER

	No idea	No Reason Given	Incorrect Reasoning	Mixed Reasoning	Correct Reasoning
Incorrect	11	19	39	7	0
Correct	0	6	1	5	22

One student responded, "I chose D, because A and B are too similar." This feature of the item led the student to the right answer. To address this implicit "clue," one of these choices was eliminated from the distractor pool. The cross-tabulation, in this case, indicated cases in which participants used construct-irrelevant variance to answer an item correctly.

Another example of construct-irrelevant variance appeared on Q12, which assesses knowledge of the conjunction fallacy. The item asks the following:

> Nora is an older woman who loves cats. In 2014, her town had 25 cat breeders and 500 post office employees. Which of the following is the most likely? Assume being a cat breeder and being a post office worker are independent of one another.

Two of the four students fixated on the adjective, "older." One student responded: "By older, what do you mean? Is she 60, 70, past her prime? Then she wouldn't be working in the post office." Another student responded, "I'm going to say that she's a cat breeder just because it's the first thing. She's an older woman and she might not have a job or be retired." Both of these students chose that Nora was a cat breeder because being employed as a post office worker is apparently less likely for

an older person. The item was edited to indicate that Nora was in her thirties, an age when most people are still employed and not retired.

4.1.2 Are students who answer confidently more likely to answer correctly? The second item investigated the extent to which confidence and correctness were associated. It would appear that confidence would be indicative of correctness, and vice versa. Not confidence responses expressed doubt or uncertainty. An example of a not confident response is "C is throwing me off. Is this a trick question? They are all true. I'm not sure what the answer is." Confident responses were worded as statements and often provided an explanation. An example of a confident answer is "It's B because the more measurements, the more reliable it will be."

TABLE VIII shows cross tabulations of confidence and answer the item correctly. Mixed answers for two-part items were also not included in this count. Although lack of confidence did seem to be indicative of choosing an incorrect response, there were 15 instances in which participants answered correctly despite not being confident, and 22 cases in which students were confident in their answers but did not answer correctly. Therefore, although there appears to be somewhat of a relationship between lack of confidence and choosing incorrectly, there were still many exceptions in which students doubted their correct answers and other times when they were overconfident. The total expressions of confidence (n=40) versus lack of confidence (n=70) indicated that students lacked self-efficacy on this test, either because they had a difficult time recalling statistical concepts or the items were confusing.

TABLE VIII

CROSS TABULATION OF CONFIDENCE TO ANSWERING CORRECTLY

	Not confident	Confident
Incorrect	55	21
Correct	15	19

4.1.3 Did students miss the problem because of difficulty or construct-irrelevant variance? Third, there were several items that all of the participants missed and provided incorrect reasoning. For Q15, "A data analyst finds an association rule for how customers purchase products in a supermarket. The rule states that 30% of customers who buy cheese also buy bread. What inference can be made based on this statement?" all of the participants incorrectly chose B, "30% of customers who buy beer also buy cheese." It seems that students oversimplified this problem. They may be imagining that the items are already in the buyer's cart rather than chosen at different points in time. The item stem was revised to "The rule states that 30% of customers who buy cheese will also pick up bread on their way out of the grocery store."

Q4 was also a very difficult problem for participants. The item asked, "The two graphs below show the number of points two different athletes scored for each game during a season. Which of the two athletes scored points more consistently?" Students answered this item incorrectly even if they could correctly provide the definition for standard deviation; they did not understand that standard deviation can be used to indicate consistency. Many students were not interpreting the graphs as a histogram, but rather as a rate curve (Figure 4). Students interpreted the line graph as a change over time, and since there was a steeper slope for A, this athlete was the less consistent of the two. After seeing this pattern for the first three respondents, the researcher anticipated this problem

and created an alternative graph with bars instead of a line. The last student also answered this item incorrectly, but changed answers upon seeing the bar histogram. However, the student chose the correct response for the wrong reason (graph B was "bumpier" than graph A, so it had more variation). It appeared that for this item, the graph was a major obstacle to indicating conceptual understanding of spread. The graphs were modified to be more easily interpretable.

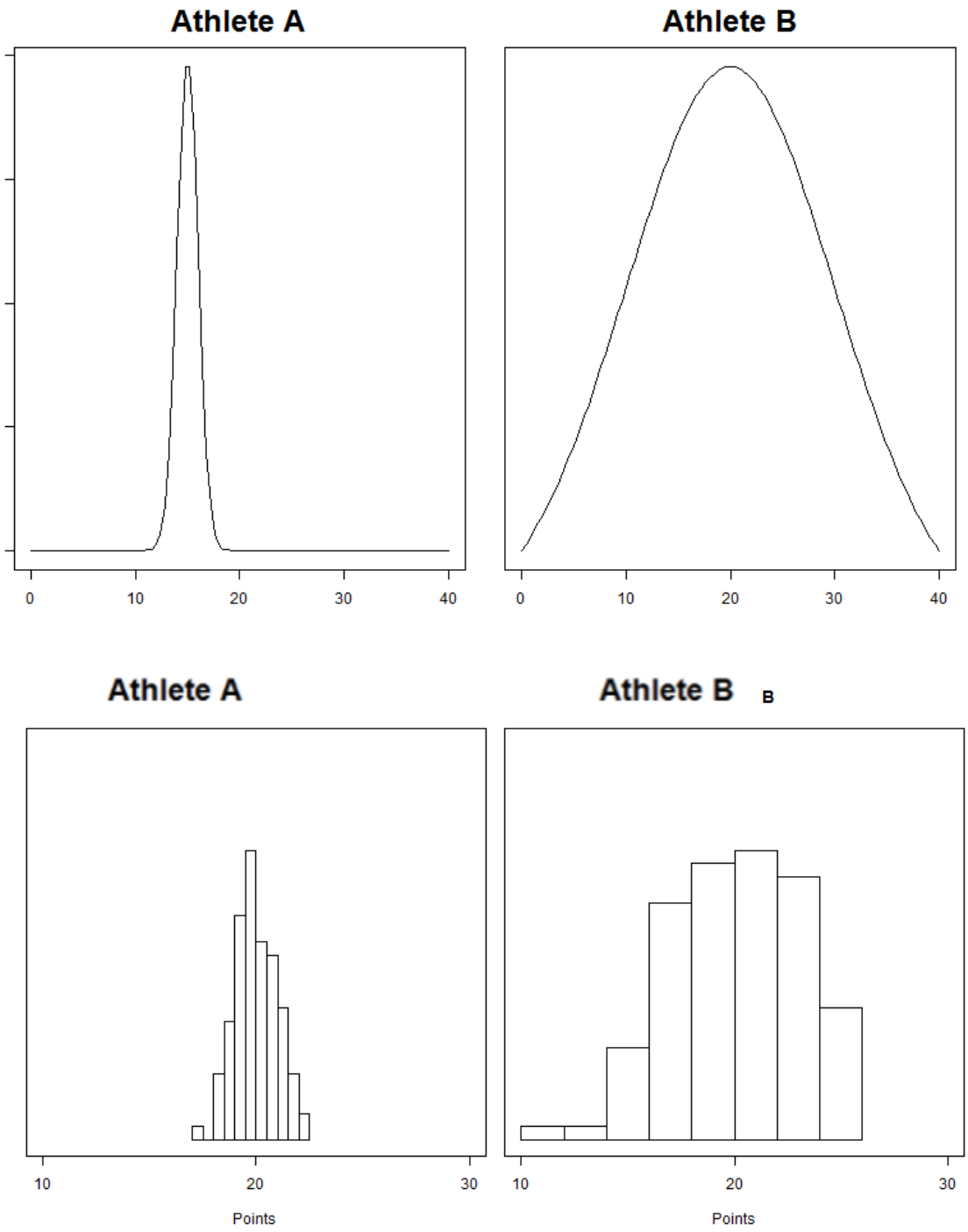


Figure 4. Comparison of graphs for Q4. Above is the original version, and below is the modified version.

Moreover, many students answered incorrectly to items in the hypothesis testing category; most of the responses indicated they had no idea how to answer the item and guessed at random. A typical explanation would be "I'm going with B. I don't know why. For these null hypothesis questions, I really don't know what I'm doing." Students expressed less confidence (21%) in answering the hypothesis testing items correctly when compared to the other groups: univariate data (55%), bivariate data (25%), probability (38%), and sampling (57%). As a result, many students picked the answers to these items randomly, rather than choose a distractor based on an educated guess or a misconception.

Hypothesis testing is not an intuitive concept; there are no experiences students encounter in real life that correspond to this reasoning. It appears that students need a higher level of understanding with this concept in order to have a misconception. A distractor analysis for this section might have problematic results. Compared to the other categories, it seems that this group would be less consistent, which may affect the reliability and factor analyses for the final psychometric analyses.

These talk-aloud studies were helpful in the revisions of the assessment. Prior to the analyses, six graduate students and one professor had reviewed the exam and provided feedback. In some cases, the group of experts found that particular items were strong (like the supermarket association rule item), but the students were not able to answer this item. Interestingly too, it appears that novices fixate and impart too much meaning on irrelevant features of items as compared to experts. For example, none of the experts said anything about Nora being an "older" woman (although perhaps that was because the experts were also older). The analysis comparing student reasoning to their answers was helpful for identifying construct-irrelevant variance and editing the items to create the next version of the assessment.

4.2 Study 2A

The purpose of the second study was to pilot test the assessment and determine the extent to which the items were reliable and valid measures of student understanding. The assessment was first administered using Mechanical Turk from December 16, 2015 to December 19, 2015 in separate batches of 10, 10, and 80. Stratifying the administration helped to ensure that any technical issues could be corrected on a smaller scale. The assessment's content was not modified during the administrations. To qualify for the study, participants had to pass two screening questions: they had to have taken one statistics course and be at least 18 years old. Four participants were eliminated from the sample because they completed the exam in less than 10 minutes, for a total of 96 participants. In this section, I present classical test theory and item response analyses, as well as a more detailed semantic and content analysis of the individual items with high alpha-if-item-deleted measures.

4.2.1 Overall test and individual item functioning. The mean observed score was 9.92 (*SD*=3.7) out of 30, or 33.1% correct. Item difficulties ranged from .13 to .76; six of the items were particularly difficult: Q1 (.13), Q8 (.15), Q23 (.15), Q25 (.16), Q6 (.19), Q15 (.19). Item discriminations ranged from 0 to .5. Nine items had low discriminations: Q18 (0), Q6 (.06), Q8 (.06), Q17 (.09), Q1 (.13), Q2 (.13), Q19 (.13), Q23 (.13), and Q11 (.19). Item difficulties and discriminations are presented in Figure 5.

For the overall assessment, Cronbach's alpha was .58 (*N*=30). For context, the reliabilities of the existing assessments of statistical understanding were .24 (the Statistical Reasoning Assessment, n=20), .55 (the Quantitative Reasoning Quotient, n=40), .77 (the Comprehensive

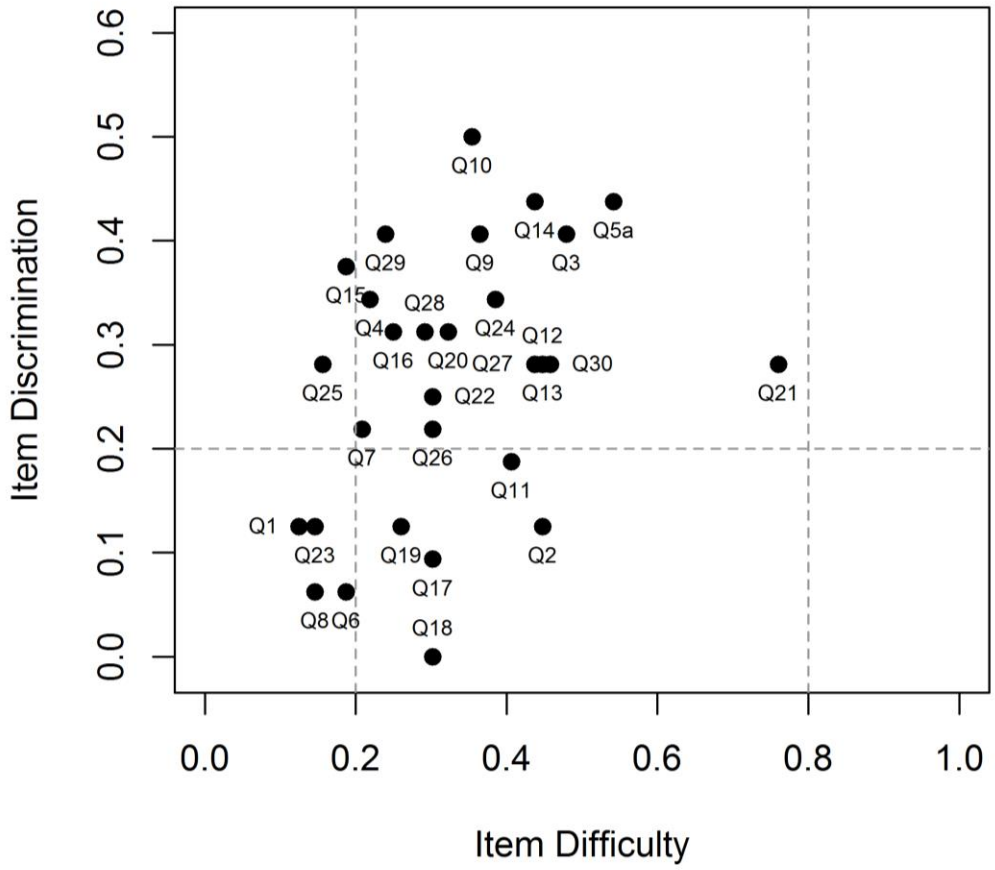


Figure 5. Item difficulty and discrimination measures for study 2A.

Assessment of Outcomes in a first Statistics Course, n=40), and .64 (the Statistics Concept Inventory, n=38).

Seven items had higher alpha-if-item-deleted than the overall test alpha (Q18, Q8, Q7, Q23, Q11, Q19, Q22), suggesting these items were testing a different construct than the other items. When these seven problematic items were removed from the item pool, the overall Cronbach's alpha increased to .66 (*N*=23). Cronbach's alphas were also calculated by demographic (Table IX) with all

30 items. Demographical features that seemed to influence performance on the assessment were level of education, number of years since the last statistics course, and the number of statistics course taken.

TABLE IX

CRONBACH'S ALPHA BY DEMOGRAPHIC FOR STUDY 2A

	α	*N*=	Mean
Level of Education			
Some college	.47	38	9.4
Finished college	.61	49	10.4
Graduate	.73	9	9.4
Race			
Caucasian	.61	73	9.9
American Indian	NA	1	11.0
Asian or Pacific Islander	.44	7	11.5
Black or African American	-.09	6	8.3
Hispanic/Latino	.43	4	8.3
Other/Would rather not say	.65	5	10.2
Gender			
Female	.47	43	9.2
Male	.60	53	10.5
Number of statistics classes taken			
One	.52	58	9.4
Two	.46	32	10.7
Two or more	.64	38	10.7
Three or more	.90	6	10.7
Number of years since statistics class			
One	.46	26	9.7
Two	.58	45	10.5
Two or more	.62	70	10.0
Three or more	.67	25	9.1

The standard error of estimate for the assessment was 2.4. A student who earned the mean score of 10 would have a 68% confidence interval of scoring between 7.6 and 12.4.

Item response theory. The study 2A results were also analyzed using one-parameter IRT. Figure 6 shows the plotted item response curves. Overall, these curves show that the test was difficult for examinees; participants had to have an ability measure around 1 in order to have a 50% chance of answering the items correctly. The majority of the items had an upper asymptote at .8, meaning that even those examinees with the highest ability level would only have an 80% chance of answering the item correctly. Q21 was the exception with an inflection point at -1.3; for this item, students with an ability of -1.3 would have a 50% chance of getting the right answer. Q1 was the hardest item, with an inflection point at 2.1.

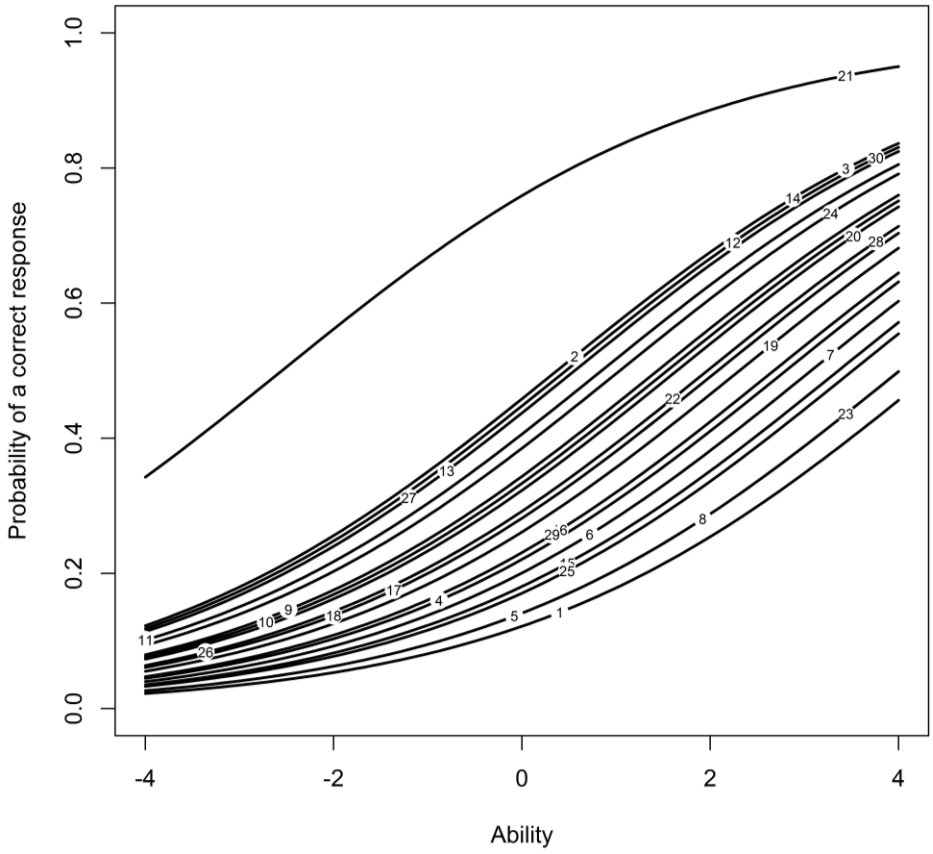


Figure 6. One-parameter item response curves for study 2A.

4.2.2 Structural analyses. Two subscale analyses were performed on the preliminary results: tetrachoric correlation and subscale alphas. Many of the items had low inter- item correlations, as per the tetrachoric correlation matrix (Figure 7); the white space indicates low or negative correlation between item pairs. Ideally, items within conceptual groups should be more strongly related, but the matrix indicates that there is no clear cohesion within the groups.

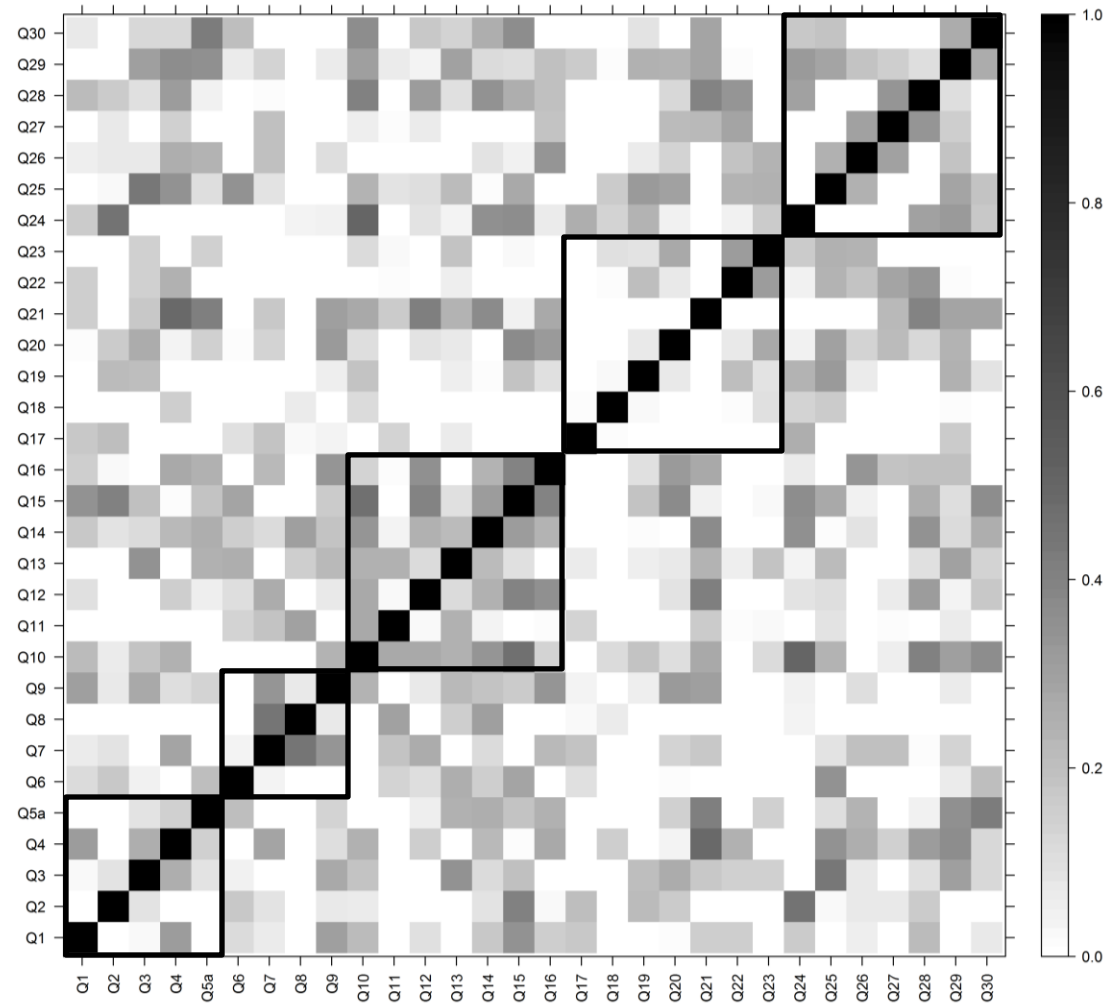


Figure 7. Tetrachoric correlation matrix for study 2A.

The subscale alphas for each conceptual group ranged from .06 to .50 (Table X). Sampling had the lowest subscale measure of the groups. Because Cronbach's alpha is dependent on the number of items, additional alphas were calculated using Spearman-Brown's prophecy formula. Given a new length of 10 items, the range would increase to .12 to .59. For reference, the Statistics Concept Inventory's subscale alphas ranged from .27 (N=5) to .47 (N=10) (Jorion et al., 2014).

These analyses showed that probability as a group was fairly reliable, whereas sampling was not. Many of the items with high alpha-if-item-deleted were from the sampling group, so improving these items should cause the alpha of this overall subscale to increase.

TABLE X

SUBSCALE ALPHAS FOR STUDY 2A

Concept	A	n	If n=10
Univariate	.14	5	.24
Correlation	.23	4	.43
Probability	.50	7	.59
Sampling	.06	5	.12
Hypothesis	.27	7	.35

4.2.3 Item modifications. The researcher looked at the seven items with high alpha-if-item deleted scores in greater detail: Q8, Q11, Q17, Q18, Q19, Q22, and Q23.

Q8, an item about correlation, asked the following:

An organization reports that students who attend preschool are less likely to drop out of high school. They also report that students who attend preschool are more likely to score better on measures of reading. What can be determined from these two statements?

A. Attending preschool is correlated to dropping out of high school.

B. Scoring better on measures of reading is negatively correlated to dropping out of high school.

C. Unless the p-value for the hypothesis test is statistically significant, nothing can be concluded about how preschool leads to a decreased high school dropout rate.

D. Scoring better on measures of reading leads to greater chances of not dropping out of high school.

The correct answer is A, but was only chosen by 15% of the respondents. This item also had a high alpha-if-item-deleted (α=.6), a low discrimination (.06) and low difficulty (.15). Choice B, the most frequently chosen distractor (39%), taps into the correlation is causation misconception. Choice C, the second-most chosen distractor (32%), may actually be a plausible response. This distractor was removed. Another factor leading to the difficulty of this item could the length (179 words). The researcher formulated a new item to use fewer words (68 words).

Q11 was the following:

There is a 10% chance that an earthquake will hit a particular city each month.

Based on this information, what can we infer about this city?

A. There will not be an earthquake in the next year.

B. There will always be an earthquake once every ten months.

C. If there are no earthquakes during the first six months of 2016, there is a higher probability that there will be an earthquake during the last six months of 2016.

D. There will probably be one earthquake during the first ten months of the year 2016.

The item had a slightly high alpha-if-item-deleted measure (α=.59) and low discrimination (.19). The majority of participants chose the correct answer, D (41%). However, many high-scorers chose the distractor B, “There will be an earthquake once every ten months.” Q11 and Q10 were designed to test similar misconceptions, and Q10 performed well in the item analyses. Distractor B was modified to more closely resemble the distractor in the near-isomorph Q10, “If there was an earthquake during the last six months of 2016, then there would not have been any earthquakes in the first six months.”

Q17 asked about sampling:

A technician selects 10 oranges from an orchard. The men weight of this sample is 8 ounces. Based on this information, what can we infer?

A. The population mean should also be 8 ounces, since sampling ensures that the group will be representative of the population.
B. The population mean should be approximately 8 ounces, since it is possible that there will be sampling error.
C. It is difficult to make a judgement about the population mean without more information.
D. The more oranges the farmer adds to the sample, the closer the sample mean will be to the population mean.

The correct answer, C, was chosen by 30% of the respondents. This answer might have been less plausible to savvy test takers who see such a response as an incorrect cop-out. This item also had a high alpha-if-item-deleted (α=.59) and low discrimination (.09). Most participants chose distractor B (32%). This answer is incorrect because there will always be sampling error, a perhaps too-subtle semantic difference. The distractor was changed to "The sample is normally distributed." The correct answer was changed to "The technician would have to weigh all the oranges to find the population mean."

Q18 asked the following:

A researcher is conducting an investigation on sleep. She asks a random sample of 20 participants how many hours they sleep per night.

If the researcher increases the sample size, which of the following is most likely to happen?

A. The distribution of the new sample will look more like a normal distribution.
B. The distribution of the new sample will look more like the population distribution.
C. The mean of the new sample will get closer to the population mean.

D. The mean of the new sample will be equal to the mean of the old sample.

This item taps into the misconception that the greater the sample, the more the distribution looks like a normal distribution, which conflates sampling distributions and samples. The distractor that it would look more like a population distribution and a normal distribution was added as a result of the protocol studies; one of the students said they wanted to choose the answers associated with both the normal and population distributions.

The majority of examinees chose the correct answer (30%). However, the discrimination measure was 0 and the alpha-if-item-deleted was .61 (greater than .58). The context may be confusing participants, since average hours of sleep is normally distributed. So technically, the distractors about the new sample looking like a normal or population distribution is plausible. To address this confusion, the context was changed to sampling counts of different words in a document, which would produce a right-skewed distribution. In addition, the item was edited to take out that it would look like both the population and normal distribution.

Q19 was also modified:

A researcher asks 25 randomly sampled employees from a large company how many hours they work per week. The sample is currently not normally distributed. The researcher is concerned that this could bias the results. What would you suggest to the researcher?

A. Increase the sample size so that the distribution looks closer to a normal distribution.

B. Survey at least 30 employees to ensure that the sample will be normally distributed.

C. Survey all the employees—otherwise, there is no way to ensure the sample will not be biased.

D. The current sample is adequate since representative samples are not always normally distributed.

This item had a same alpha-if-item deleted (α=.58) to the overall alpha and a low discrimination (.12). This item was similar to Q18 in that it asks about sample size and distributions. The problem with this item seemed to be similar; the example context was of a phenomenon that was normally distributed (hours worked per week). The example was changed to ask how many years they spent in school, where the majority of employees finished either college or graduate school. Such a sample would not be normally distributed, but instead bimodal, with maximum values around 16 years and 21 years.

Q22 was a sampling problem:

For a group project, you are administering a survey to students enrolled at your school. There are 30,000 students currently enrolled at your school. One person in your group says that 30 students should be enough to constitute a representative sample for your survey. Is this a sufficient sample size for the survey?

A. Yes, it is a sufficient size.

B. No, it is not a sufficient size.

Because…

A. a good sample should be a high percentage of the population.

B. any sample drawn from the population should be representative.

C. the sample should be representative as long as it is randomly selected from the student population.

The item had an equal alpha-if-item-deleted to the overall alpha and was difficult for participants (.3). The majority of participants chose the answer that 30 was not a representative sample size because "a good sample should be a higher percentage of the population." One possible issue with the item is the wording, particularly "sufficient," "random," and "representative." The researcher changed the item context and flattened the item so it was no longer in two parts.

Q23 was also problematic:

You are conducting a survey on political affiliation in a specific city.

What are ways to ensure that you pick a random sample?

A. Ensure everyone in the population has an equal chance of being picked for the survey.

B. Randomly pick new people when others decline to participate.

C. Randomly pick people within representative subgroups of the population.

D. Sample a relatively large percentage of the population.

This item had a high alpha-if-item-deleted (α=.59), was difficult (.15), and poorly discriminating (.12). Few participants chose B, the correct answer (15%). Most respondents chose the incorrect answer of A (56%). A professor pointed out that such conceptual understanding of sampling was too sophisticated for first-year statistics students. The item was changed to test another sampling concept. In addition to the edits on these seven items, Q4 and Q13 were flattened so that they were no longer two-part items.

4.2.4 Summary and next steps. Overall, the performance on this version of the assessment was low and many examinees performed no different from chance. Moreover, the overall reliability was poor, meaning that there was not much variance among test takers. Given these results, the researcher decided to proceed with a different strategy. First, the researcher modified the problematic items with feedback from three other professors. Second, using the modified item set, the researcher conducted an additional study using Mechanical Turk but restricted the participant qualifications to those who have taken 2 or more statistics courses. The sample size for this additional study was 100. The results of this third study indicated how to proceed with the research.

4.3 Study 2B

Given the low reliability measure of the first administration of the assessment, the researcher ran an additional study of 100 participants. The purpose of this study was to re-check item statistics and make any appropriate item modifications before the larger administration. Consequently, the results of this section do not include an IRT analysis.

Seven of the items had been edited from the previous version. In addition, the prerequisites were changed so that participants had to have taken two or more statistics classes instead of one. The researcher also added an attention check item to ensure that participants were reading the items carefully and not just clicking through the survey in order to get paid. The survey was administered in two batches in January 2016. Eleven participants were rejected for failing the attention check. Table XI indicates demographics for this sample.

4.3.1 Overall test and individual item functioning. The mean observed score was 10.85 (*SD* = 4.39) or 36.2% correct, an improvement of 3.1% from the previous version. Item difficulties ranged from .11 to .69 and item discriminations ranged from 0.04 to 0.63 (Figure 8). Items with low discriminations included Q13 (0.04), Q8 (0.15), Q5 (0.11), Q2 (0.19) and Q17 (0.19).

The overall reliability was .72 (*N*=30), an improvement from the previous version (α = .58). Three items had higher alpha-if-deleted measures: Q6 (.72), Q13 (.73), and Q27 (.73). When these three items were removed, the reliability measure increased to .74.

4.3.2 Structural analyses. Subscale alphas for each category are given in Table XII. Subscale alphas ranged from .16 (hypothesis testing) to .56 (probability). Comparing the subscale alphas from the previous version, most of the alphas improved. Sampling improved substantially as the result of editing five of the seven items. Univariate and correlation also had a modest increase in reliability. However, probability did not improve even though one item had been edited. Q13 detracted from the overall reliability; when this item was taken out, the subscale alpha increased to

TABLE XI

CRONBACH'S ALPHA BY DEMOGRAPHIC FOR STUDY 2B

	α	*N*=	Mean
Level of Education			
Some college	0.56	21	9.4
Finished college	0.67	62	11.2
Graduate	0.85	17	11.6
Race			
Caucasian	0.75	72	10.9
American Indian	NA	1	9.0
Asian or Pacific Islander	0.62	12	11.1
Black or African American	0.33	9	9.0
Hispanic/Latino	0.61	4	10.9
Other/Would rather not say	0.89	2	12.0
Gender			
Female	0.69	44	10.5
Male	0.73	56	11.1
Number of statistics classes taken			
One	-0.21	5	7.0
Two	0.67	74	10.6
Two or more	0.72	84	11.0
Three or more	0.78	21	12.4
Number of years since statistics class			
One	0.36	10	10.4
Two	0.77	10	11.1
Two or more	0.71	90	10.9
Three or more	0.72	80	10.9

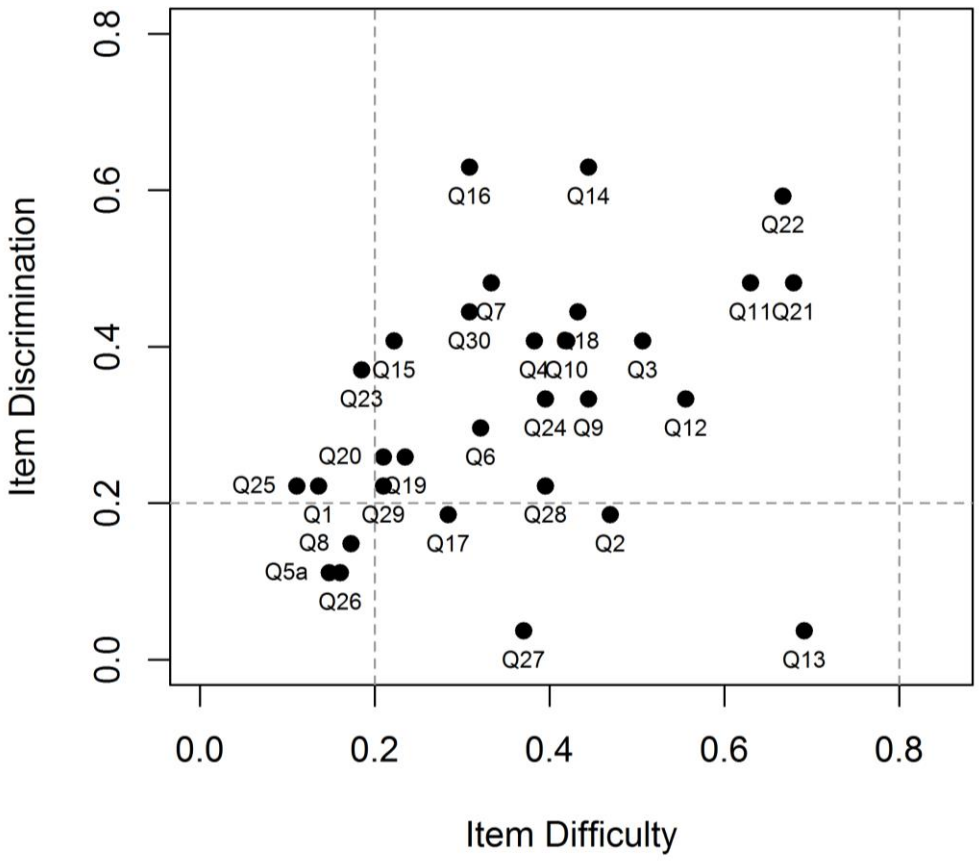


Figure 8. Item difficulty and discriminations for study 2B.

TABLE XII

SUBSCALE ALPHAS FOR STUDY 2B

Test Version	2B	2A	
Concept	A	α	n
Univariate	.23	.14	5
Correlation	.34	.23	4
Probability	.35	.50	7
Sampling	.56	.06	5
Hypothesis	.16	.27	7

.49 (compared to .50 of the previous version). For the hypothesis section, when Q29 and Q30 were removed, the subscale alpha increased to .25 (compared to .27 of the previous version).

Despite these still someone low reliabilities, many of item pairs within categories were strongly related. Table XIII shows a table of item pairs and correlation. There were also some items with high correlations that were not in the same categories, suggesting that participants may be using similar skills to answer the items.

TABLE XIII

TETRACHORIC CORRELATIONS FOR STUDY 2B

Items	Relationship	Correlation
Q10, Q11	Probability theory	.45
Q15, Q16	Probability/logic	.49
Q26, Q28	Hypothesis testing	.53
Q6, Q7	Correlation	.42
Q21, Q22	Sampling	.65
Q7, Q11	“Logic” (Correlation, Probability)	.50
Q7, Q15	“Logic” (Correlation, Probability)	.60
Q4, Q29	“Standard deviation” (Univariate, Hypothesis testing)	.53
Q10, Q18	“Joint events” (Probability, Sampling)	.43
Q14, Q30	“Sample Size” (Probability, Hypothesis testing)	.46
Q14, Q22	“Sample Size” (Probability, Sampling)	.42

4.3.3 Item modifications. Three items with poor reliabilities and discriminations were scrutinized for the next version of the assessment: Q1, Q8, and Q13. Q1, the item about measures of central tendency, was thrown out because it could be argued that both median and mode are appropriate measures for income distribution. Instead, a new item was created about which measure of central tendency was most appropriate for ordinal data.

Q8, a new correlation item from the previous version, was also thrown out because the amount of words and logic involved put a strain on cognitive load. Therefore, a simpler correlation item was created, using the high-functioning Q7 as a model.

Q13 had been flattened from the previous version:

Given that there is an equal chance of choosing a boy or girl, which of the following sequences is more likely?

A. {Boy, Boy, Boy, Boy, Boy} because choosing all boys is more probable.
B. {Boy, Boy, Boy, Girl, Girl} because samples should be characteristic of the larger population.
C. Either because any sequence of the same length is equally possible.

Although Q13 is technically correct, flattening the item caused it to lose discriminatory power. Therefore, the researcher decided to delete the item and add a new one similar to Q12, which had strong item properties.

Several other items were examined and modified as needed. Q5 was problematic because it was a two-part item, which is contributed to its difficulty. It was flattened into a one-part item. Q27 was also a poorly discriminating item; however, it was determined to be technically correct, and therefore was not changed for the next iteration. The researcher also made small changes to the wordings and distractors to several other items for the next version of the assessment (Q9, Q18, Q20, Q29, and Q30).

4.4 Study 3

The purpose of the last study was to substantiate claims about score use of the StatCI for the overall instrument, for concepts in statistics, and for statistical misconceptions. In order to investigate the degree to which the StatCI could enable users to infer participants' understanding of specific concepts, a larger sample was required. In this third study, the researcher administered the assessment to 750 participants on Mechanical Turk over the period of three days. In total, 811 participants took the assessment; data from the 71 participants who failed either of the two attention checks was removed from the sample.

The sample was composed of 53% males, 47% females. The mean age was 32.46 (*SD* = 10.0). The mean time taken for the assessment was 32 minutes and 54 seconds (*SD* = 27.5). In regards to education, 28% were still in college, 53% had graduated college, and 18% had finished a graduate degree. 78% had taken two statistics classes while the rest had taken more than two statistics classes. In regards to number of years since their last statistics class, 8% had a class within the last year, 17% had a class within the last two years, and 75% had taken their last class three or more years ago. 1.2% identified as American Indian, 9.1% as Asian, 7.2% as Black, 5.7% as Hispanic or Latino, 75.8%, and 0.02% indicated other or would rather not say. Table XIV presents the overall demographics for study 3.

4.4.1 Overall test and individual item functioning. The overall mean observed score was 12.16 out of 30 (*SD* = 4.63), or 40.5% correct. A distribution of the scores are presented in the histogram in Figure 9. The researcher conducted basic item analyses to investigate the functioning of all the items (Figure 10). Item difficulties ranged from .22 to .69. As the plot shows, there is a wide range of difficulties, suggesting that the StatCI would be appropriate for testing participants varying in ability level. Item discriminations ranged from .06 to .59. Several items had low

TABLE XIV

CRONBACH'S ALPHA BY DEMOGRAPHIC FOR STUDY 3

	α	*N*=	Mean
Level of Education			
Some college	.71	209	10.73
Finished college	.68	399	12.52
Graduate	.74	142	13.33
Race			
Caucasian	.70	558	12.46
American Indian	.54	10	9.9
Asian or Pacific Islander	.76	67	12.61
Black or African American	.62	56	10.71
Hispanic/Latino	.73	43	10.19
Other/Would rather not say	.67	16	11.75
Gender			
Female	.69	351	11.83
Male	.72	339	12.45
Number of statistics classes taken			
Two	.69	581	11.93
Three or more	.71	169	12.16
Number of years since statistics class			
One	.77	64	12.03
Two	.72	132	11.37
Two or more	.70	686	12.17
Three or more	.70	554	12.36

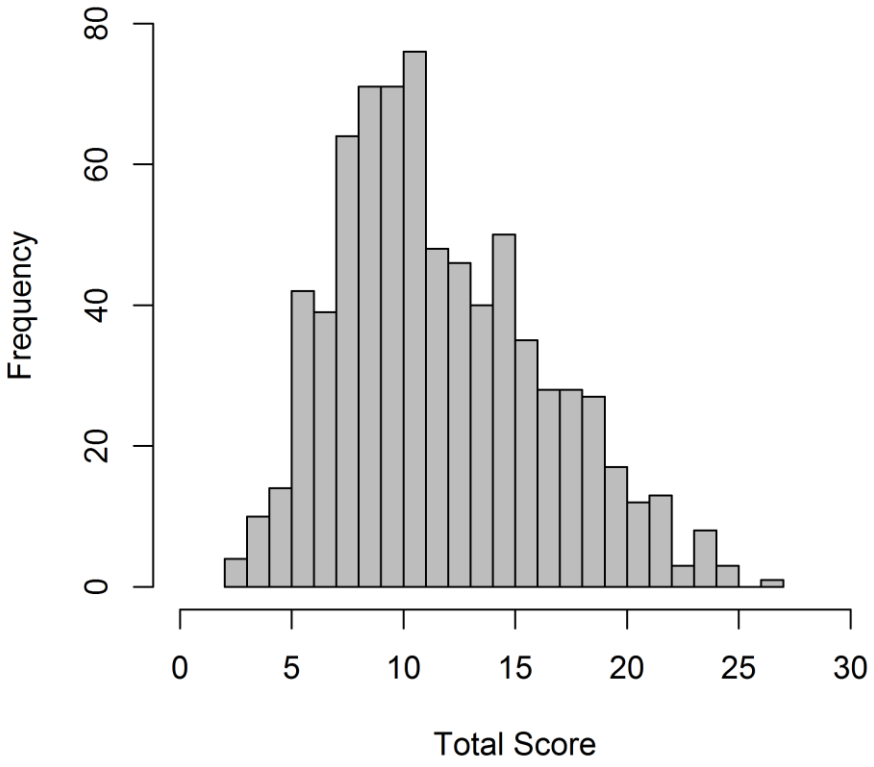


Figure 9. Histogram of the distribution of total scores.

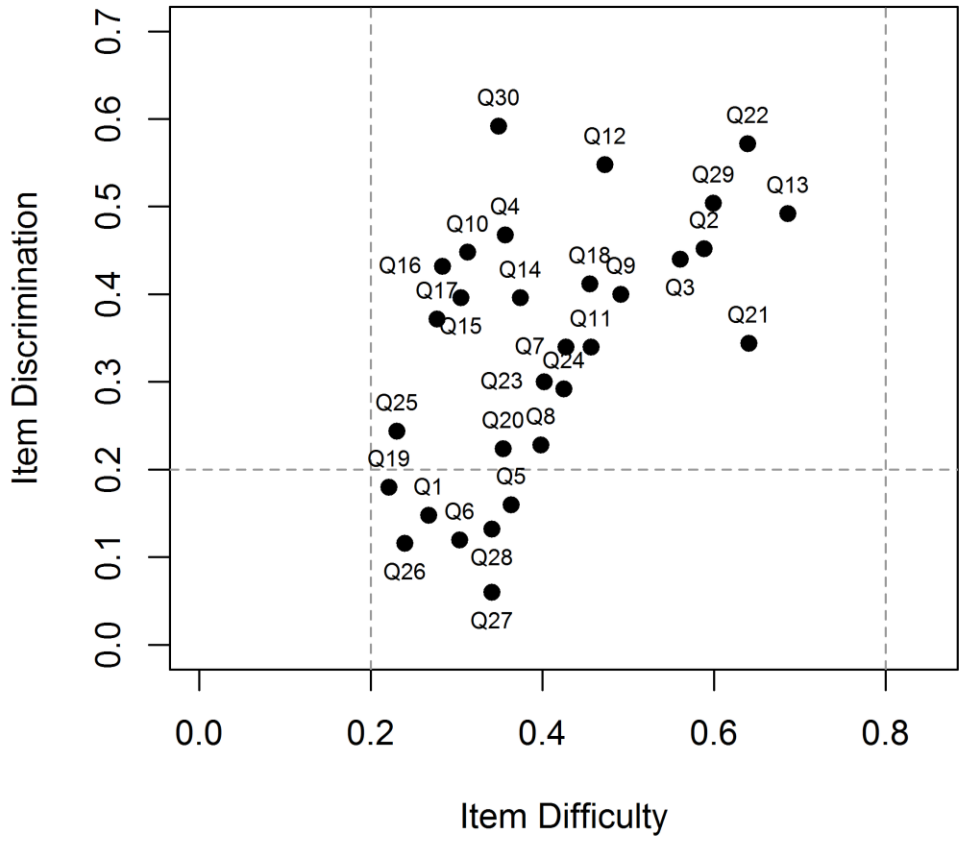


Figure 10. Item difficulties and discriminations for study 3.

discriminations: Q1 (.15), Q5 (.16), Q6 (.12), Q19 (.18), Q26 (.12), Q27 (.06), and Q28 (.13). The item with the lowest discrimination, Q27, was one that was also problematic in Study 2B, even though it was judged as technically correct by a statistical expert. Although many of the items on the assessment contribute positively to overall proficiency as represented by total performance, some of the items performed less well in this regard.

The overall reliability was .71, comparable to the previous version's reliability measure of .72. Several items had a higher Cronbach's-alpha-if-item deleted: Q5 (.71), Q6 (.72), Q26 (.71),

Q27 (.72), and Q28 (.71). When Q27 was removed from the item pool, the reliability increased to .72. This suggests that these items may not cohere conceptually with the rest of the assessment.

The standard error of estimate for the sample was 2.50, meaning that a participant who attained a score of 12 would have 68% confidence interval of having a true score between 9.5 and 14.5.

Item response theory. Item response theory was also used as a complementary method of analyzing participant performance data. The researcher compared the model-fit of the one-, two-, and three-parameter logistic IRT models using Bayesian information criterion (BIC) and Akaike information criterion (AIC); the two-parameter model had the best fit. This implies that letting the discrimination parameter vary resulted in a better fitting model, while adding a guessing parameter did not improve model fit.

Figure 11 displays the item response functions for the two-parameter logistic model. As student ability increases, so too should their probability of answering the item correctly. A well-functioning item should have a normal ogive or smooth s-shape, with a lower asymptote close to 0 and an upper asymptote close to 1. A participant with an ability level of -4.0 should have a low probability of answering the item correctly. An item that has a greater probability of being answered correctly by a participant with a very low ability level shows some evidence of guessing. Likewise, for a well-functioning item, a student with a very high ability level should have a very high probability of answer the item correctly. An item with an upper asymptote much lower than one suggests that the item is tricky for high ability students. Many of the items fit the items fit the model, however several deviated from the normal ogive model. In particular, Q6 and Q27, which were both items that had low reliabilities, had relatively flat item characteristic curves, indicating

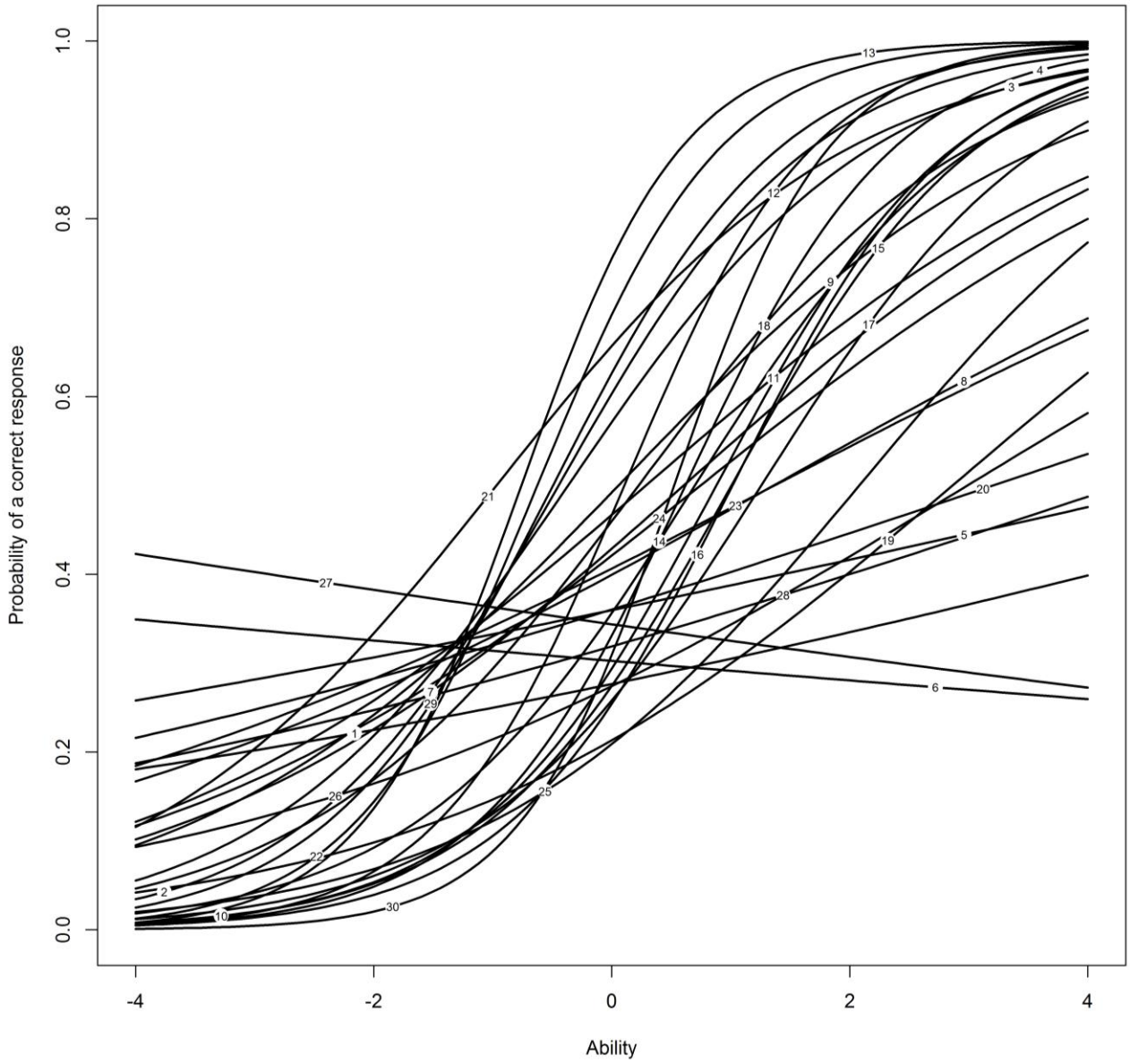


Figure 11. Two-parameter item characteristic curves for study 3.

that they do not do a good job of discriminating between high and low performing participants. Moreover, the slight decreasing curves indicate that participants have a lower probability of answering them correctly the higher their ability level. This suggests that these items were measuring a construct different from the rest of the assessment. Several other items had a .55 probability of being answered correctly by participants with a +4.0 ability level, suggesting that there are features of these items that are causing participants to slip.

Item response theory can also provide a useful framework for investigating the extent to which a test or item is informative for particular ability levels. Figure 12 shows item information functions plotted by ability levels. Q13 was the most informative item for participants between -3.0 and -0.2, and Q30 was the most informative item for participants with abilities levels between -0.2 and +2.0.

Differential item functioning analysis. One threat to test validity is when items measure irrelevant constructs in addition to the ones intended by the test developers. In some cases, however, the probability of answering an item correctly can be conditional on group membership. For example, an item that includes a term with which females are more familiar may give females an advantage over males with similar ability levels. This would suggest that the item is assessing an extraneous construct and could potentially be biased.

DIF analysis is one way to investigate potential measurement bias. This method flags items on which particular subgroups perform better after controlling for participant ability levels. DIF was tested in reference to gender, ethnicity, and age. DIF was detected using the Mantel-Haenszel method (Mantel & Haenszel, 1959), a commonly used method of calculating DIF. The effect size was calculated using deltaMH, where an effect size of 1.5 or more is classified as large (Holland & Thayer, 1988).

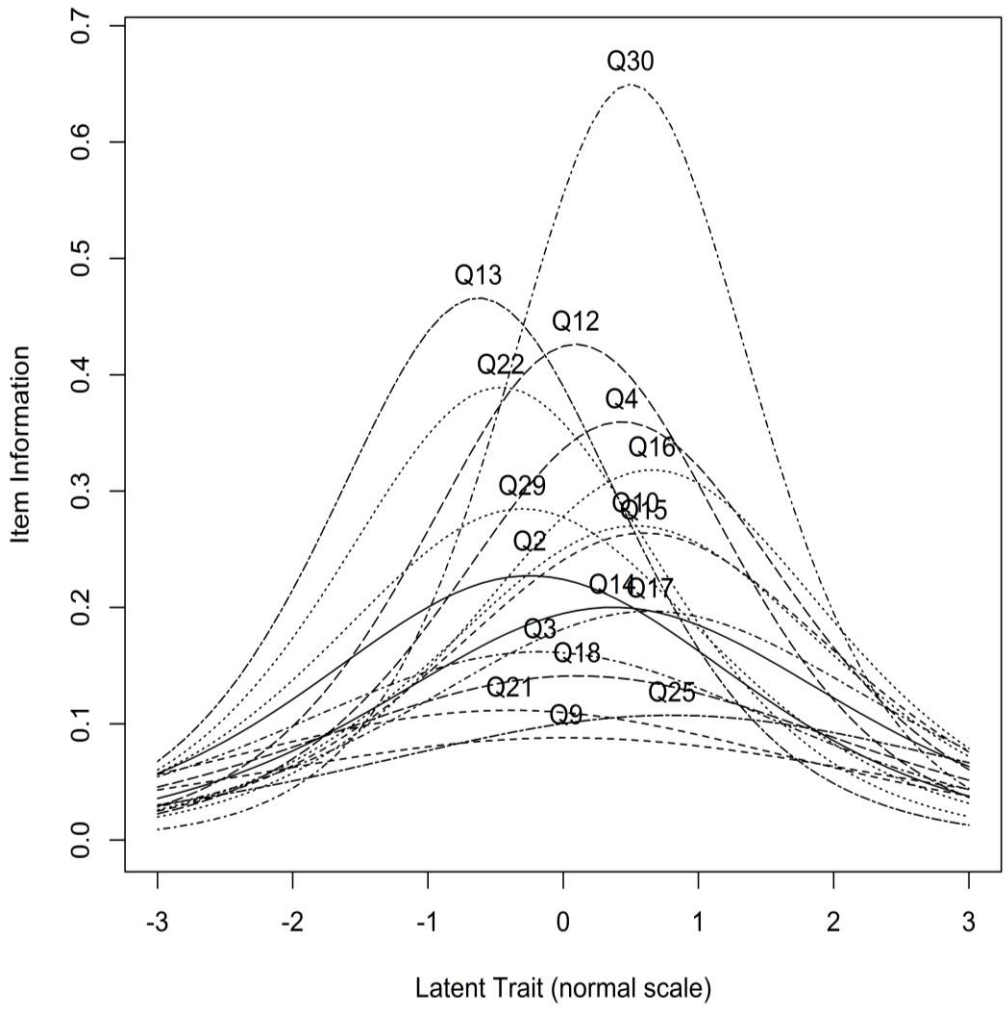


Figure 12. Item information functions for study 3.

For gender, Q14 was flagged with a large effect as being easier for males ($p < .001$, delta-value = 1.586). This item had to do with engineers and computing, a content area with which males may feel more comfortable. Q11 and Q13 were flagged as having moderate effects ($p = .006$, delta-value = 1.048; $p = .017$, delta-value = 1.069, respectively). Q11 had to do with earthquakes and Q13

with firefighters, topics that may be biased towards males.

For ethnicity, white was compared to all other ethnicities. Q3 had a moderate effect against white participants (p = .004, delta-value = -1.376). This item asked about which statistical measure would have the greatest impact on outliers.

For age, the sample was split into two: 30 years-old and younger, and 31 year-old and older. Q22 had a large effect against 30-year olds. The item asked, "Your employer wants you to determine an adequate sample size for a study. The employer wants the sample to be cost-efficient, yet a sufficient size to make inferences about the population. Which would you suggest?" It may be that participants less than 30 years old have less experience in the workplace. On the other hand, two other items, Q9 and Q19, also pertained to situations in the workplace.

Three categories were analyzed for DIF and only two items indicated large effect sizes and two with moderate effect sizes, which supports the construct validity of the remaining items.

4.4.2 Structural analyses. To investigate the extent to which the assessment can substantiate claims about participant understanding of specific concepts, four structural analyses were conducted: tetrachoric correlations, subscale alphas, exploratory factor analysis, and confirmatory factor analysis.

Tetrachoric correlations. The StatCI item pairs are shown graphically by a heat map (see Figure 13). The map has a mix of dark and light areas; the probability cluster of items had the darkest area, and several item pairs were also shaded darker. The rows for problematic items, Q6 and Q27, were light, indicating that they were not well correlated to the rest of the items.

Removing items. Before conducting the rest of the analyses, three problematic items were removed: Q5, Q6, and Q27. These three items had poor item discriminations, poor tetrachoric correlations to the rest of the items, and poor model fit indices for the two-parameter IRT model.

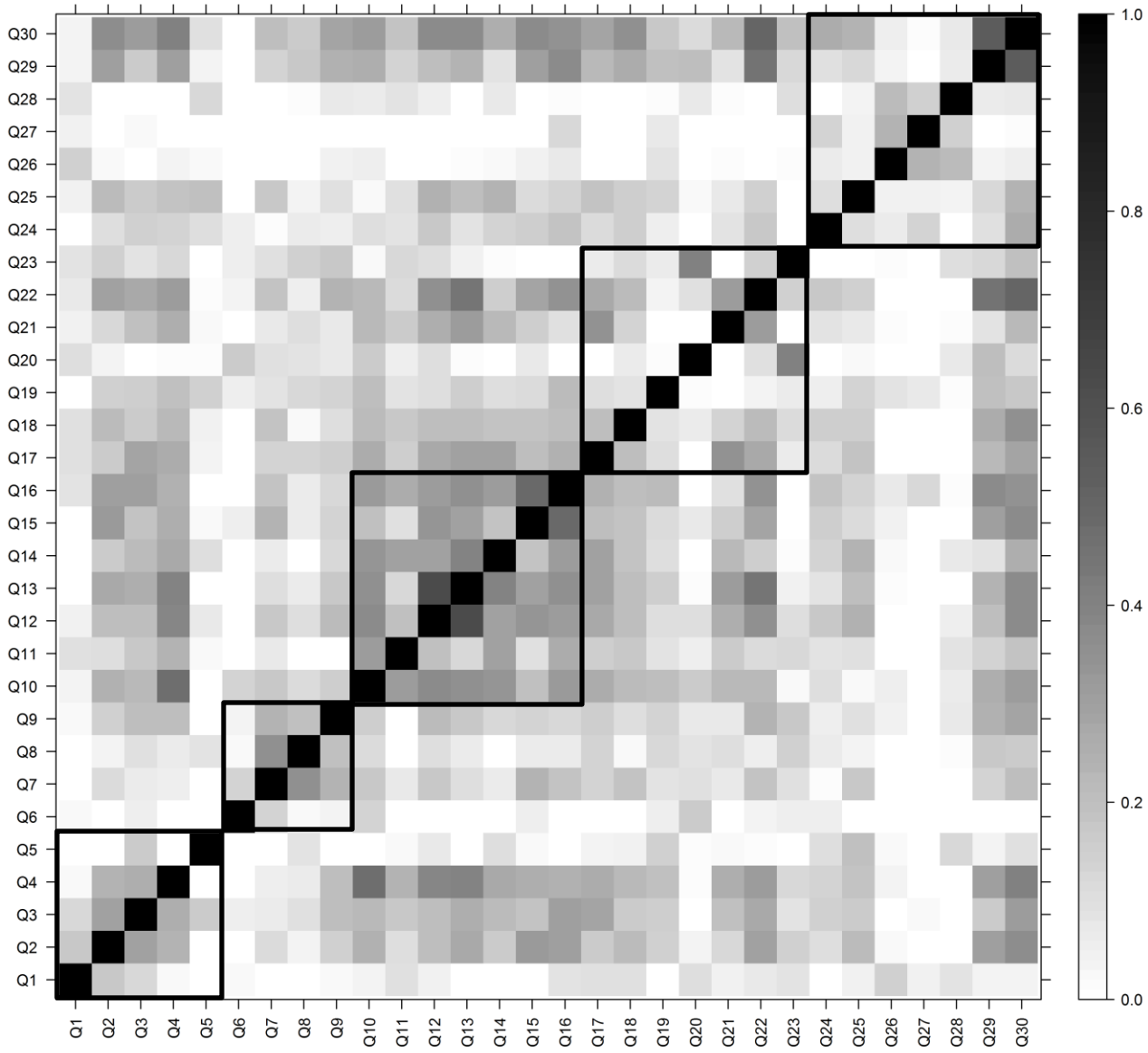


Figure 13. Tetrachoric correlations for study 3.

Although these items had been had been substantiated by content experts, it appeared that too many high-performing participants chose other distractors for these items.

Subscale alphas. Subscale alphas were estimated for each of the five concepts (see Table XV). Subscale values ranged from .32 to .63. Because the groups varied in test length, Spearman-Brown prophecy formula was also calculated for each subscale for comparison purposes. Given that alpha measures are a function of test length, these subscale alphas demonstrate reasonable reliabilities for assessments of this length.

TABLE XV

SUBSCALE ALPHAS FOR STUDY 3

Concept	α	α (2B)	N	If n=10
Univariate	.57	.23	8	.63
Correlation	.32	.34	4	.54
Probability	.63	.35	8	.67
Sampling	.45	.56	9	.48
Hypothesis	.36	.16	7	.44

Exploratory factor analysis. To examine the extent to which the participant data supported the theoretical structure for the assessment, an exploratory factor analysis was run. An oblimin rotation was used since there were strong inter-item correlations for items from differing conceptual clusters. Because the items were scored dichotomously, tetrachoric correlations were analyzed for the factor analysis.

A parallel analysis (Horn, 1965) indicated that an eleven-factor solution would yield the optimal amount of components (the same number of factors as in the Q-matrix) with the three items removed. A ten-factor solution was run since one of the components did not yield high factor loadings (Table XVI). The ten-component factor solution explained 49% of the total variance. Factor loadings less than .20 were suppressed. The resulting ten factors had substantial overlap with the intended conceptual groupings of the original Q-matrix (Appendix E). Bolded values denote items that do not align with the designated category. Three groups were not manifested in the groupings: measures of dispersion (however, Q5 was removed, so only Q4 remained in this category), sample size, and large sample theory. Probability was split into three groups: probability theory, probability logic, and events in probability. In addition, another category emerged: skewed distributions, which included Q3 about the most appropriate measures for distributions with outliers, and Q19 regarding a skewed, non-normally distributed sample.

Because the items loaded on ten different factors (rather than on one), these results support that claim that the assessment measures differentiable conceptual understanding commensurate to the original categories in the Q-matrix. There were some exceptions to this mapping; some items that were not originally in the same Q-matrix grouping ended up loading together on one factor (such as Q3 and Q19). Moreover, several items grouped on more than one factor (Q4, Q9, Q11, Q15, Q22, and Q30). This suggests that participants are using overlapping skills to answer these items.

The factor correlation matrix (Table XVII) shows that all of the groups were positively related. Those groups that were the most highly related were ones within the same subgroups. For example, FK3a- Probability Theory had a .37 and .39 correlation with FK3b- Events in Probability

TABLE XVI

EXPLORATORY FACTOR RESULTS FOR STUDY 3

	FK3a- Probability Theory	FK3b- Events in Probability	FK4c- Data Gathering	FK6b- Hypothesis Testing	FK3b- Probability Logic	FK6a- P-values	FK4a- Sampling	FK2- Correlation	FK1- Univariate	Skewed Distributions
Q13	1									
Q12	.46									
Q10		1								
Q4		**.36**							.21	
Q11		.22			.25		**.21**			
Q23			1							
Q20			.36							
Q29				.99						
Q30				.26					.46	
Q16					.83					
Q15					.5				**.26**	
Q14					.21					
Q25						.88				
Q21							**.68**			
Q17							.39			
Q8								.69		
Q7								.51		
Q9								.24	**.23**	
Q26										
Q24									**.3**	
Q22				**.24**			.25		**.29**	
Q2									.24	
Q3										.31
Q19										.28

TABLE XVII

FACTOR CORRELATION MATRIX FOR STUDY 3

	FK3a- Probability Theory	FK3b- Events in Probability	FK4c- Data Gathering	FK6b- Hypothesis Testing	FK3b- Probability Logic	FK6a- P-values	FK4a- Sampling	FK2- Correlation	FK1- Univariate	Skewed Distributions
FK3a- Probability Theory	1									
FK3b- Events in Probability	.37	1								
FK4c- Data Gathering	.09	.11	1							
FK6b- Hypothesis Testing	.24	.27	.18	1						
FK3b- Probability Logic	.39	.36	.04	.38	1					
FK6a- P-values	.27	.11	.02	.17	.24	1				
FK4a- Sampling	.39	.29	.04	.13	.22	.16	1			
FK2- Correlation	.09	.13	.16	.19	.12	.11	.08	1		
FK1- Univariate	.31	.12	.13	.36	.31	.19	.27	.16	1	
Skewed Distribution	.14	.11	-.05	.04	.14	.11	.15	-.05	.02	1

and FK3b- Probability Logic, respectively. Therefore, although the smaller grained Q-matrix groupings were separable, the superordinate categories were still related. FK3a- Probability Theory was also highly related to FK6b-Hypothesis Testing (.38) and FK4a-Sampling (.39), which agrees with the researcher's hypothesis that the skills required to answer these items correctly are inter-related.

Confirmatory factor analysis. The exploratory factor analysis helped to investigate the Q-matrix structure, which parsed the original 5 conceptual categories into smaller grained skills. The

researcher used a confirmatory factor analysis to investigate the functioning of the 5 conceptual categories, particularly to test the hypothesis that the 5-factor model fit the data and to evaluate alternative structural models. Given the number of latent and observed variables, as well as the anticipated effect size and desired statistical power level, 750 cases surpassed the minimum sample size of 200. A series of models were tested, and regression weights and model fit indices were used to make modifications to the model. Based on the recommendations of Hooper, Coughlan, and Mullen (2008), the researcher focused on the following fit indices: χ^2 test, the root mean square error approximation, the standardized root mean square residual, the comparative fit index, and the Parsimonious Normed Fit Index, as well as the Akaike information criterion. The researcher's first model was based on the Q-matrix. All the latent variables were allowed to covary, since the developer hypothesized that participants were drawing upon an overall knowledge of statistical concepts and thinking in addition to understanding of particular concepts. Overall, this model fit the data moderately well; although the model fit statistics met the recommended values, several observed variables had low or negative regression weights (Figure 14).

A second model was tested that eliminated observed values with the low regression weights: Q1 (.08), Q20 (.11), Q26 (.03), and Q28 (.01). It also detached several observed values from the latent variables: Q10 from sampling (.01), Q11 from sampling (-.06), Q17 from univariate (.04), Q18 from univariate (.13), Q21 from sampling (-.06), Q29 from univariate (-10) and Q30 from univariate (-12).[3] This model fit the data better, as shown in the difference of fit statistics in Table XVIII. The

[3] When factors are correlated, standardized coefficients are regression coefficients and can be greater than one in magnitude (Jöreskog, 1999).

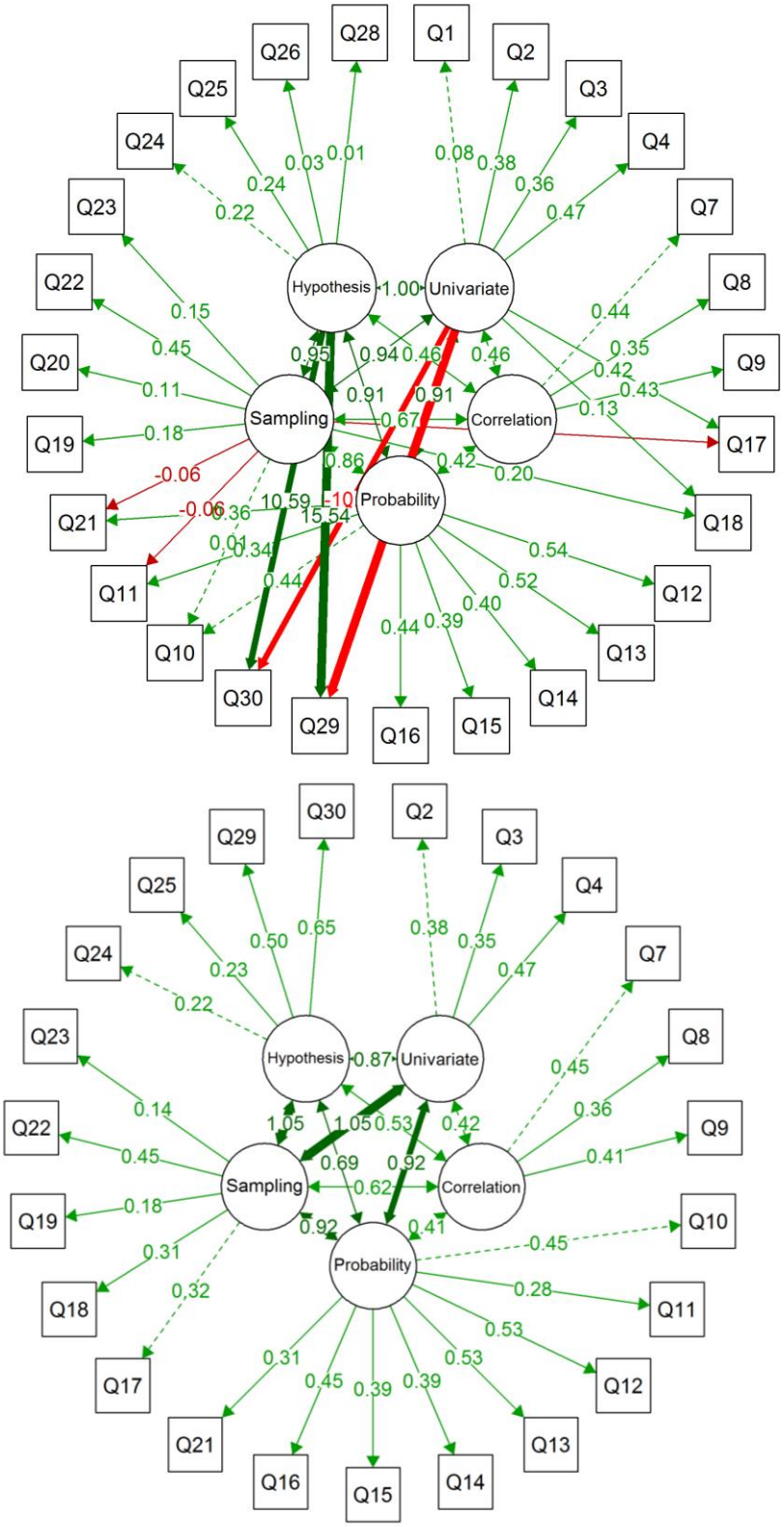


Figure 14. Confirmatory factor analysis models for study 3. Top is model 1; bottom is model 2.

TABLE XVIII

CONFIRMATORY FACTOR ANALYSIS MODEL FIT INDICES

Measure	Recommended value	Model 1	Model 2	Model 3
Number of parameters		71	56	56
Degrees of freedom	–	307	220	220
Global (absolute) indexes				
χ^2	Low relative to df	515	319	331
Root-mean-square-error Approximation	< .07	.027	.024	.026
Standardized root mean square residual	< .08	.037	.035	.037
Comparative (incremental) fit indexes				
Comparative fit index	< .95	.89	.93	.93
Comparisons among multiple models				
Akaike information criterion	Lower values	25833	22033	22215

final model indicates high loadings among the five latent variables, providing evidence for the existence of a broad conceptual statistics understanding (Figure 15). This model displayed good model-data fit, with indexes meeting the recommended cutoff values (CFI > .9, RMSEA < .05). All of the latent variables were correlated, with sampling highly related to univariate, probability, and hypothesis testing (Table XVIV). The "correlation" latent variable was the least related to the other four categories. Two correlations were extremely high (Sampling-Univariate, 1.05 and Hypothesis-Sampling, 1.14). Correlation coefficients may be greater than one when they are not statistically distinguishable.

TABLE XIV
INTER-FACTOR CORRELATION MATRIX OF MODEL 2

	Univariate	Correlation	Probability	Sampling	Hypothesis
Univariate	1				
Correlation	0.42	1			
Probability	0.93	0.42	1		
Sampling	1.05	0.63	0.97	1	
Hypothesis	0.85	0.53	0.67	1.14	1

Given that the concepts were so highly related, the researcher tested a third model that simplified the construct space further using a bifactor model. To run this model, the researcher constrained the covariances of the latent variables to 0 and included a general factor. Such a model allows the researcher to examine how both the general factor and the subscales contribute uniquely to the overall model. One of the original assumptions is that the StatCI assesses overall statistical understanding, which is why the researcher used an oblique rotation in the exploratory factor analysis. Once the model was run, the researcher pruned connections with low factor loadings (<.10). In this way, the final model would be more parsimonious, and the relations between the items and the groupings would be clearer. Items in the univariate and sampling group had low loadings on the subgroup latent factor, but loaded more highly on the general factor; therefore, the univariate and sampling latent variables were removed from the model. Some items only had high loadings on the general factor, others loaded onto the subgroup latent factor, while several items grouped on both the general and subgroup factor. When comparing the fit indices (Table XV), this model performed slightly worse than model 2 even though the degrees of freedom were identical.

These models indicate that items are related to particular subgroups as well as a general factor. The advantage of running a confirmatory factor analysis is that it indicates which subgroups are performing particularly well, and which items are more related to the hierarchal factor. The relationships between the items and subgroups are multifaceted and overlapping. Based on this assessment, student understanding of statistics is composed of both separable dimensions and well as an overall level of conceptual understanding with a high degree of relatedness between concepts. This supports the claim that participants' scores on the StatCI can be used to indicate their understanding of specific concepts as well as overall understanding of statistics.

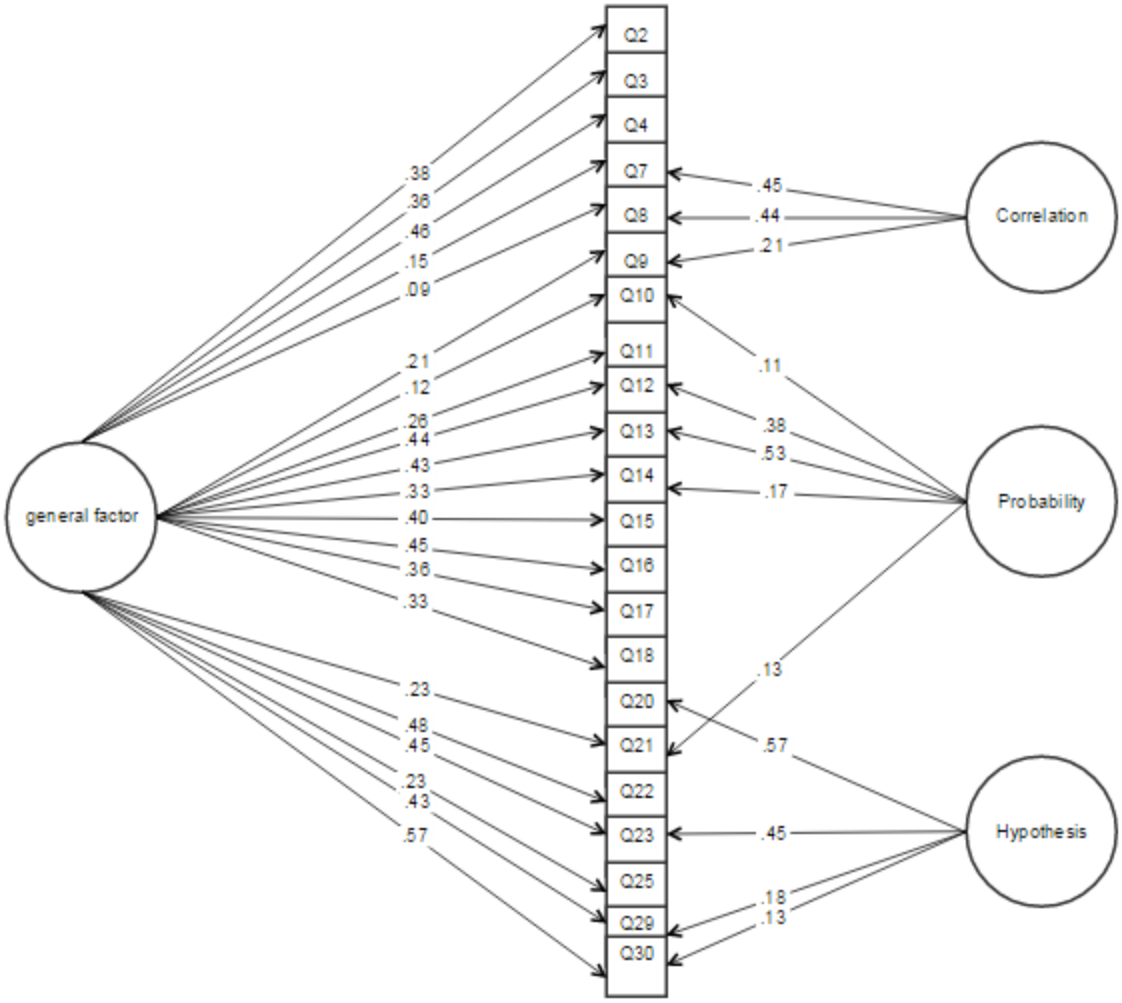


Figure 15. Confirmatory factor analysis model for study 3. Model 3 is a bifactor model.

Structural analyses with participants scoring higher than chance. An assessment may not yield particularly strong structural properties when the participants answering the items are guessing or do not know basic statistics. To score above chance, a participant had to answer more than 7 items correctly (since there were 4 options for each item). Out of the 750 participants, 641 scored above 7 on the StatCI (85.47%). The researcher re-ran the structure analyses with this subsample to

investigate the effects of answering pattern (indicating knowledge threshold) on the inventory's structural properties.

The researcher re-ran an exploratory factor analysis on this subsample with the same procedure as the previous analysis (Table XX). For ease of comparison to the previous analysis, the same number of factors were used. Three items with low loadings were removed: Q1, Q9, and Q18. The exploratory factor analysis with the larger sample had six items removed (Q1, Q5, Q6, Q18, Q27, and Q28). This solution explained 45% of total variance (less than the 49% total variance explained by the larger sample). There were many similarities to the original exploratory factor analysis. One notable difference was that there were more items that loaded on FK6a-P-values, since Q27 and Q28 were retained in this model. However, p-values was also negatively related to the other categories, as indicated in the factor correlation matrix (Table XXI).

The researcher also re-ran model 2 and 3 of the previous confirmatory factor analysis with the subsample (Figure 16). The models fit the larger sample slightly better than the smaller sample for the most part; for the larger sample, the root mean square error approximation and the root mean square residual were smaller and the comparative fit index was larger. The Akaike information criterion was lower for the smaller sample. However, this may be a function of having a larger N and some overfit in the first model. Since the confirmatory factor analysis results remained relatively stable across samples, the sample of participants who scored less than seven did not have a significantly detrimental impact on the overall structural properties of the instrument.

TABLE XX

EXPLORATORY FACTOR ANALYSIS RESULTS FOR STUDY 3 SUBSAMPLE

	FK3a- Probability Theory	FK6b- Hypothesis Testing	FK3b- Events in Probability	FK4c- Data Gathering	FK3b- Probability Logic	FK2- Correlation	FK4a- Sampling	General	FK6a- P-values	FK1- Univariate
Q13	1									
Q12	.47									
Q29		1								
Q10			.99							
Q4	.21		.29							
Q6			.23							
Q23				1						
Q20				.37						
Q16					.92					
Q15					.38					
Q7						.87				
Q8						.38				
Q21							.62			
Q17							.44			
Q11			.23		.21		.29			
Q5								.46		
Q19								.43		
Q25								.41		
Q3								.33		
Q14							.21	.24		
Q28									.68	
Q26									.34	
Q27					.23				.23	
Q22		.3								.41
Q30		.33								.38
Q24										.36
Q2										.26

TABLE XXI

FACTOR CORRELATION MATRIX FOR STUDY 3 SUBSAMPLE

	FK3a- Probability Theory	FK6b- Hypothesis Testing	FK3b- Events in Probability	FK4c- Data Gathering	FK3b- Probability Logic	FK2- Correlation	FK4a- Sampling	General	FK6a- P-values	FK1- Univariate
FK3a- Probability Theory	1									
FK6b- Hypothesis Testing	.16	1								
FK3b- Events in Probability	.33	.19	1							
FK4c- Data Gathering	-.03	.1	.01	1						
FK3b- Probability Logic	.3	.27	.27	-.06	1					
FK2- Correlation	.08	.1	.13	.03	.13	1				
FK4a- Sampling	.31	.07	.18	-.07	.11	0	1			
General	.15	.16	.17	.02	.22	.05	.18	1		
FK6a- P-values	-.15	-.08	-.05	.01	-0.1	-.15	-.13	-.02	1	
FK1- Univariate	.25	.25	.07	.03	.2	.11	.23	.13	-.1	1

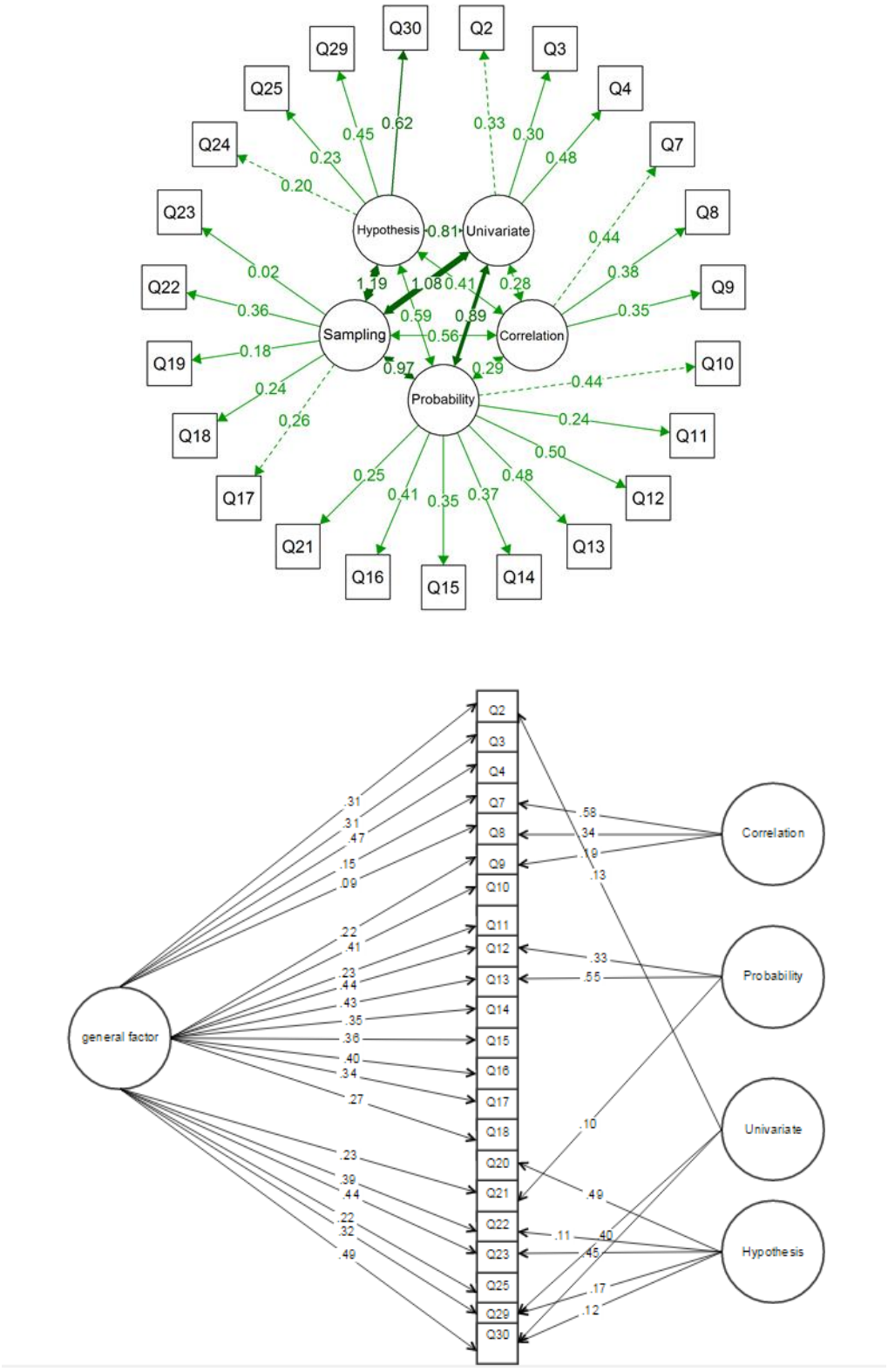


Figure 16. Confirmatory factor analysis models for study 3 using a subsample of participants who correctly answered more than 7 items.

4.4.3 Diagnostic analysis. Diagnostic classification analysis was also used to investigate the diagnostic strength of the StatCI. This class of models predicts the probability of observed responses from latent categorical variables using the developer's Q-matrix. Unlike IRT that assigns respondents to a single score on a continuous latent scale, diagnostic classification models assign respondents to a discrete multidimensional skills profile (i.e., master or non-master). There are three types of diagnostic models: compensatory, non-compensatory, and general (Templin & Henson, 2010). Compensatory models allow a strong, positive interaction in skill attributes; mastery of one skill can compensate for lack of another skill. Non-compensatory models posit a negative interaction between items; all attributes must add on to the other in order to produce the correct answer. General models allow a mix of compensatory and non-compensatory models for each item on the same test.

First, the Q-matrix was revised by removing the three items previously identified to have poor reliabilities (Q5, Q6, and Q27). The category, "measures of spread," was consequently removed as well because of the low numbers for item membership. Next, to assess relative fit, three different models were tested on the data: a compensatory model (DINO, Templin & Henson, 2006), a non-compensatory model (NC-RUM, DiBello et. al, 1995), and a generalized model (G-DINA, de la Torre, 2011). Lower values for the model fit indices indicate better fitting models. MADRCOV is an absolute model fit of effect size; the smaller the effect size, the better the model fit. The generalized model had the best fit measures for all but two of the fit indices (Table XXII).

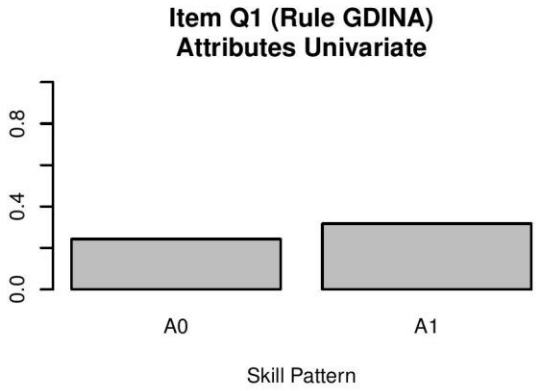
TABLE XXII

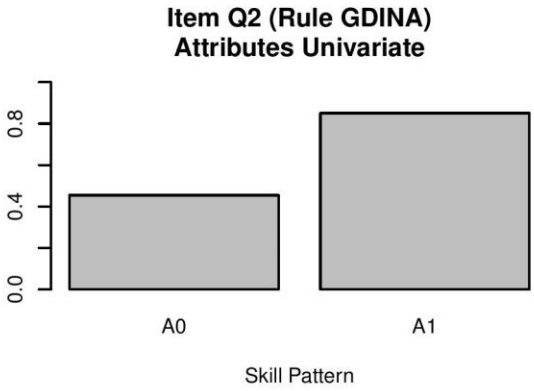
DIAGOSTIC CLASSIFICATION MODEL FIT INDICES

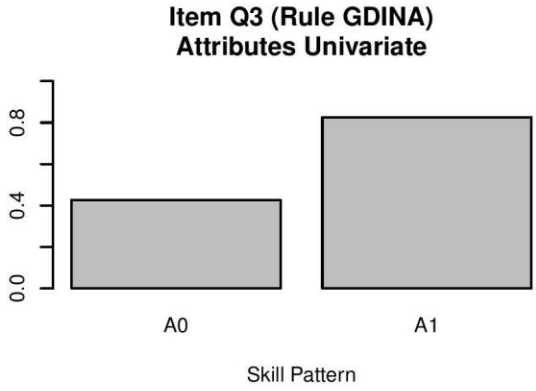
Measure	DINO	NC-RUM	G-DINA
Number of parameters	85	106	131
Log-likelihood	-10962.84	-10937.23	**-10857.10**
AIC	22095.67	22086.45	**21976.20**
BIC	**22488.38**	22576.18	22581.43
AICc	22117.69	22121.73	**22032.16**
CAIC	**22573.38**	22682.18	22712.43
MADRCOV	0.742	1.011	**0.604**

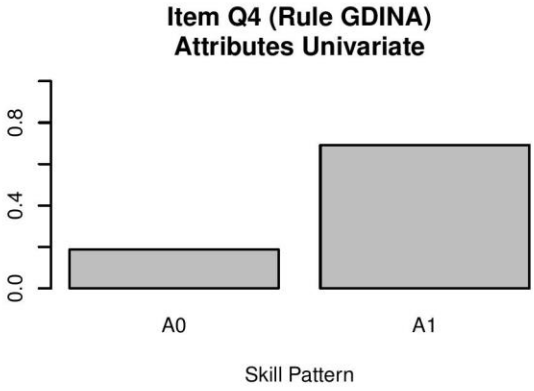
After estimating the G-DINA model, the Q-matrix was validated by recalculating the item parameters and item discrimination parameters; the Q-matrix rows were determined by maximizing the estimated item discrimination index (de la Torre, 2008). Figure 17 show each item and their respective skill pattern probability profiles. Each bar is associated with a certain combination of skills, where each number following the "A" indicates the skill and whether the participant has mastered that skill (where 1 indicates mastery and 0 non-mastery). For example, the developer proposed two concepts associated with Q10: probability and logic. The "A00" bar indicates the probability that a participant answers the item correctly given that they have not mastered any of the two proposed skills. The "A10" bar indicates the probability of answering the item correctly when the participant has mastered the first but not the second skill. Participants have a .60 probability of answering Q10 correctly if they have mastered both skills.

Ideally, an item should have a very low probability of being answered correctly if a participant did not master the skills required for the item; the item should also have a very high probability of being answered correctly if the participant has mastered all of the skills necessary to


Item Q1 (Rule GDINA)
Attributes Univariate
0.0
0.4
0.8
A0
A1
Skill Pattern


Item Q2 (Rule GDINA)
Attributes Univariate
0.0
0.4
0.8
A0
A1
Skill Pattern

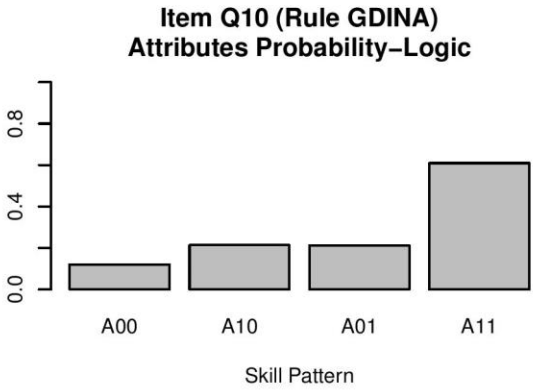

Item Q3 (Rule GDINA)
Attributes Univariate
0.0
0.4
0.8
A0
A1
Skill Pattern

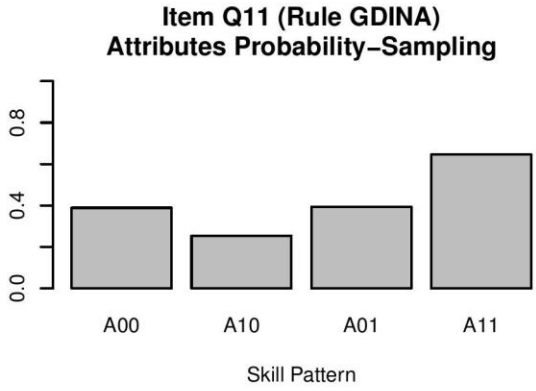

Item Q4 (Rule GDINA)
Attributes Univariate
0.0
0.4
0.8
A0
A1
Skill Pattern

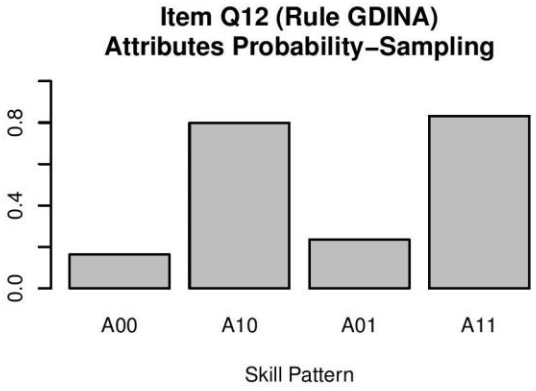

Item Q7 (Rule GDINA)
Attributes Correlation
0.0
0.4
0.8
A0
A1
Skill Pattern


Item Q8 (Rule GDINA)
Attributes Correlation
0.0
0.4
0.8
A0
A1
Skill Pattern

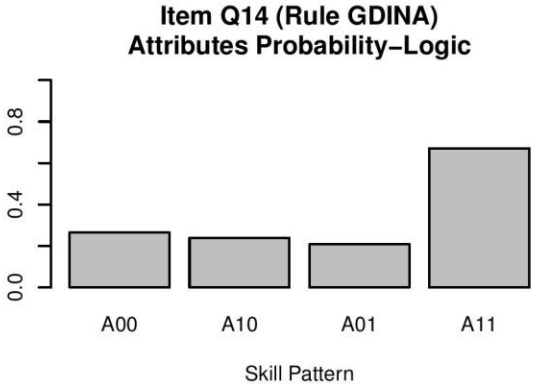

Item Q9 (Rule GDINA)
Attributes Correlation
0.0
0.4
0.8
A0
A1
Skill Pattern

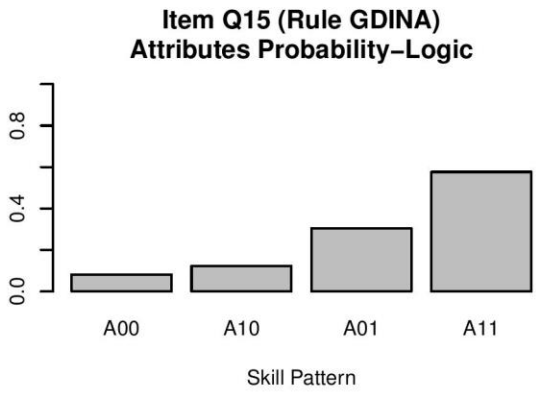

Item Q10 (Rule GDINA)
Attributes Probability−Logic
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

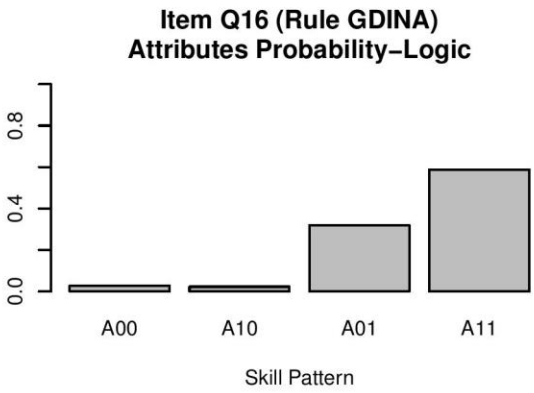

Item Q11 (Rule GDINA)
Attributes Probability−Sampling
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

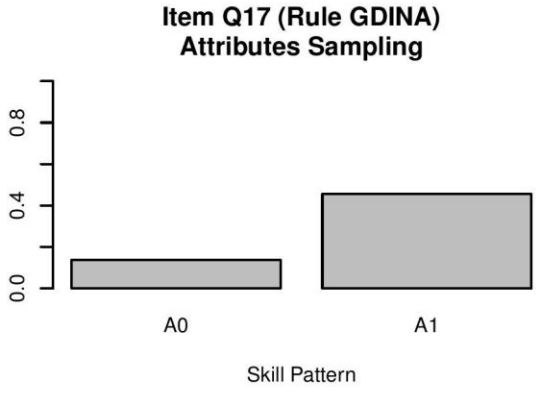

Item Q12 (Rule GDINA)
Attributes Probability−Sampling
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

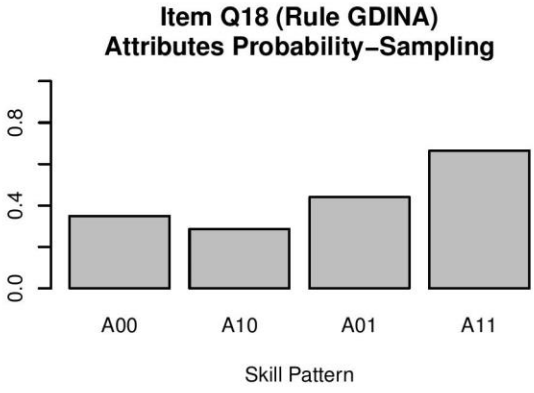

Item Q13 (Rule GDINA)
Attributes Probability
0.0
0.4
0.8
A0
A1
Skill Pattern

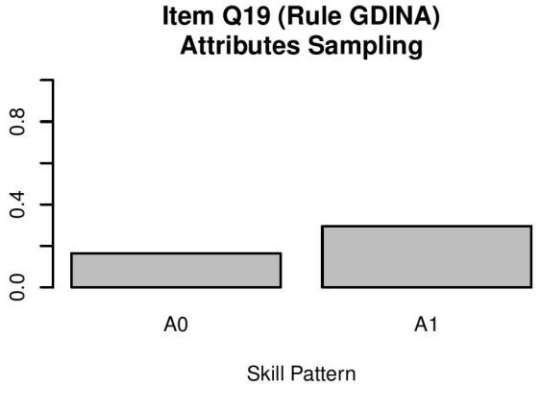

Item Q14 (Rule GDINA)
Attributes Probability−Logic
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

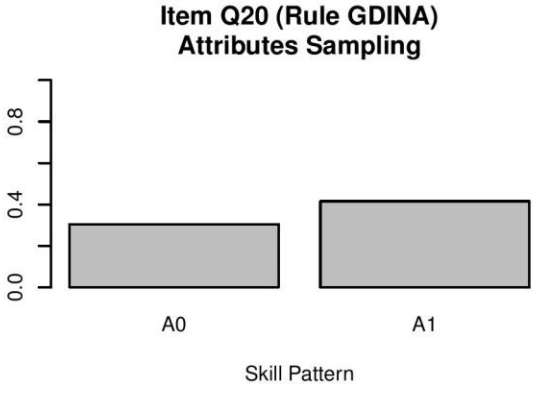

Item Q15 (Rule GDINA)
Attributes Probability−Logic
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

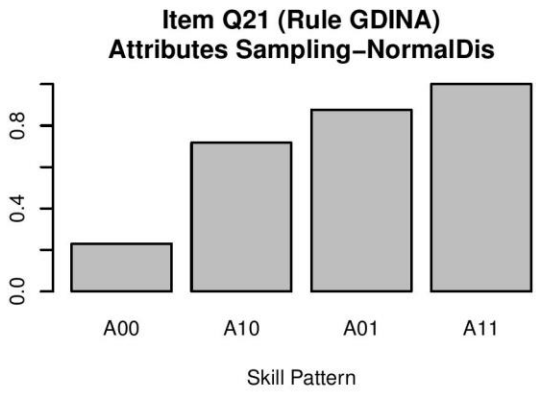

Item Q16 (Rule GDINA)
Attributes Probability−Logic
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

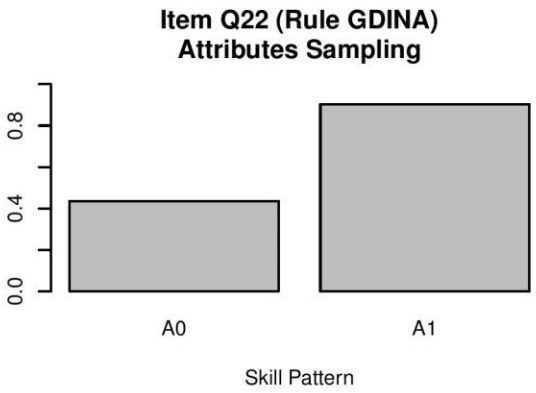

Item Q17 (Rule GDINA)
Attributes Sampling
0.0
0.4
0.8
A0
A1
Skill Pattern

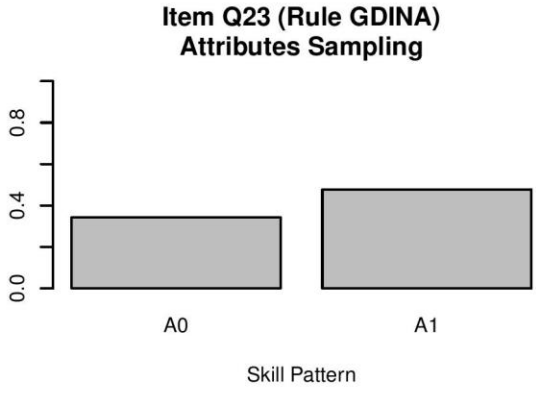

Item Q18 (Rule GDINA)
Attributes Probability−Sampling
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern

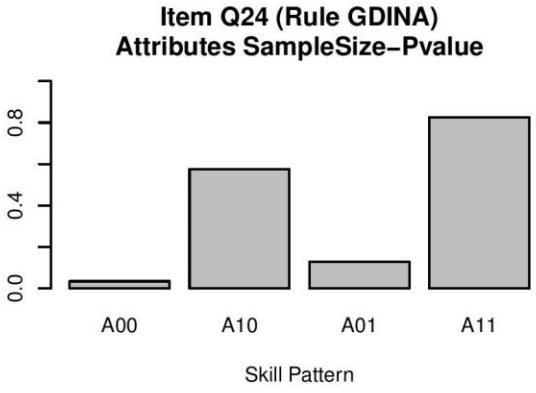

Item Q19 (Rule GDINA)
Attributes Sampling
0.0
0.4
0.8
A0
A1
Skill Pattern

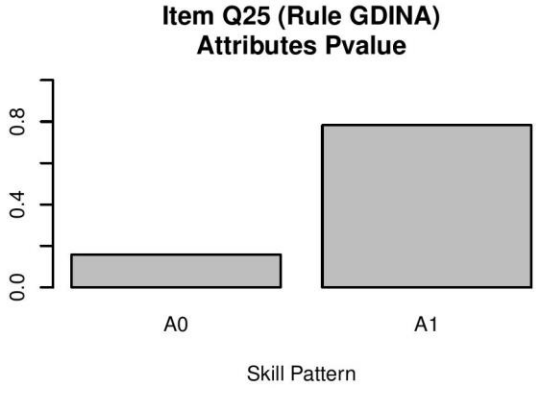

Item Q20 (Rule GDINA)
Attributes Sampling
0.0
0.4
0.8
A0
A1
Skill Pattern

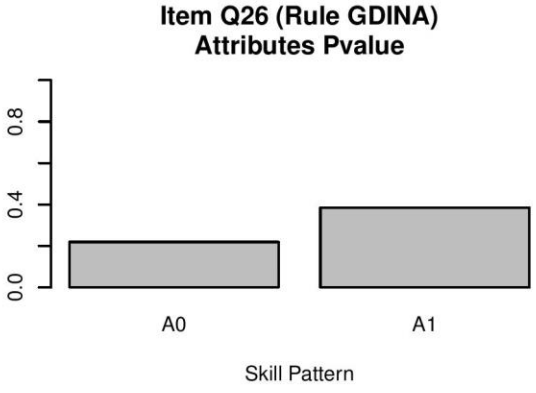

Item Q21 (Rule GDINA)
Attributes Sampling−NormalDis
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern


Item Q22 (Rule GDINA)
Attributes Sampling
0.0
0.4
0.8
A0
A1
Skill Pattern


Item Q23 (Rule GDINA)
Attributes Sampling
0.0
0.4
0.8
A0
A1
Skill Pattern


Item Q24 (Rule GDINA)
Attributes SampleSize−Pvalue
0.0
0.4
0.8
A00
A10
A01
A11
Skill Pattern


Item Q25 (Rule GDINA)
Attributes Pvalue
0.0
0.4
0.8
A0
A1
Skill Pattern


Item Q26 (Rule GDINA)
Attributes Pvalue
0.0
0.4
0.8
A0
A1
Skill Pattern

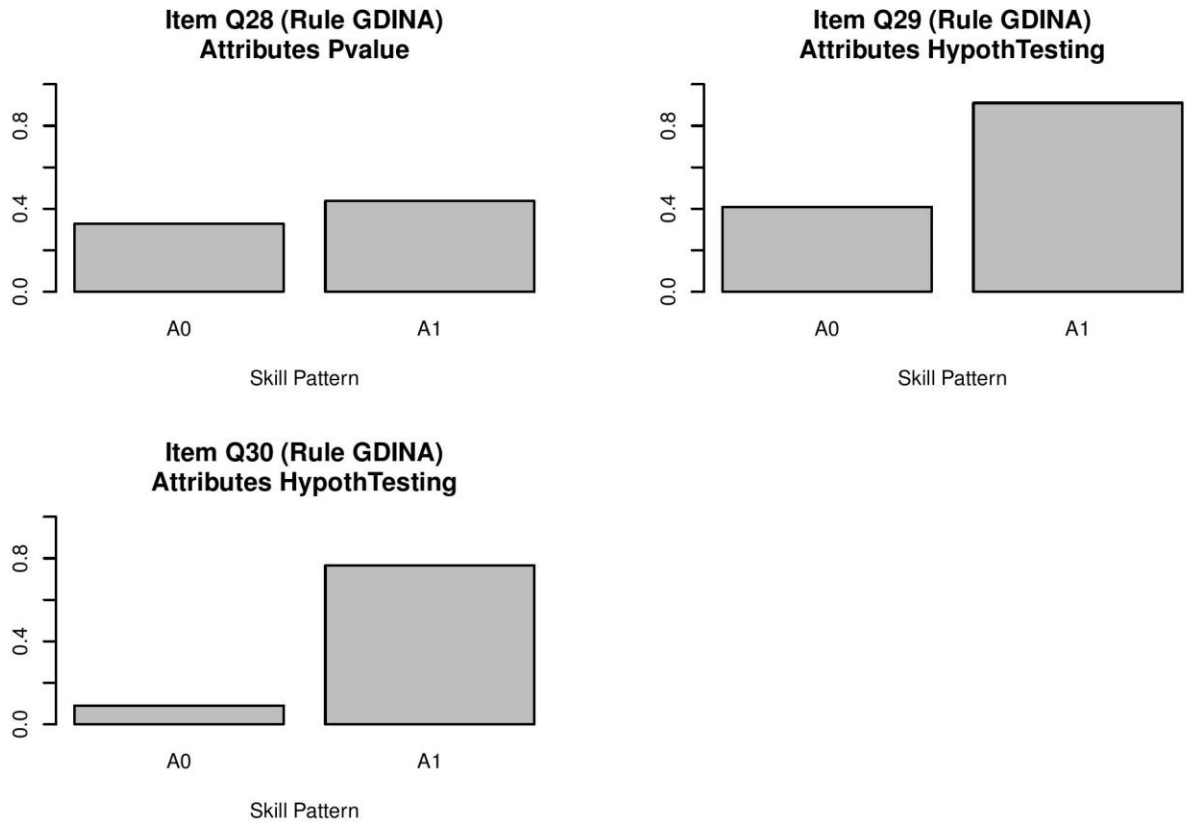


Figure 17. Conceptual understanding pattern profiles.

answer the item. Q4, for example, follows this ideal pattern. The plot for Q28, on the other hand, indicates that participants do not have a much higher probability of answering the item correctly if they have all the skills necessary for answering the item. These graphs also indicate interaction among skills and the probability of answering the item correctly; some of the item skills are compensatory, such as in Q14. In this case, the probability of answering the item correctly is more than the sum of mastering either the first item or the second item.

Table XXIII indicates the difficulty of the attributes. Sample size was the easiest skill, mastered by 67% of the participants, whereas p-values was the most difficult skill, mastered by only 12% of the participants.

TABLE XXIII
DIFFICULTY BY CONCEPT

Univariate	.34
Correlation	.37
Probability	.46
Sampling	.44
Sample Size	.67
P-values	.12
Hypothesis Testing	.38
Normal Distributions	.38
Logic	.59

This model was also used to generate student skill profiles, in which the probability of a student mastering each concept can be calculated given a response pattern. Two students with the same scores may have different mastery probabilities depending on their response patterns. For example, for two participants in study 3 who scored 15/30, participant A had a 23% probability of mastering correlation while participant B had a 91% probability of mastering correlation. Appendix I is a sample report for a random participant from study 3.

Overall, diagnostic classification modeling helped to refine and validate the Q-matrix design. The resulting skill pattern profiles provide information on how distinct concepts contribute to probability of mastery and how certain skills interact with one another. It would also be possible to create performance profiles for each participant, which would be useful for professors seeking information about the mastery levels of their students. Based on these results, the researcher created a refined Q-matrix (Table XXIV). The measures resulting from this analysis indicate the extent to which mastery of particular concepts contributes to the probability of answering the item correctly. This provides evidence for the diagnostic capability of the StatCI and the relationship between concepts and items on the assessment.

TABLE XXIV

NEW Q-MATRIX

	Univariate	Correlation	Probability	Sampling	Sample Size	P-values	Hypothesis Testing	Normal Distrib	Logic
Q1	1	0	0	0	0	0	0	0	0
Q2	1	0	0	0	0	0	0	0	0
Q3	1	0	0	0	0	0	0	0	0
Q4	1	0	0	0	0	0	0	0	0
Q7	0	1	0	0	0	0	0	0	0
Q8	0	1	0	0	0	0	0	0	0
Q9	0	1	0	0	0	0	0	0	0
Q10	0	0	1	0	0	0	0	0	1
Q11	0	0	1	1	0	0	0	0	0
Q12	0	0	1	1	0	0	0	0	0
Q13	0	0	1	0	0	0	0	0	0
Q14	0	0	1	0	0	0	0	0	1
Q15	0	0	1	0	0	0	0	0	1
Q16	0	0	1	0	0	0	0	0	1
Q17	0	0	0	1	0	0	0	0	0
Q18	0	0	1	1	0	0	0	0	0
Q19	0	0	0	1	0	0	0	0	0
Q20	0	0	0	1	0	0	0	0	0
Q21	0	0	0	1	0	0	0	1	0
Q22	0	0	0	1	0	0	0	0	0
Q23	0	0	0	1	0	0	0	0	0
Q24	0	0	0	0	1	1	0	0	0
Q25	0	0	0	0	0	1	0	0	0
Q26	0	0	0	0	0	1	0	0	0
Q28	0	0	0	0	0	1	0	0	0
Q29	0	0	0	0	0	0	1	0	0
Q30	0	0	0	0	0	0	1	0	0

4.4.4 Distractor analysis. To support the claim that participant scores can be used to indicate their propensity for misconceptions, the researcher analyzed the distractor data in three ways. First, cross tabulations and chi-squares were used to investigate the hypothesis that respondents hold similar misconceptions across items. Second, to investigate how ability levels corresponded to distractors, the researcher analyzed the data using two types of item response theory-based polytomous scoring methods: the partial credit model and the nominal response model.

Cross tabulations. The researcher analyzed cross tabulations of item pairs with common distractors. The researcher hypothesized that participants would hold similar misconceptions across items. However, the results indicated that certain misconceptions were more prevalent than others. In other words, participants were more likely to hold *particular* misconceptions across items; other misconceptions were less robust across items. Four examples are provided as an illustration in this section. First, for the contingency table of Q12 and Q13, there was no difference in cell counts from random chance, except for the two distractors corresponding to the equiprobability bias (Table XXV).

TABLE XXV

CONTINGENCY TABLES FOR SELECTED PROBABILTIY DISTRACTORS

	Q13	Base Rate Fallacy	Conjunction	Equiprobability Bias
Q12	Base Rate Fallacy	10	9	7
	Conjunction	6	12	24
	Equiprobability Bias	16	27	**81**

χ^2 = 15.745, df = 4, p-value = .003

	Q21	N/A	Equiprobability Bias	Reliability not related to sample size
Q12	Base Rate Fallacy	8	15	7
	Conjunction	8	10	9
	Equiprobability Bias	20	**53**	40

χ^2= 3.6293, df = 4, p-value = .459

	Q14	Larger sample = more error	Equiprobability Bias	N/A
Q12	Base Rate Fallacy	18	21	5
	Conjunction	16	35	8
	Equiprobability Bias	40	**105**	33

χ^2= 6.8759, df = 4, p-value = .137

Second, for the correlation pair, Q7 and Q8, there did not seem to be strong evidence that participants were choosing distractors corresponding to correlation and causation; however, there was a large number of responses for the misconception associated with if A is correlated to B and B is correlation to C, then A is correlated to C (Table XXVI).

TABLE XXVI

CONTINGENCY TABLE FOR SELECTED CORRELATION DISTRACTORS

	Q8	If XY + YZ are correlated >> XZ is correlated	Correlation = Causation	Correlation = Causation
Q7	Correlation = Causation	42	44	53
	If XY + YZ are correlated >> XZ is correlated	**78**	53	15
	Correlation = Causation	6	5	4

χ^2= 88.377, df = 4, p-value = .001

Third, there was also a high degree of relationship between two distractors tapping into the misconception that larger samples lead to normal distributions. For Q18, this distractor was, "The distribution of the new sample will look more like a normal distribution." For Q19, the distractor was, "Increase the sample size so that the distribution comes closer to a normal distribution." Table XXVII shows the cross tabulations for each distractor.

TABLE XXVII

CONTINGENCY TABLE FOR SELECTED SAMPLING DISTRACTORS

	Q19	Increasing ss = normal distribution	Need to survey population	Rerun if sample is biased
Q18	Increasing ss = normal distribution	**137**	35	14
	Increasing ss = no change	36	24	8
	Same means	38	20	14

χ^2= 18.113, df = 4 p-value = .004

Fourth, another misconception that showed consistency across two items was the bigger the sample, the more likely the subsamples will be exactly equal to the population distribution. For Q20, the associated distractor is, "The more students the researcher includes in the survey, the more likely that there will be exactly the same number of males and females in the sample." For Q21, "The more fish the biologist includes in the sample, the more likely there will be exactly 70% blue fish and 30% red fish." Table XXVIII shows the cross tabulations for these two items.

TABLE XXVIII

CONTINGENCY TABLE FOR SELECTED SAMPLING DISTRACTORS

	Q23	Sample must be representative	Gambler's Fallacy	More in sample, more exact
Q20	Sample must be representative	19	18	37
	Gambler's Fallacy	26	23	53
	More in sample, more exact	23	22	**125**

χ^2= 17.368, df = NA, p-value = .002

More examples of how response counts were linked across examples could be generated; the differences within the answering distributions for these four are particularly pronounced. These cross tabulations indicate that certain misconceptions hold across multiple items, suggesting that the StatCI could be used to diagnose learner misconceptions.

Partial Credit Model. Item response theory can also be used to investigate ability levels associated with distractors (Briggs, Alonzo, Schwab, & Wilson, 2006). The researcher ordered the distractors in a hypothesized progression and analyzed the extent to which the responses to the multiple-choice answers show evidence for a progression of responses. One example is illustrated in this section.

The first three items tested understanding of choosing the most appropriate measure of central tendency depending the data. The hypothesized learning progression was that novices will tend to use the very broad heuristic that mean is always the most appropriate way of measuring central tendency. The responses to the multiple-choice answers show evidence for this trajectory

(Figure 18). The blue curve, labeled 4, is the correct answer, while the black and red curves are associated with the belief that mean is always the best measure of central tendency. For Q1 and Q2, participants with the lowest theta had the highest probability of choosing "Mean because it is the best measure for summarizing numeric data." Participants with slightly higher ability levels demonstrated that they were using a more sophisticated heuristic for choosing the measure of central tendency. For Q1, mid-range thetas had a higher probability of choosing mode, probably because they recognized that the data was discrete. For Q2, participants with mid-range thetas were more likely to choose median, likely because they recognized that zip codes are also discrete data. Although Q3 also assessed measures of central tendency, most participants chose mean and participants with lower thetas were more likely to choose median. These results support the developer's hypothesis, but also shows that certain items are more effective at differentiating pre-identified learning progressions.

The partial credit model may be useful when the developer can hypothesize the knowledge trajectory a priori, but often this is not always the case. The polytomous scoring method in the following section can be useful when the developer does not have an assumption about the relationship between common misunderstandings represented by the distractor choices.

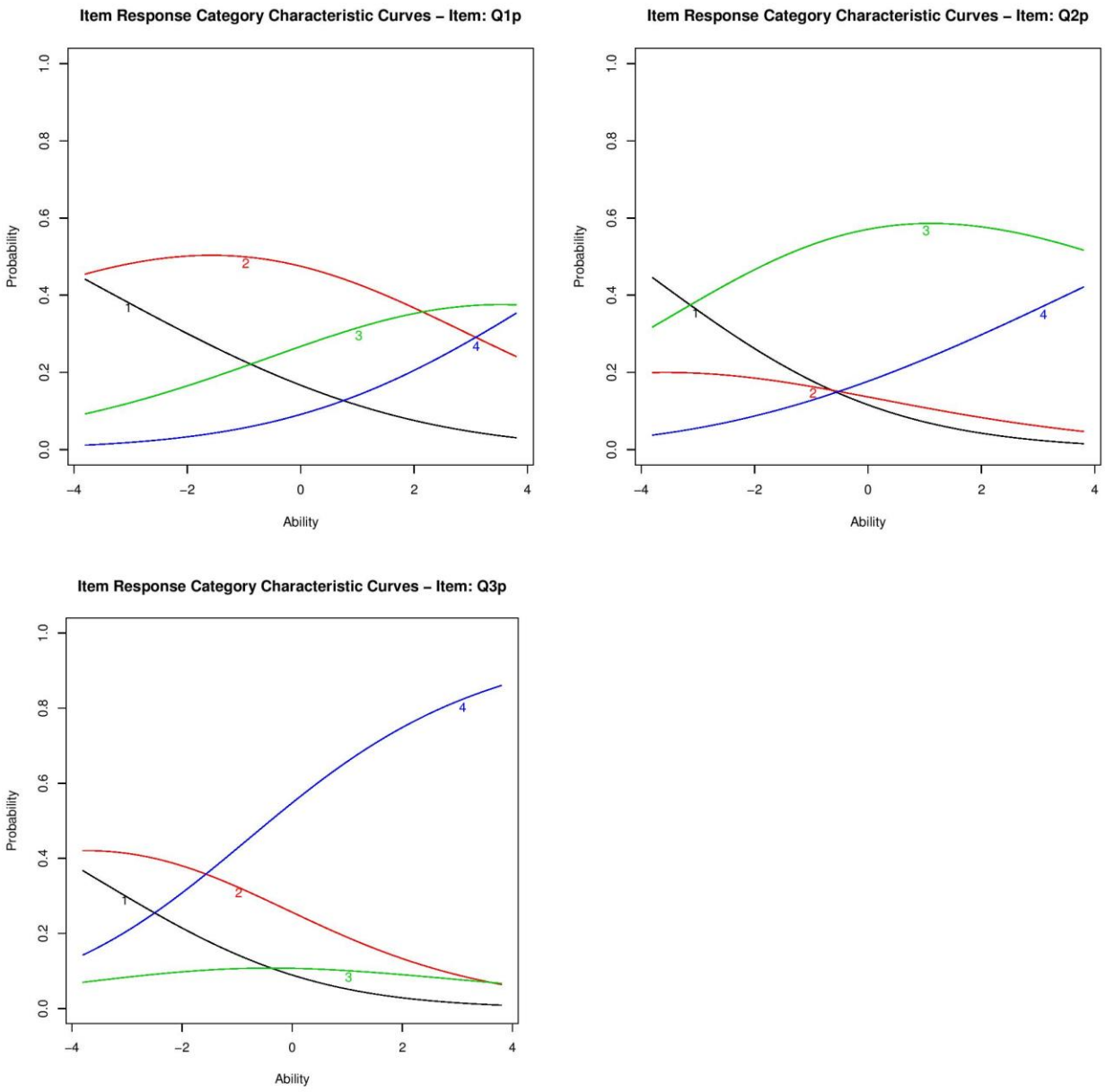


Figure 18. Item response curves for Q1, Q2, Q3 in the univariate category.

Nominal response model. Bock's (1972) nominal response model is another method that can be used to investigate the relationship between ability level and selected response. Unlike the partial credit model, this polytomous scoring model does not assume an inherent order in the answer choices. All of the items were examined using this model. Three examples of items that had distinct category characteristic curves are presented here to illustrate the application of this method.

Q17 asks, "A technician selects 10 oranges from an orchard. The mean weight of this sample is 8 ounces. Based on this information, what can we infer?" Category response curves are graphed in Figure 19. Respondents with the lowest ability level tended to choose B, "The sample is normally distributed." Respondents with the next highest ability level tended to choose A, "The population mean should also be 8 ounces, since sampling ensures that the sample will be representative of the population." Participants with ability levels slightly higher than average were more likely to pick response D, "If the technician adds one other orange to the sample, the new sample mean will be closer to the population mean." Respondents with the highest ability level generally chose the correct answer C, "The technician would have to weigh all the oranges to find the population mean."

These answers illustrate an increasingly sophisticated understanding of the relationship between samples and populations. Novice statistic learners often believe in the simplistic heuristic that all samples will be normally distributed (confusing sampling distributions and samples). More knowledgeable learners may believe that samples generally look like the population distribution. Learners that think that adding one case to the sample will make it look more like the population distribution are *sometimes* but not *always* correct; this misunderstanding is a misapplication of the law of large numbers.

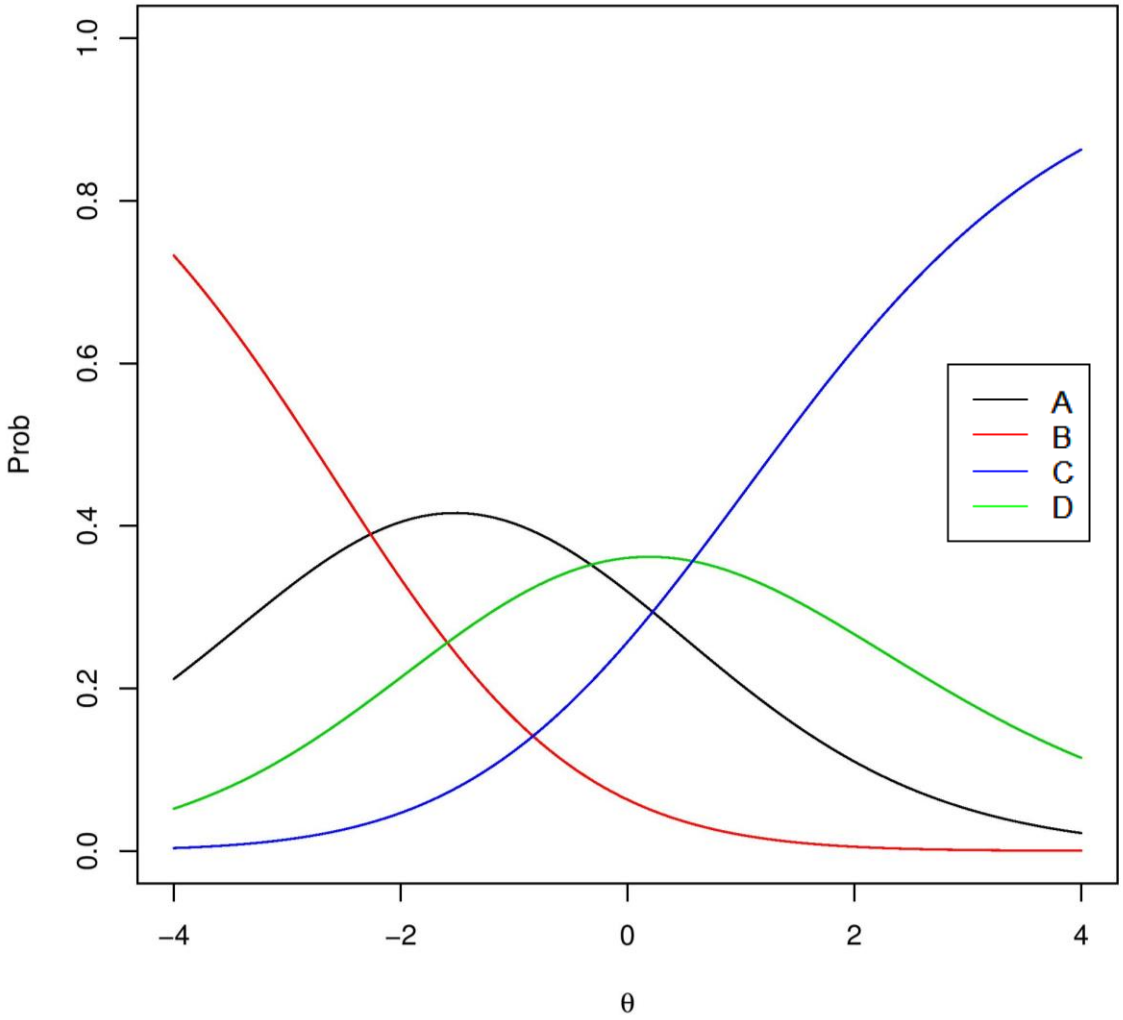


Figure 19. Category characteristic curves for Q17.

Most of the other items on the StatCI did not have such cleanly differentiable category characteristic curves as Q17, even the items that had strong reliabilities and performance indices. The category character curves of Q30, one of the most informative and discriminating items on the inventory, are shown in Figure 20. The item asks, “A teacher wants to support the claim that an

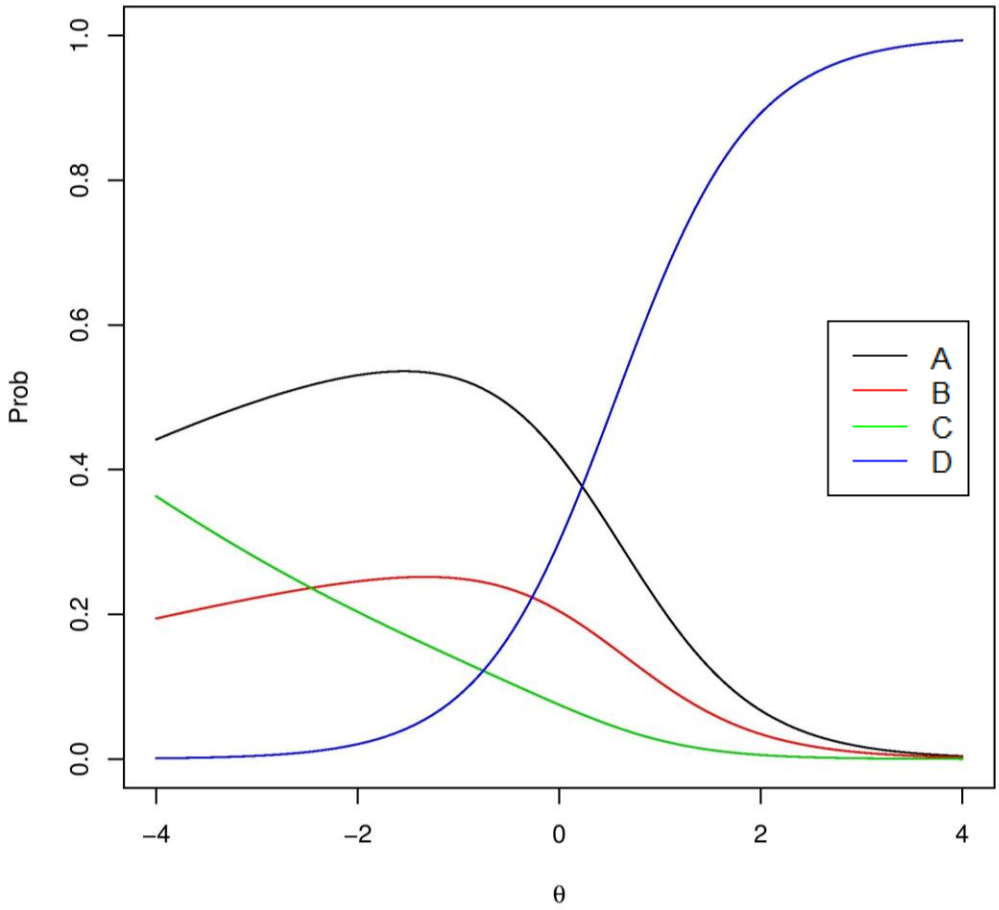


Figure 20. Category characteristic curves for Q30.

intervention has a statistically significant impact on student test scores. She calculates the means of a pre- and post- test for a random sample and finds there is a difference between the two means. Does this support the claim?" For this item, people with the highest ability tend to answer the item correctly (that the teacher would also need to determine if the difference was due to sampling error).

However, the other three distractors had similar probabilities of being picked by respondents of lower ability levels. As ability level increased, respondents were less likely to choose the distractor C (any intervention will lead to a statistically significant difference). Respondents had a higher probability of answering A, that if there is a difference in means would indicate the intervention had an impact on test scores, which is a common misconception among students who do not understand the concept of hypothesis testing. The probability of respondents choosing B, that the teacher would need to run the tests again to verify any difference, is low (.2 for respondents of lower ability), but lingers for candidates with average ability levels. This case illustrates that just because an item may be highly reliable and discriminating when it is scored right/wrong, it may not be as informative for teasing out differences in ability level by distractor.

The last example illustrates the item curves for Q12 and Q13, which both have identical mapped misconceptions (Figure 21). Respondents with the lowest ability level were more likely to choose the distractor associated with the equiprobability bias (D in both cases). Respondents with slightly higher ability levels chose the conjunctive bias (C), and few people chose the distractor corresponding to the base rate fallacy (A). The category characteristic curves look similar across these items, although respondents had a higher probability of choosing C for Q12, which shows that respondents are sensitive to an item's context. Overall, these analyses show that the StatCI's distractor data can be mapped to learner ability and can also provide clues about the trajectory of learners in statistics.

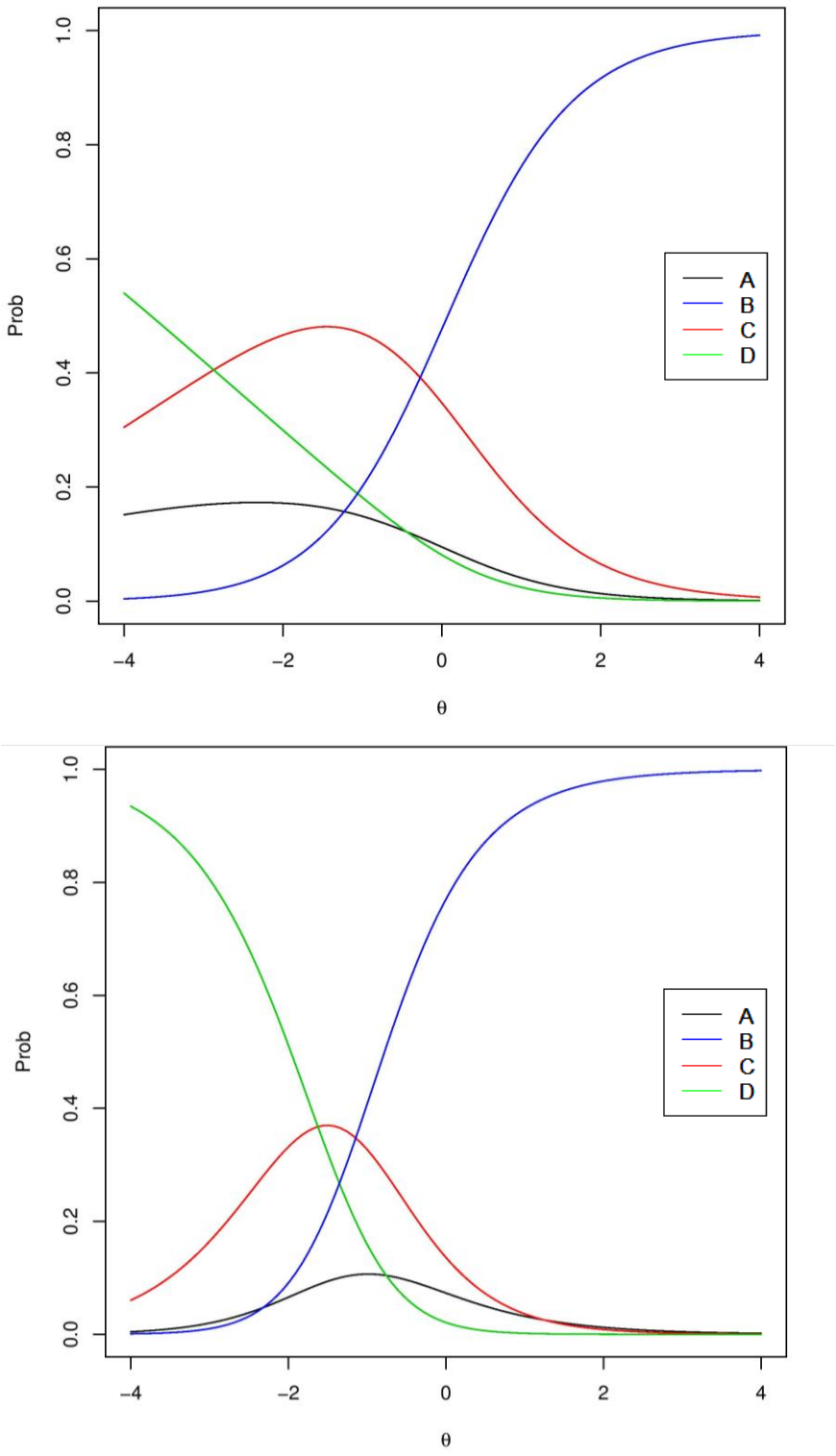


Figure 21. Category characteristic curves for Q12 and Q13.

V. DISCUSSION

5.1 Summary

Although there have been previous attempts at designing an assessment of statistical reasoning, none have been constructed using an ECD framework to guide construct representation and the design process, including explication of an explicit measurement model. Only one such instrument has demonstrated reasonable reliability of measurement. Thus, there is a need for new, high-quality assessments of conceptual understanding of statistics. To create such an assessment for the domain of introductory statistics, the researcher leveraged an evidentiary framework based on theories and research regarding learner cognition. The observation element was elaborated via an evidence-centered design assessment template. The initial instrument underwent several revisions informed by the feedback of several professors and graduate students knowledgeable in statistics. To help further refine the wording of the items and choice of distractors, the researcher conducted think-aloud studies with four undergraduate students.

The researcher administered the assessment to 100 participants on Amazon Mechanical Turk, and then investigated the evidence for the validity and reliability of the new assessment. Because the reliability of the first online version was low ($\alpha = .58$) and several items did not perform well, the researcher revised the instrument and administered it to a new sample of 100 participants, using the results to edit the assessment. The screening criteria for the new version required participants to have taken at least two statistics classes instead of just one and included one attention check. This version had a higher reliability ($\alpha = .72$), and overall the items performed much better. Finally, the researcher administered this last version to 750 participants; extensive analyses were conducted on the participant performance data to obtain evidence of multiple claims regarding the validity of the StatCI.

5.1.1 Overall quality of the StatCI. The final assessment had a reliability measure of .71. The data supported the claim for the assessment's overall score as a measure of student understanding of statistics (Claim 1). Most of the items had an appropriate and wide range of item difficulties and discriminations. The structural analyses provided evidence for the claim that the assessment could measure differentiable conceptual understanding in statistics. Many of the subscales aligned with the concepts that the developer originally defined. Diagnostic analyses show that the StatCI can help to evaluate student mastery of particular concepts (Claim 2). Finally, distractor responses indicated that the assessment could diagnose common misconceptions in statistics (Claim 3).

5.1.2 Use of the StatCI. Given that the researcher was able to provide evidence for the three claims made about the StatCI, the instrument would be appropriate for a variety of purposes. First, the StatCI could be used as a formative assessment in undergraduate statistics classrooms to evaluate students' overall statistical understanding, understanding of particular concepts, as well as to diagnose common misconceptions. The results could inform students and teachers about strengths and gaps in learner understanding. Second, instructors could use the StatCI to evaluate the efficacy of learning interventions or compare student conceptual understanding across different instructors or departments. Instructors could also use pre- and post- test measure to assess changes in conceptual understanding. However, caution should be exercised when using a simple difference measure, which may be an impoverished metric to evaluate conceptual change. Professors should also consider student responses to see if instruction resulted in any changes in conceptual understanding or misunderstanding. Third, the items could be used as clicker questions, since distractor responses provide valuable information about student thinking. Despite the assessment's potential formative use in the classroom, professors may not want to use this as a summative measure of understanding without first aligning the instrument to the instruction and curriculum.

5.2 Proposed Methodology for CI Design

This research presents three main steps for a principled approach to concept inventory design. The first step is to apply an Evidence-Centered Design framework to build the assessment. This includes creating a Domain Analysis, a Domain Model, and a Conceptual Assessment Framework. This process not only lays out the foundation for detailing the important concepts, assessment items, and distractors, it also ensures that the developers are clearly specifying the claims they wish to make about student understanding of the domain.

Once a preliminary version of the assessment is created, the second step is to consult with a range of subject matter experts, from experts to novices. The experts should check the items for face validity and answer the assessment themselves. Ideally, the experts should do well on all of the items, and those with the most experience should do the best on the assessment. Next, the developer should conduct student think-aloud studies to help with distractor development, fine-tuning the difficulty level, and identifying construct-irrelevant variance. Problematic items that the developer modifies should be checked again by experts.

The third step is to administer the assessment. The developer should first use a small sample (n = 100) and check the reliability and validity of the items using psychometric analysis. Depending on the resulting measures, the developer should administer the items on a larger sample (n = 700), particularly to examine the structural properties of the assessment. This participant data should be analyzed for the extent to which it provides evidence for the developer's claims about what the inventory should show about student understanding in the domain. Given that there is substantial validity evidence to support these claims, the assessment will then be ready for use by professors in classrooms.

5.3 Challenges of Developing an Assessment of Conceptual Understanding

5.3.1 Unexpected outcomes. Creating an assessment of higher-order thinking is a

challenging process. Initially, several items were designed with two parts: the first part asked "what," and the second part asked why participants chose the first answer. The aim of this design was to avoid false positives, in which participants were choosing the right answer for the wrong reasons. However, these items were difficult to score and generally had poor psychometric properties. Should learners receive credit only if they answered both parts correctly? Should they receive partial credit for answering only one part correctly? Consequently, the researcher revised these two-part items by "flattening" them into one-part items, with the choice and reasoning both in the same distractor (see for example Q1-Q5). However, the initial step of posing these items in two parts helped to determine which pairs were more likely to be chosen together, which served as the basis for the final distractors.

Having experts validate the items was a necessary but not sufficient means of ensuring that respondents would use the skills the items were designed to measure. In the protocol studies, learners sometimes fixated on irrelevant features of the items, demonstrating the moderating effect of context on novice understanding. The choice of any single word in an item could affect participants' interpretations of the question and how they should consequently answer the item. For example, the use of the world "old" to describe an agent in Q12 made some students think that the agent was retired, which would affect the probability of the outcome. A helpful way to investigate the extent to which the items behaved as expected was to map novice thinking onto a binary classification scheme (true/false negative/positive), as shown in Study 1.

Another unexpected outcome during the item-design process was the effect of distractors on an item's difficulty and discrimination measures. For example, in Study 2A, Q30 had average reliability and item properties. The researcher determined from the distractor responses that one of the distractors was confusing to respondents (some high ability respondents were choosing the wrong answer). The item choices were revised and the stem was kept the same. In Study 3, this

item had strong psychometric properties and had the highest item information functioning of all the items on the assessment. This suggests that there is hope for "poorly" functioning items, so long as they are well-grounded in the domain model.

On the other hand, there were a few items that several experts agreed were conceptually sound and worded appropriately, but kept performing poorly in the analyses regardless of any revisions, such as items Q27 and Q13 (from studies 1-2B). Q27 may have been too wordy for participants, causing increased cognitive load and poor performance on the item. Alternatively, since the test-takers were mTurkers, their thresholds for effort may have been set low, so they may not have put effort into overly wordy items. Thus, in developing the item pool for an assessment, it seems that there will be inevitable casualties no matter how meticulously each item is constructed. One strategy that seemed to help in item revisions was to identify items with strong psychometric properties and use these as a model for creating new items. Q13 from Study 3 was modeled after Q12, and had strong psychometric properties on the first iteration.

There were also some unexpected outcomes in the structural analyses. First, some of the concepts were not cleanly separable, particularly within the probability, sampling, and hypothesis testing categories. This may be because there are overlapping concepts among these categories, such as outcomes of repeated events and sample size. Moreover, participants' fluency in logic may have interacted with their ability to answer these items correctly. Second, one category, p-values, was not discernable via the exploratory factor analysis; the items within this category were not highly correlated. However, this factor did emerge for the subsample of participants who scored greater than chance on the overall assessment. This may reveal an idiosyncrasy about the particular concept. Most people do not leverage understanding of p-values in their everyday lives, but encounter it through formal learning. As a result, this knowledge may depreciate over time, even for those students who earned high grades in statistics. P-values is also a concept that students have

difficulty mastering. In contrast, probability is a concept people encounter and leverage all the time, which may be why participants answered more consistently across items and misconceptions. This suggests that statistical knowledge falls in line with both the knowledge-in-pieces and knowledge-as-theory views of learning, depending on the concept.

5.4 Limitations and Future Study

One notable limitation of the study was that the sample was from mTurkers, not students in an undergraduate statistics class. The reliability of the instrument increased from .58 to .72 from the first to the second version; the change in sample requirements and the attention check likely played a part in this change. Although these participants have an incentive to do well on the tasks since their work can be rejected for poor quality of answers, participants are to some extent motivated by payment. As a result, there may be little incentive for participants to try hard on the items. For example, several professors and graduate students reviewed Q27, yet this item had consistently poor psychometric properties. This was likely because the item had more words than the other items and was therefore more cognitively demanding.

Moreover, the mTurk sample was diverse in that they came from all over the United States and had very different experiences learning statistics. For some participants, it had been three years since they had taken a statistics class. Consequently, the way in which participants answered items was less cohesive than it would have been had the sample been students from one classroom. It is also not clear whether conceptual understanding fades with time, although for the third study, the sample that indicated they had had more time pass since their last statistics class performed better on average.

Therefore, to improve the assessment for future use, it should be administered to a sample of undergraduate students, especially ones who had an incentive to do well on the assessment. Re-testing on such a sample would likely result in more cohesive performance statistics. Any

remarkable deviations from the analyses performed for this research could indicate idiosyncrasies resulting from particular learning experiences with a teacher or curriculum. Correlations of scores on the inventory and class grades could be indicative of the extent to which grades are reflective of conceptual mastery. For those statistics classes that do emphasize conceptual understanding, scores on the inventory could serve as additional criterion-related validity evidence.

Results from such an administration would support previous analyses regarding which areas still needed improvement. For example, if Q27 had poor item statistics from a second administration, it would need to be edited so that it was less wordy. Other items that had poor item statistics should be modified in accordance with psychometric indices. A poor discrimination index might be fixed with better wording of a distractor. An item flagged for DIF should be edited to have an alternate context.

Several additional analyses could provide additional evidence interpretive uses of the inventory. First, a protocol study could be run on those items flagged for DIF to investigate the influence of the item's context on participant responses. Second, it would be useful to provide support for the Q-matrix using expert inter-rater reliabilities or student cognitive interviews. Finally, the number of items per category could be modified so that more important concepts would have more items. Subject matter experts could weigh the relative importance of each concept, which could be used to find the percentage of the test that should be devoted to each concept.

In addition, the product of this research was not only to design an assessment, but also a domain model and design pattern template. Experts in statistics could adapt and add to these documents, which could then serve as the basis for new items. Developers could use these to generate more items with varied contexts and compare psychometric results of these new items to the ones on this inventory.

Finally, a productive avenue for future study would be to use distractor performance data to create learning trajectories in different conceptual areas in statistics. The nominal response model showed that certain misconceptions were associated with different ability measures, and some of these patterns held across items (such as for Q12 and Q13). It would be interesting to see if these patterns held for different samples, and if these results could be cross-validated with other items. The literature on statistical learning trajectories is just nascent; more work in this area could be a great benefit to curriculum and assessment developers, especially given the increasing importance of statistics in education.

5.5 Conclusion

Now more than ever before, statistics plays an increasingly ubiquitous role in daily life. The advent of sophisticated technologies allows us collect, store, and analyze data in order to verify theories and model complex phenomena; statistics serves as a tool to make usable information out of the flood of data that surrounds us. And with this increasing importance of statistics in the everyday, so too is the importance of learning and leveraging the core concepts of the domain.

To this end, the goal of this research was to create a resource for instructors seeking a diagnostic instrument for evaluating learner progress, classroom misconceptions, and educational interventions related to the teaching and learning of introductory statistics. Previous attempts at creating assessments of statistical conceptual understanding were limited in two ways: (1) they were not developed using a principled design approach and (2) the reliability and validity evidence properties of these assessments were lacking. This research details the process to design an inventory: from the creation of a domain model and design pattern template grounded in research literature, to the development of initial items, administering the assessment through student protocol studies and psychometric analyses, and multiple series of refinements and iterations to strengthen the

validity argument for the use of this instrument. Given the overall evidence supporting the three claims the developer sought to validate, the StatCI would have value for any instructors seeking to diagnose student understanding of statistics. The result of this research was a diagnostic assessment of statistical conceptual understanding, a design pattern template for the basis of other instruments, as well as a methodology for applying evidence-centered design to developing assessments of conceptual understanding, with applications to the field of statistics, learning theory, and assessment design.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. AAT 3212015)

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*(2), 167-207.

Arieli-Attali, M. (2013, Oct, 20-25). Formative Assessment with Cognition in Mind: The Cognitively Based Assessment of, for and as Learning (CBALTM) Research Initiative at Educational Testing Service. Paper presented at the *International Association for Educational Assessment*, Tel Aviv. http://www.iaea.info/documents/paper_5b92751b.pdf

Bao, L., & Redish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *Physics Education Research Section of American Journal of Physics, 69* (7), 45-53.

Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement, 8*(2-3), 70-91.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368.

Best, J. B. (1982). Misconceptions about psychology among students who perform highly. *Psychological Reports, 51*(1), 239-244.

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and

preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27*(2), 151-169.

Bock, R.D. (1972). Estimating item parameters and latent proficiency when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30*(2), 213-225.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press.

Brewer, J.K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics,* 10, 252-268.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Callaert, H. (2004). In search of the specificity and the identifiability of stochastic thinking and reasoning. In M. A. Mariotti (Ed.), *Proceedings of the third conference of the European society for research in mathematics education*. Pisa: Pisa University Press.

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295-323). Springer Netherlands.

Chi, M. T. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences, 14*(2), 161-199.

College Board. (2010). *Statistics Course Description Effective Fall 2010*. Retrieved from

http://media.collegeboard.com/digitalServices/pdf/ap/ap-statistics-course-description.pdf

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics.* Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Conley, D. T., Drummond, K. V., de Gonzalez, A., Rooseboom, J., & Stout, O. (2011). *Reaching the goal: The applicability and importance of the Common Core State Standards to college and career readiness*. Eugene, OR: Educational Policy Improvement Center.

Crocker, L. & Algina, J. (2006). *Introduction to classical and modern test theory*. *N*ew York: Wadsworth Publishing Co.

DeBarger, A., DiBello, L., Minstrell, J., Feng, M., Stout, W., Pellegrino, J., Haertel, G., Harris, C., & Ructinger, L. (2011). *Evaluating the diagnostic validity of a Facet-based formative assessment system*. Paper presented at the Society for Research on Educational Effectiveness. Doi: http://files.eric.ed.gov/fulltext/ED540295.pdf

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman and P. Pufall (Eds.), *Constructivism in the Computer Age.* Hillsdale, NJ: Lawrence Erlbaum, 49-70.

diSessa, A.A., Elby, A., & Hammer, D. (2002). J's epistemological stance and strategies. In G. Sinatra and Pintrich (Eds.), *Intentional conceptual change* (238-290). Mahwah, NJ: Lawrence Erlbaum Associates.

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*(6), 843-900.

Falk, R., & Greenbaum, C. W. (1995). Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology, 5*(1), 75-98.

Fife, J. H. (2013). *Automated Scoring of Mathematics Tasks in the Common Core Era: Enhancements to M-rater in Support of CBAL™ Mathematics and the Common Core*

Assessments. ETS Research Report Series. Retrieved from http://www.ets.org/Media/Research/pdf/RR-13-26.pdf

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) report*. Alexandría: American Statistical Association.

Garfield, J. (1998). The statistical reasoning assessment: Development and validation of a research tool. In the *Proceedings of the 5th International Conference on Teaching Statistics.*

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1-2), 99-125.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education, 10*(2), 456-467.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*(1), 83–92.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213-224.

Graf, E. A., Harris, K., Marquez, E., Fife, J. H., & Redman, M. (2010). Highlights from the Cognitively Based Assessment of, for, and as Learning (CBAL) project in mathematics. *ETS Research Spotlight, 3*, 19-30.

Gray, G. L., Costanzo, F., Evans, D., Cornwell, P., Self, B., & Lane, J. L. (2005). The dynamics concept inventory assessment test: A progress report and some results. In *Proceedings of the 2005 American Society for Engineering Education Annual Conference & Exposition* (electronic). Portland, OR.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53, 1043-1055.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158.

Huang, C.W. (2003). *Psychometric analyses based on evidence-centered design and cognitive science of learning to explore students' problem-solving in physics*. Doctoral dissertation, University of Maryland, College Park.

Huff, C. (2014). *"Who are These People?": Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents* (Doctoral dissertation, Department of Government, Harvard University).

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice.* *C*ambridge, MA: MIT Press.

Ioannides, C., & Vosniadou, S. (2002). Exploring the changing meanings of force: From coherence to fragmentation. *Cognitive Science Quarterly, 2*(1), 5-61.

Ipeirotis, P. (2009). Turker Demographics vs Internet Demographics. Retrieved from http://www.behind-the-enemy-lines.com/2009/03/turker-demographics-vs-internet.html

Ipeirotis, P. (2010). Demographics of Mechanical Turk. C*eDER-10–01 working paper*, New York University.

Jöreskog, K. G. (1999). How large can a standardized coefficient be? Unpublished Technical Report. *Retrieved from: http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf.*

Jorion, N., Gane, B., James, K., Schroeder, L., DiBello, L., & Pellegrino, J. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education , 104*(4), 454-496.

Jorion, N., Gane, B. D., James, K., Schroeder, L., Pellegrino, J., & DiBello L. (2014, May). Quantitative analyses of student performance on concept inventories. In J. W. Pellegrino (Chair*), Evaluating and Improving Concept Inventories as Assessment Resources in STEM Teaching and Learning*. Symposium conducted at the meeting of the American Educational Research Association, Philadelphia, PA.

Kahneman, D., Slovic, P. And Tversky, A. (1982). *Judgment under uncertainty: Heuristics and Biases.* Cambridge, UK: Cambridge University Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M., & M. (2013). Validity and fairness in the testing of individuals. *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*. Emerald Group Publishing, Bingley, 17-53.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. In Invited paper in *Proceedings of PME-NA XXXIII* (the North American Chapter of the International Group for the Psychology of Mathematics Education).

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

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Liu, H. J. (1998). *A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States*. Doctoral dissertation, University of Minnesota, Minneapolis.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84.

Masters, J. (2013, April). *Assessing Knowledge and Misconception related to Area Measurement: Validity Evidence*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Masters, J., & Famularo, L. (2015, April). *The Challenges of Measuring Misconceptions in Middle Grades Statistics*. Poster presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In J. Pellegrino, L. Jones, & K. Mitchell (Eds.) *Grading the nation's report card: Research from the evaluation of NAEP*. Washington DC: National Academy Press.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. *Designing for science: Implications from everyday, classroom, and professional settings*, 415-443.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R., & Riconscente, M. (2005). *Evidence-centered Assessment Design: Layers, Structures, and Terminology* (PADI Technical Report 9).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational

assessments. *Measurement: Interdisciplinary Research and Perspectives,* 1, 3-67.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International statistical review, 65*(2), 123-137.

Morgan, B., Baggett, W., & Rus, V. (2014). Error Analysis as a Validation of Learning Progressions. In *Proceedings of the 7th International Conference on Educational Data Mining.*

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward Modern Psychometrics: Application of Item Response Theory Models in Personality Research. In *Handbook of Research Methods in Personality Psychology*, ed. R. W. Robins, New York: Guilford Press.

Morsanyi, K., Primi, C., Chiesi, F., & Handley, S. (2009). The effects and side-effects of statistics education: Psychology students'(mis-) conceptions of probability. *Contemporary Educational Psychology*, *34*(3), 210-220.

Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition & Instruction, 18*, 209-237.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education, 8*(2), 195-209.

Oosterhof, A. (1994). *Classroom applications of educational measurement* (2nd ed.). New York: Macmillan Publishing Company.

Özdemir, G., & Clark, D. B. (2007). An overview of conceptual change theories. *Eurasia Journal of Mathematics, Science & Technology Education, 3*(4), 351-361.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The*

science and design of educational assessment. Washington, DC: National Academies Press.

Pellegrino, J. W., DiBello, L. V., Jorion, N., James, K., & Schroeder, L. (2013). Components of a comprehensive approach to validity. In J. W. Pellegrino (Chair), *Evaluating the validity of concept inventories as aids for STEM teaching and learning*. Symposium conducted at the 2013 Annual Meeting of the American Education Research Association, San Francisco, CA.

Pellegrino, J. W., DiBello, L. V., Miller, R., Streveler, R., Jorion, N., James, K., Schroeder, L., & Stout, W. (2013). An analytical framework for investigating concept inventories. In J. Pellegrino (Chair), *The Conceptual Underpinnings of Concept Inventories. S*ymposium conducted at the meeting of the American Educational Research Association, San Francisco, CA.

Riconscente, M. M., Mislevy, R. J., & Hamel, L. (2005). *An introduction to PADI task templates*. (PADI Technical Report 3). Menlo Park, CA: SRI International and University of Maryland. Retrieved May 1, 2006, from http://padi. Sri. Com/downloads/TR3_Templates.pdf.

Roschelle, J. (1994). Collaborative Inquiry: Reflections on Dewey and Learning Technology. *Computing Teacher*, *21*(8), 6-8.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010, April). Who are the crowdworkers?: shifting demographics in mechanical turk. In CHI'10 extended abstracts on *Human factors in computing systems* (pp. 2863-2872). ACM.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in science Teaching, 35*(3), 265-296.

Shaughnessy, J. M. (2007). Research on statistics learning. *Second handbook of research on mathematics teaching and learning*, 957-1009.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Steif, P. S., & Hansen, M. A. (2007). New practices for administering and analyzing the results of concept inventories. *Journal of Engineering Education, 96*(3), 205-212.

Stout, William. (2013). *Big Ideas and Enduring Understandings in Statistics*. Unpublished document. Department of Statistics, University of Illinois at Urbana-Champaign.

Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests. *Journal of Statistics Education, 17*(2), 3.

Southerland, S. A., Abrams, E., Cummins, C. L., & Anzelmo, J. (2001). Understanding students' explanations of biological phenomena: Conceptual frameworks or p-prims? *Science Education, 85*(4), 328-348

Sundre, D. L., & Thelk, A. D. (2010). Advancing Assessment of Quantitative and Scientific Reasoning. *Numeracy, 3*(2), 2.

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in Web-based, networked public goods experiments. *PLoS One, 6*(3), e16836.

Tempelaar, D. (2004). Statistical reasoning assessment: An Analysis of the SRA instrument. Proceedings of the ARTIST Roundtable Conference, Lawrence University. [Online: http://www.rossmanchance.com/artist/Proctoc.html]

Thissen-Roe, A., Hunt, E., & Minstrell, J. (2004). The DIAGNOSER project: Combining assessment and learning. *Behavior Research Methods, Instruments, & Computers, 36*(2), 234-240.

TurkPrime. (2015, March). *The New New Demographics on Mechanical Turk: Is there Still a Gender Gap?* Retrieved from http://blog.turkprime.com/2015/03/the-new-new-demographics-on-mechanical.html

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1986). Judgment under uncertainty: Heuristics and biases. In Arkes, Hal R. (Ed), Hammond, Kenneth R. (Ed), et al. *Judgment and decision making: An interdisciplinary reader*. (pp. 38-55). Cambridge, England UK: Cambridge University Press.

Vallecillos, A., & Batanero, C. (1996). Conditional probability and the level of significance in the tests of hypotheses. In *PME CONFERENCE* (Vol. 4, pp. 4-371). The Program Committee of the 18th PME Conference.

Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science*, 18, 123-183.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design: Association for Supervision & Curriculum Development.* Alexandria, VA.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawerence Erlbaum Associates.

APPENDICES

Appendix A: William Stout's Big Ideas and Enduring Understandings

Skills defined, Including Coding Rules

1. A. **Understanding elementary probability.** Understanding probability rules, conditional probability and independence, probability modeling of real world. Reasoning probabilistically to solve elementary probability problems.
   B. **Properties and uses of the small set of discrete and continuous probability distributions important to statistics** (normal, exponential, chi square, t; geometric, binomial, likely Poisson in an engineering stats. Course, possibly F if ANOVA included in course). This includes knowledge of the finite-sample-size sampling distributions of commonly used indices, but not their large sample distribution behavior, which falls under Skill 3 below. Example of #1B: when sampling from a normal popn., then (Xbar - mu)/(s/sq rt(n)) is t distributed with n-1 d.f.: a fact that may be used to solve a prob. Problem even if n small. (not part of example above) Knowledge of what types of data tend to come from a particular distribution population included (like heights being normal)

   Coding: If the problem states the index to use (and hence no deciding which index to use required) and then requires one to know its distribution, then the skills assigned should NOT include #5, and should include #1B (or #3 if a large sample distribution result)
   Coding: If sampling distribution of an index result used plus use of that result to help decided which hypothesis testing procedure to use, include both 1B and 2. Ditto for use for deciding on CI procedure, using both 1B and 6 then.

2. **Understanding hypothesis testing**, including how to set one up so as to satisfy the practitioner's real world needs. Knowing when each of the standard hypothesis tests is appropriate to use, based on the probability modeling assumptions made and sampling procedure used (such as the observations being independent, the sample size being large, and the population being normal). Interpreting hypothesis testing results, including considerations of level of significance, p value, and power.

   Coding: Mere mention of the t distribution does not mean that the question involves hypothesis testing. For #2 to be correct, one must require specific knowledge of hypothesis testing.
   Coding: Do not include #1B (or #3) unless specific sampling distribution results needed. But do include #1B or 3 if a specific sampling distribution theory result needed.

3. **Understanding large sample theory results**, including understanding central limit theorem and law of large numbers as they impact the large sample behavior of the sample average and of a few other important statistics. Understand how sample size relates to large sample theoretic results and their statistical applications (getting beyond the abused rule that the CLT applies if and only if the sample size n> 20). Understand that the sample mean has standard error of sigma/sq rt(n) and the resulting implication that the prob. Is small for an observed sample mean being far away (in sigma/√n units) from the population mean. Note that Skill 3

often applies to items that require either Skill 2 or 6 as well. (when a large sample distribution result used plus specifics about the hyp testing or CI used)
Coding: Specific information about CLT or LLN needed means #3 is included

4. **Understanding univariate data graphing** either of random samples (either finite population or experimental replication based) or of other data sets not describable as procured by a random mechanism (such as a so-called sample of convenience like polling the readers of a magazine by inviting e-mail responses or the heights of the 10 tallest buildings in NYC) . Interpreting graphs of data, including understanding sample data shape (relative to the shape of the population). Reading and interpreting of graphs of various types, such as histogram, cumulative distribution function, box and whiskers.

   Coding: bivariate data graphing is coded as #7.
   Coding: If knowledge that shape of graph should be like that of a particular distribution is used, then include 1B (or 3 if several sample means graphed and sample size is moderate, say > 8)
   Coding: If knowledge is required that graph (of sample) shape varies from population shape considerably for small sample size and not much for large sample size, then Skill 8 is required (in addition to #4 if reading graph skill required as well)

5. **Understanding important indices** (called statistics) and understanding deeply the descriptive statistical roles they each can play including understanding how robust (low sensitivity to influential outliers) they are. Knowledge of how influential outliers are for various indices.

   Coding: Do not include #5 unless specific information about an index is required. If mention of index is part of problem context but no special knowledge of index's properties required, do not include #5.
   Coding: If a result about the distribution of an index is required, include #3 or #1B (whichever is appropriate).

6. **Understanding confidence intervals,** including knowing the probability modeling assumptions and sampling procedures, including sample size concerns, that make each of the standard CI procedures appropriate. Interpreting CI results including interpreting what an x% CI actually means.

7. **Carrying out elementary regression** (but not memorizing the formulas for estimating the line or correlation, etc.) and correlation-based analyses, including interpreting the results. Robustness of and influential outlier considerations for bivariate regression data inference, including robustness considerations for the slope estimator and the correlation estimator (r). Graphing of bivariate data and interpretation of bivariate scatter plot data.

8. **Understanding sampling**. Sampling requirements to produce data acceptable for effective statistical analysis ("good" data). That is, good data should be obtained via sampling from a finite population via some randomization mechanism (e.g., obtaining three "pick 3" balls in a lottery drawing) or a sample obtained via independent replications of some physical experiment (e.g., throwing a pair of dice). Specifics about finite population random sampling methods (stratified, multistage, etc.). Sample shape = population shape + error (with size of error decreasing with sample size): From # 3: understanding sample data shape (relative to the shape of the population. The larger the sample size the more accurate the inference is belongs in #8, unless a computation about how accurate is required, such as using (#3) SD($\bar{X}$) =sigma/√n and/or using the central limit theorem. If specific knowledge about type of inference (hypothesis testing or CIs) required as well, include 2 or 6 as well.

   Coding: #1 not appropriate unless specific probability knowledge required. Vague probabilistic reasoning about sample differing from population is definitely #8.
   Coding: shapes that various samples typically have (like heights being normal) belongs as #1, not #8, even though result is about sampling. This is more probability distribution knowledge (like constant failure waiting times are exponential) than details of sampling (like when stratified sampling more appropriate than simple random sampling).

Suggested guiding principles for skills selection:

The actual skills chosen must work in three ways:

a. They need to link closely to item performance so they can be informative for the formative assessment, while being sufficiently few in number so as to work well psychometrically by providing good statistical inference of examinees' skills profiles, this latter forcing a coarseness of granularity.
b. Second, and important but illusive, they must tie in well with the Big ideas/Enduring Understandings (BI/EU). For example, understanding percentile seems a poor skill in that it is too fine grained for the SCI inventory, while a skill based on having a deep understanding of hypothesis testing does tie in well.
c. Skills must work well for classroom formative assessment purposes (cannot be too few skills nor too coarse grained: must tie in well with short term learning objectives of the instructor). Note that one goal for skills is that they must be more fine-grained than BI/EUs.

A new guiding principle for item quality evaluation: Effective items, in addition to measuring the specified skills well, should also function as pathways to the BI/EUs. What can this statement mean operationally? For example, good items in this regard will require some BI/EU-based deep thinking to avoid being drawn to attractive distractors. That is, one is more likely to choose the item's correct option if, and perhaps only if, one has that illusive deeper
understanding that links cognitively to the appropriate BI/EUs , even if indirectly. Other requirements for what constitutes a good item from the BI/EU perspective can be stated of course.

Appendix B: Design Pattern and Misunderstanding Bank to aid in generating questions

<table>
<tr><td colspan="2">Rationale</td><td colspan="2">Conceptual understanding is difficult to assess, and students often maintain erroneous beliefs even after formal instruction. Professors tend to stress procedural knowledge and assume their students have already mastered the necessary conceptual understandings. In order to facilitate conceptual change, professors need a tool to measure students' prior beliefs and understandings.</td></tr>
<tr><td colspan="2">Overview</td><td colspan="2">This Design Pattern employs an ECD framework to item development. In addition, a misconception bank has been embedded in the focal knowledge section, designating misconceptions or confusions; normative thinking is sometimes mentioned within parenthesis. Numbers in the brackets are references from the literature review. The Q-matrix is based on William Stout's Big Ideas and Enduring Understandings in statistics (2012); this document was also adapted for FKs.</td></tr>
<tr><td colspan="2">Construct Labels for Q-matrix</td><td colspan="2">FK1a- Measures of central tendency
FK1b- Measures of dispersion
FK2- Correlation
FK3a- Probability theory
FK3b- Events in Probability
FK4a- Sampling
FK4b- Sample size and probability
FK4c- Data gathering
FK5- Large sample theory
FK6a- P-values
FK6b- Hypothesis testing</td></tr>
<tr><td rowspan="2">Focal Knowledge, Skills, and Abilities
Descriptions
Misunderstand-ings</td><td rowspan="2">1- Univariate Data</td><td>FK1a</td><td>Ability to summarize data using estimates of central tendency, i.e., mean, median, mode; knowing how outliers impact each measure.
M1a.1: An average is the same thing as normal/mode/ median (this disregards variability).
M1a.2: The mean is always the most appropriate measure of central tendency.
M1a.3a: Mode is most affected by skew.
M1a.3b: Median is most affected by skew.</td></tr>
<tr><td>FK1b</td><td>Ability to summarize data using measures of dispersion, i.e., variation and standard deviation.
M1b.1: Variation refers to the bumpiness of a distribution (rather than the spread of a distribution).
M1b.2: Variation refers to how much the data values differ from each other (rather than deviation from central tendency).
M1b.3: Variance is not related to sample size.</td></tr>
</table>

APPENDIX B (continued)

	2- Correlational	FK2	***Ability to interpret studies using correlations and recognize limitations; Knowing when it is justified to make a claim about causation from results of a statistical analysis.*** M2.1: Correlated values are linked by a math function such as proportion. M2.2: If there is a negative correlation, there is no correlation. M2.3: If XY and YZ are correlated, then XZ must be correlated. M2.4: Correlation implies causation.
	3- Probability	FK3a	***Ability to apply probability theory to statistical problems accurately.*** M3a.1: Representative misconception— The likelihood of a sample is based on how closely it resembles the population. M3a.2: Equiprobability bias—Viewing several outcomes of an experiment as equally likely. M3a.3: Availability heuristic—Basing judgments of an outcome on personal experience, where the strength of association becomes basis for probability. M3a4: Time axis —Knowing the result of an event will affect a previous outcome. M3a.5: Base rate fallacy –Ignoring base rates and rely on information about personality to determine outcome. M3a.6: Outcome orientation—Judging a probability as a yes or no decision rather than a series of events. M3a.7: Conjunctive fallacy – Assuming that a joint probability of two events is more likely than one of the events. M3a.8: Gambler's fallacy—Believing that chance is a self-correcting process.
		FK3b	***Ability to interpret the relationship between events. In particular, determine if relationship is dependent or independent. Ability to interpret the outcomes of multiple events.*** M3b.1: P(A|B) is the same as P(B|A). M3b.2: P(A|B) is the same as P(A & B). M3b.3: A and B is more likely than A (Conjunction fallacy). M3b.4: Conditionality is causality. M3b.5: Confusing relative frequency and probability.
	4- Sampling	FK4a	***Ability to identify how "good" samples are created. Ability to differentiate between sample drawn from a population and sampling. Recognize properties of a normal distribution.*** M4a.1: Confusing the original sample with the result of the sampling process. M4a.2: As sample size increases, the closer the distribution will look like the normal distribution. Random sampling is a self-correcting process. M4a.3: A normal distribution is the shape of an inverted U (failing to show that the tails extend without touching the x-axis). M4a.4: A normal distribution will necessarily result from a sample size of 30 / will result from empirical data.

APPENDIX B (continued)

		FK4b	***Ability to recognize the nature of the relationship between sample size and probability.*** M4b.1: Gambler's fallacy (non-independence of unrelated events). M4b.2: Representative misconception – The likelihood of a sample is based on how closely it resembles the population.
		FK4c	***Ability to identify qualities of "good" data gathering. Requires sampling with some randomized mechanism with independent replication*** M4c.1: To be a representative sample, the sample must represent a large portion of the population. M4c.2: Law of small numbers—Any sample drawn from the population will be highly representative of the population.
	5- Large Sample	FK5	***Ability to state the implications of large sample theory on probability and recognize importance in practice of statistical research.*** M5.1: Overlooking sample size when regarding sampling error and predictive accuracy of a statistical test. M5.2: Statistical significance always has practical importance (may not be the case for large sample size or when multiple tests are conducted for an experiment without adjustment for multiple comparisons). M5.3: As sample size increases, data has a tendency to regress towards the mean – and have the same parameters as the population. M5.4: A (large enough) sample can represent the characteristic of the population (without hypothesis testing—extrapolates the law of large numbers to small samples)—aka Representative Heuristic ("People who rely on the representative heuristic tend to estimate the likelihood of events by neglecting the sample size or by placing undue confidence in the reliability of small samples")
	6- Statistical Significance	FK6a	***Understanding the practical significance of p-values and alpha. Understanding that differences in p-values can be explained by sample size, standard deviation, study design, and chance.*** M6a.1: A p-value is deterministic—the null hypothesis is true or false. Statistical tests are probabilistic proofs (similar to the mathematical proof by contradiction based on modus tollens). M6a.2: The p-value is the probability of the null (or alternative) hypothesis. M6a.3: The p-value is the probability of obtaining the same (or more extreme) data. M6a.4: A p-value is the probability that the statistic is correct if the null hypothesis is correct. M6a.5: A p-value is the probability of making an error when rejecting the null hypothesis. M6a.6: A p-value is strength of treatment. // Outcomes with lower p-values have a stronger treatment effect that those with high p-values. M6a.7: A p-value is the probability that the event happened by chance (incomplete because it does not specify the nature of the conditional).

APPENDIX B (continued)

			M6a.8: Mixing up the meaning of significance level with significance, p-value, critical region, and confidence interval. M6a.9: Alpha is the probability that the null hypotheses is true. M6a.10: Alpha is the probability of rejecting the null hypothesis. M6a.11: Alpha is the probability that the null hypothesis is rejected when the null hypothesis is wrong. M6a.12: Alpha is the probability of the null hypothesis assuming its rejection. M6a.13: Alpha is the probability that the null hypothesis is rejected even though the null hypothesis is correct; Beta is that the null hypothesis is not rejected although the null hypothesis is incorrect. [92] M6a.14: A significance level of .05 means that on average, 5 times out of every 100 times we reject the null hypothesis, we will be wrong. // When we reject at the 95% level, we are saying that the chances are 95 out of 100 that it is false.
		FK6b	***Understanding the practical significance of hypothesis testing.*** M6b.1: *Confusing null and alternative hypothesis.* M6b.2: A hypothesis can refer to both the population and sample. M6b.3: Hypothesis testing is not related to the decision making process. M6b.4: A statistic will be very significant if there is a high p-value. M6b.5: Statistical significance always has practical importance (may not be the case for large sample size or when multiple tests are conducted for an experiment without adjustment for multiple comparisons). Conversely, a finding of no difference does not have practical importance (may be due to insufficient power). M6b.6: There will necessarily be a meaningful (interpretive) difference between one study that has a significant p-value and another that does not. // *Confusing practical and meaningful significance.*
Additional Knowledge, Skills, and Abilities *Things that have to be dealt with because of the context of the items.*		AK1	Understand and interpret tables and graphs
		AK2	Apply variables and functions (for modeling)
		AK3	Leverage literacy skills
		AK3	Apply number sense
		AK4	Apply familiarity with real-world situation
		AK5	Leverage logic
Potential Observations		PO1	Accuracy of answer selected by student
		PO2	Distractor selected by student
		PO3	Pattern of answers selected by student
		PO4	Degree of certainty student shows in selecting answer (protocol studies)
		PO5	Rationale student provides for answer (protocol studies)

APPENDIX B (continued)

Potential Work Products	PW1	Answers to selected or open response (***solve***). Examples: **-** ***Reason*** probabilistically to solve statistical problems **-** ***Interpret*** statistical output **-** ***Predict*** what will happen and ***justify*** the corresponding selection with reasoning
	PW2	Recordings/transcripts of students working through problems
	PW3	Computer records of students' interaction with online assessment (time it took for students to complete answer, whether students switched answers, whether students used any supplementary tools)
Characteristic Features *Features that all or almost all questions have.*	CF1	Task asks students to evaluate interpretations or predict outcomes by presenting a situation that requires understanding the targeted focal knowledge, skills, or abilities
Variable Features *Types of context*	VF1	Which type of distribution to focus on
	VF2	Which types of information representations to interpret, such as text, diagrams, tables
	VF3	What kind of statistical test to use
	VF4	What type of context (superficial features) to interpret
National educational standards	CCSS.MATH.CONTENT.HSS.ID.A.2 – **FK1** Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets. CCSS.MATH.CONTENT.HSS.ID.A.3 – **FK1a** Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers). CCSS.MATH.CONTENT.HSS.ID.C.8 – **FK2** ~~Compute (using technology) and~~ Interpret the correlation coefficient of a linear fit. CCSS.MATH.CONTENT.HSS.ID.C.9 – **FK2b** Distinguish between correlation and causation. CCSS.MATH.CONTENT.HSS.IC.A.1 – **FK4a** Understand statistics as a process for making inferences about population parameters based on a random sample from that population. CCSS.MATH.CONTENT.HSS.IC.B.3 – **FK4**, **FK5** Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each. CCSS.MATH.CONTENT.HSS.IC.B.5 – **FK6** Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant. CCSS.MATH.CONTENT.HSS.CP.A.2 – **FK3b** Understand that two events *A* and *B* are independent if the probability of *A* and *B* occurring together is the product of their probabilities, and use this characterization to determine if they are independent.	

	CCSS.MATH.CONTENT.HSS.CP.A.5 – **FK3a** Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations. F*or example, compare the chance of having lung cancer if you are a smoker with the chance of being a smoker if you have lung cancer.* CCSS.MATH.CONTENT.HSS.MD.B.7 – **FK3a** (+) Analyze decisions and strategies using probability concepts (e.g., product testing, medical testing, pulling a hockey goalie at the end of a game).

Appendix C: Connections among Statistics Common Core Benchmarks,

Focal KSAs, and Potential Observations

Benchmark	Focal KSA
Summarize, represent, and interpret data on a single count or measurement variable	Interpreting univariate data
CCSS.MATH.CONTENT.HSS.ID.A.1 Represent data with plots on the real number line (dot plots, histograms, and box plots).	[Procedural]
CCSS.MATH.CONTENT.HSS.ID.A.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.	FK1a- Ability to summarize data using estimates of central tendency, i.e., mean, median, mode. FK1b - Ability to summarize data using measures of dispersion, i.e., variation and standard deviation.
CCSS.MATH.CONTENT.HSS.ID.A.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).	FK1a - Knowing how outliers impact measures of central tendency.
CCSS.MATH.CONTENT.HSS.ID.A.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.	[Procedural]
Summarize, represent, and interpret data on two categorical and quantitative variables	Interpreting bivariate data
CCSS.MATH.CONTENT.HSS.ID.B.5 Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.	[Procedural]
CCSS.MATH.CONTENT.HSS.ID.B.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	

APPENDIX C (continued)

CCSS.MATH.CONTENT.HSS.ID.B.6.A Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.	[Procedural]
CCSS.MATH.CONTENT.HSS.ID.B.6.B Informally assess the fit of a function by plotting and analyzing residuals.	[Procedural]
CCSS.MATH.CONTENT.HSS.ID.B.6.C Fit a linear function for a scatter plot that suggests a linear association.	[Procedural]
Interpret linear models	
CCSS.MATH.CONTENT.HSS.ID.C.7 Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.	[Procedural]
CCSS.MATH.CONTENT.HSS.ID.C.8 Compute (using technology) and interpret the correlation coefficient of a linear fit.	FK2- Ability to interpret studies using correlations and recognize limitations.
CCSS.MATH.CONTENT.HSS.ID.C.9 Distinguish between correlation and causation.	FK2- Knowing when it is justified to make a claim about causation from results of a statistical analysis.
Understand and evaluate random processes underlying statistical experiments	Understanding how sampling applies to statistical inferences
CCSS.MATH.CONTENT.HSS.IC.A.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.	FK4a- Ability to identify how "good" samples are created. Ability to differentiate between sample drawn from a population and sampling. Recognize properties of a normal distribution.
CCSS.MATH.CONTENT.HSS.IC.A.2 Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.	[Out of scope]
Make inferences and justify conclusions from sample surveys, experiments, and observational studies	

APPENDIX C (continued)

CCSS.MATH.CONTENT.HSS.IC.B.3 Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.	FK4a- Ability to identify how "good" samples are created. Ability to differentiate between sample drawn from a population and sampling. Recognize properties of a normal distribution. FK4b- Ability to recognize the nature of the relationship between sample size and probability. FK4c- Ability to identify qualities of "good" data gathering. Requires sampling with some randomized mechanism with independent replication. FK5- Ability to state the implications of large sample theory on probability and recognize importance in practice of statistical research.
CCSS.MATH.CONTENT.HSS.IC.B.4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.	[Procedural]
CCSS.MATH.CONTENT.HSS.IC.B.5 Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.	FK6- Understanding that differences in p-values can be explained by sample size, standard deviation, study design, and chance; Understanding the practical significance of p-values and alpha; Understanding the practical significance of hypothesis testing.
CCSS.MATH.CONTENT.HSS.IC.B.6 Evaluate reports based on data.	[Out of scope]
Understand independence and conditional probability and use them to interpret data	
CCSS.MATH.CONTENT.HSS.CP.A.1 Describe events as subsets of a sample space (the set of outcomes) using characteristics (or categories) of the outcomes, or as unions, intersections, or complements of other events ("or," "and," "not").	[Procedural]

CCSS.MATH.CONTENT.HSS.CP.A.2 Understand that two events *A* and *B* are independent if the probability of *A* and *B* occurring together is the product of their probabilities, and use this characterization to determine if they are independent.	FK3b- Ability to interpret the relationship between events. In particular, determine if relationship is dependent or independent. Ability to interpret the outcomes of multiple events.
CCSS.MATH.CONTENT.HSS.CP.A.3 Understand the conditional probability of *A* given *B* as *P*(*A* and *B*)/*P*(*B*), and interpret independence of *A* and *B* as saying that the conditional probability of *A* given *B* is the same as the probability of *A*, and the conditional probability of *B* given *A* is the same as the probability of *B*.	[Procedural]
CCSS.MATH.CONTENT.HSS.CP.A.4 Construct and interpret two-way frequency tables of data when two categories are associated with each object being classified. Use the two-way table as a sample space to decide if events are independent and to approximate conditional probabilities.	[Procedural]
CCSS.MATH.CONTENT.HSS.CP.A.5 Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations. F*or example, compare the chance of having lung cancer if you are a smoker with the chance of being a smoker if you have lung cancer.*	FK3a- Ability to apply probability theory to statistical problems accurately.
Use the rules of probability to compute probabilities of compound events.	
CCSS.MATH.CONTENT.HSS.CP.B.6 Find the conditional probability of *A* given *B* as the fraction of *B*'s outcomes that also belong to *A*, and interpret the answer in terms of the model.	[Procedural]
CCSS.MATH.CONTENT.HSS.CP.B.7 Apply the Addition Rule, P(A or B) = P(A) + P(B) - P(A and B), and interpret the answer in terms of the model.	[Procedural]
CCSS.MATH.CONTENT.HSS.CP.B.8 (+) Apply the general Multiplication Rule in a uniform probability model, P(A and B) = P(A)P(B|A) = P(B)P(A|B), and interpret the answer in terms of the model.	[Procedural]
CCSS.MATH.CONTENT.HSS.CP.B.9 (+) Use permutations and combinations to compute probabilities of compound events and solve problems.	[Procedural]

Calculate expected values and use them to solve problems CCSS.MATH.CONTENT.HSS.MD.A.1 (+) Define a random variable for a quantity of interest by assigning a numerical value to each event in a sample space; graph the corresponding probability distribution using the same graphical displays as for data distributions. CCSS.MATH.CONTENT.HSS.MD.A.2 (+) Calculate the expected value of a random variable; interpret it as the mean of the probability distribution. CCSS.MATH.CONTENT.HSS.MD.A.3 (+) Develop a probability distribution for a random variable defined for a sample space in which theoretical probabilities can be calculated; find the expected value. CCSS.MATH.CONTENT.HSS.MD.A.4 (+) Develop a probability distribution for a random variable defined for a sample space in which probabilities are assigned empirically; find the expected value.	[Procedural] [Procedural] [Procedural] [Procedural]
Use probability to evaluate outcomes of decisions CCSS.MATH.CONTENT.HSS.MD.B.5 (+) Weigh the possible outcomes of a decision by assigning probabilities to payoff values and finding expected values. CCSS.MATH.CONTENT.HSS.MD.B.5.A Find the expected payoff for a game of chance. CCSS.MATH.CONTENT.HSS.MD.B.5.B Evaluate and compare strategies on the basis of expected values. F*or example, compare a high-deductible versus a low-deductible automobile insurance policy using various, but reasonable, chances of having a minor or a major accident.* CCSS.MATH.CONTENT.HSS.MD.B.6 (+) Use probabilities to make fair decisions (e.g., drawing by lots, using a random number generator). CCSS.MATH.CONTENT.HSS.MD.B.7 (+) Analyze decisions and strategies using probability concepts (e.g., product testing, medical testing, pulling a hockey goalie at the end of a game).	[Procedural] [Procedural] [Procedural] FK3a- Ability to apply probability theory to statistical problems accurately.

Appendix D: Task Template

Component	Value
Summary	Assessment for first year undergraduate course on statistics. The assessment has 30-multiple choice items, with four choices per item. Distractors are linked to common student misconceptions.
Student Model Summary	(1) One overall measure of proficiency (2) Subscores of performance on 5 specified concepts, including diagnostic profile indicating probability of concept mastery (3) Reports of possible misconceptions
Student Models	(1) Overall proficiency model – a univariate, continuous model (2) Subscores—a multivariate model (3) Misconceptions—polytomous response models and cross tabulations
Measurement Model Summary	Overall: univariate and dichotomous Subscores: multivariate and dichotomous Misconceptions: polytomous and univariate
Evaluation Procedures Summary	Automated scoring with answer key
Work Product Summary	Selected answer responses in Qualtrics.
Template-level Task Model Variables	Content area. Specific construct being assessed. Complexity of content. Familiarity of students with content.
Task Model Variable Settings	Probabilities, distributions, sample size, p-values, accompanying graphics
Materials and Presentation Requirements	Computer and internet access to take assessment.
Activities Summary	Students fill out test online and review results
Tools for Examinee	Computer and internet access
Design Patterns	Design Pattern for Statistical Conceptual Understanding and Misconceptions

Appendix E: Original Q-Matrix

Item	FK1a- Measures of central tendency	FK1b- Measures of dispersion	FK2- Correlation	FK3a- Probability theory	FK3b- Events in Probability	FK4a- Sampling	FK4b- Sample size	FK4c- Data gathering	FK5- Large sample theory	FK6a- P-values	FK6b- Hypothesis testing
Q01	1	0	0	0	0	0	0	0	0	0	0
Q02	1	0	0	0	0	0	0	0	0	0	0
Q03	1	0	0	0	0	0	0	0	0	0	0
Q04	0	1	0	0	0	0	0	0	1	0	0
Q05	1	1	0	0	0	0	0	0	0	0	0
Q06	0	0	1	0	0	0	0	0	0	0	0
Q07	0	0	1	0	1	0	0	0	0	0	0
Q08	0	0	1	0	1	0	0	0	0	0	0
Q09	0	0	1	0	1	0	0	0	0	0	0
Q10	0	0	0	1	0	1	0	0	0	0	0
Q11	0	0	0	1	0	1	0	0	0	0	0
Q12	0	0	0	1	1	0	0	0	0	0	0
Q13	0	0	0	1	1	0	0	0	0	0	0
Q14	0	0	0	0	1	0	0	0	1	0	0
Q15	0	0	0	0	1	0	0	0	0	0	0
Q16	0	0	0	0	1	0	0	0	0	0	0
Q17	0	1	0	0	0	1	0	0	1	0	0
Q18	0	0	0	0	0	1	0	0	1	0	0
Q19	0	0	0	0	0	1	1	1	0	0	0
Q20	0	0	0	0	1	1	0	1	1	0	0
Q21	0	0	1	0	1	0	1	0	0	0	0
Q22	0	0	0	0	0	1	1	1	0	0	0
Q23	0	0	0	0	1	0	1	1	0	0	0
Q24	0	0	0	0	0	0	0	0	0	1	0
Q25	0	0	0	0	0	0	0	0	0	1	1
Q26	0	0	0	0	0	0	0	0	1	1	0
Q27	0	0	0	0	1	1	0	0	0	1	1
Q28	0	0	0	0	0	0	0	0	0	1	1
Q29	1	0	0	0	0	0	0	0	0	0	1
Q30	1	0	0	0	0	1	0	0	0	0	1

Appendix F: Statistics Concept Inventory

Assessment of Statistical Conceptual Understanding

Pre1 What year were you born?
❍ [select year]

Q1 A group of critics ranked 100 restaurants on a scale of 1 to 5, 1 being a poor restaurant and 5 being an excellent restaurant. What would be the most appropriate measure of central tendency for this set of rankings?

FK1a: Ability to summarize data using estimates of central tendency.

❍ Mean because it provides the most precise value.
The mean is the most precise measure of central tendency.

❍ Mean because it is the best measure for summarizing numeric data.
The mean is always the most appropriate measure of central tendency.

❍ Median because it is the best measure for summarizing ordinal data.
Correct.

❍ Mode because it is the best measure for summarizing discrete data.
Mode is the most appropriate measure of central tendency for ordinal data.

Item statistics	**Q1**
Category	Univariate
Mean	0.27
AlphaIfDeleted	0.71
Discrimination	0.09

	Percent Chosen
A	16%
B	48%
C	**27%**
D	9%

Q2 Which is the most appropriate summary statistic for a set of zip codes?

FK1a: Ability to summarize data using estimates of central tendency.

- Mean because it is the best measure for summarizing numeric data.
  The mean is always the most appropriate measure of central tendency.
- Median because the measure indicates the middle value.
  Median is the most appropriate measure of central tendency for discrete data.
- Mode because the measure is the most appropriate for discrete data.
  Correct.
- Median because it is the best measure for summarizing ordinal data.
  Confusing nominal and ordinal data.

Item statistics	Q2
Mean	0.59
AlphaIfDeleted	0.70
Discrimination	0.30

	Percent Chosen
A	12%
B	12%
C	**59%**
D	17%

Q3 An outlier will have the greatest impact on which statistical measure?

FK1a: Knowing how outliers impact each measure of central tendency.

- Mode because it often times not accounting for all the data.
  Mode is most affected by skew.
- Median because it is dragged in the direction of the extreme scores.
  Median is most affected by skew.
- Mode because the measure is the most appropriate for discrete data.
  Mode is most affected by skew.
- Mean because it averages all the observations.
  Correct.

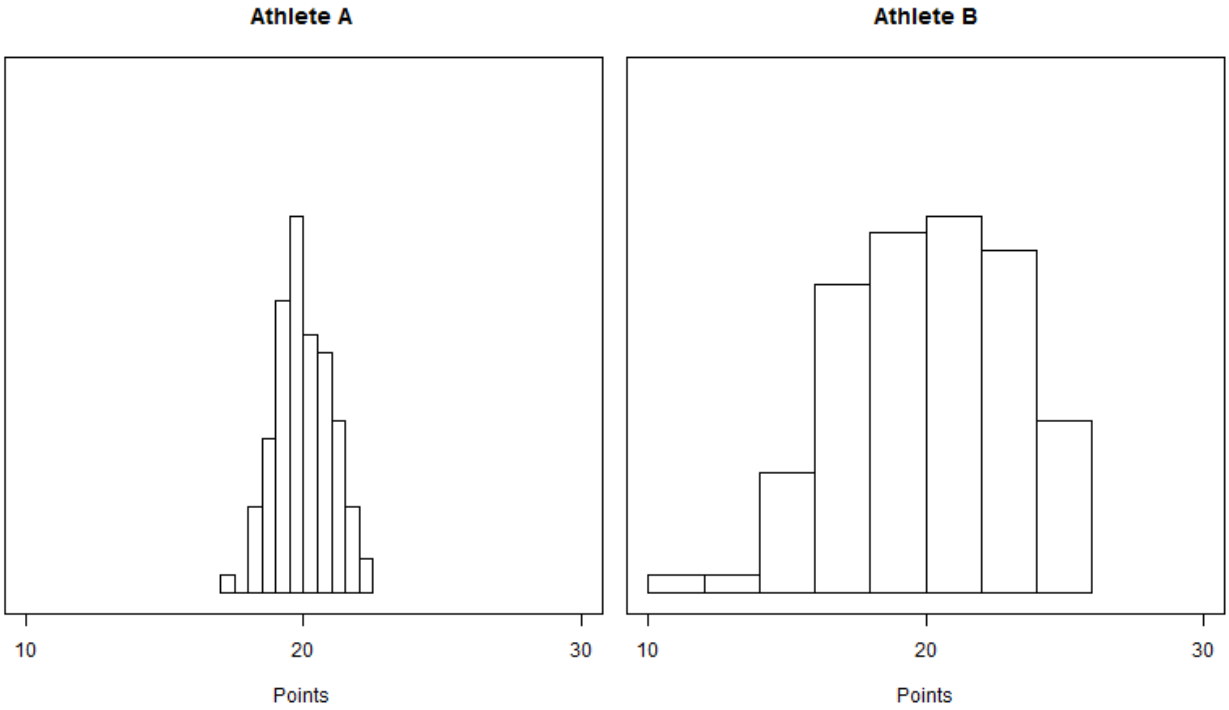
Item statistics	Q3
Mean	0.56
AlphaIfDeleted	0.70
Discrimination	0.30

	Percent Chosen
A	9%
B	28%
C	8%
D	**56%**

Q4 The two graphs below show the number of points two different athletes scored per game over a season. Which of the two athletes scored points more consistently?

FK1b: Ability to summarize data using measures of dispersion, i.e., variation and standard deviation.

- ❍ Athlete A because the scores are less spread out.
  Correct.
- ❍ Athlete A because the scores fall along a normal distribution.
  Smaller SD = more normally distributed.
- ❍ Athlete B because the athlete scored more points on average across games.
  Smaller SDs = greater the N.
- ❍ Athlete B because the scores are more spread out.
  Smaller SD = more spread.



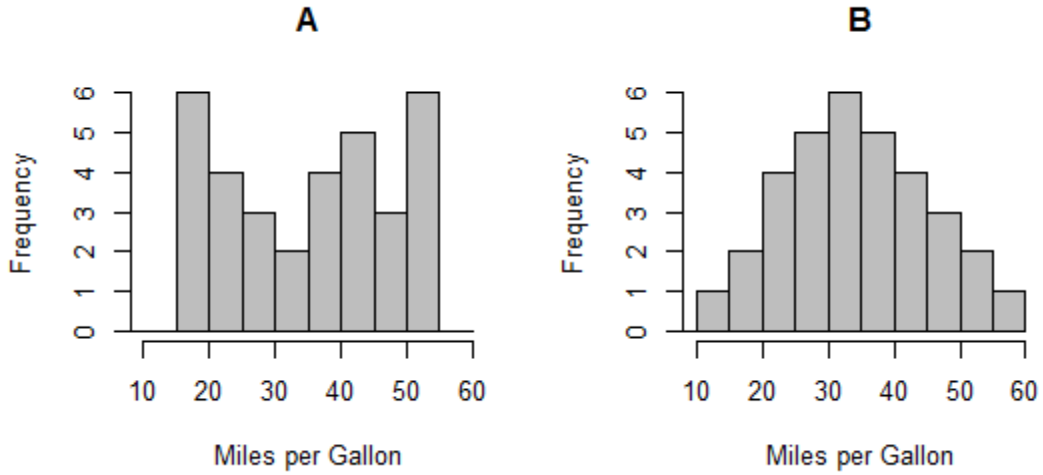
Item statistics	Q4
Mean	0.36
AlphaIfDeleted	0.69
Discrimination	0.36

Percent Chosen	
A	**36%**
B	14%
C	42%
D	8%

Q5 Both of the distributions below have the same mean of 33.5. Which of these two histograms is the most likely to have the greatest standard deviation?

FK1b: Ability to summarize data using measures of dispersion, i.e., variation and standard deviation.

- ❍ Graph A because the observed values have a greater average distance from the mean.
  Correct.
- ❍ Graph A because it is less normally distributed.
  Smaller SD = more normally distributed.
- ❍ Graph A because the distribution is bumpier.
  Variation refers to the bumpiness of a distribution (rather than the spread of a distribution).
- ❍ Graph B because the Miles per Gallon has a larger range (15 - 55 for A versus 10 - 60 for B).
  Variation refers to how much the data values differ from each other (rather than deviation from central tendency).



Item statistics	Q5
Mean	0.36
AlphaIfDeleted	0.71
Discrimination	0.06

Percent Chosen	
A	**36%**
B	31%
C	28%
D	5%

Q6 According to an article, there is a -0.8 linear correlation between commute time and happiness. What does this tell you about the two variables?

FK2: Ability to interpret studies using correlations. Knowing when it is justified to make a claim about causation from results of a statistical analysis.

- ❍ Shorter commute times lead to lower happiness.
  Correlation implies causation.
- ❍ Shorter commute times lead to higher happiness on average.
  Correlation implies causation.
- ❍ Commute time and happiness are strongly related.
  Correct.
- ❍ Commute time and happiness are not related.
  If there is a negative correlation, there is no correlation.

Item statistics	Q6
Mean	0.30
AlphaIfDeleted	0.72
Discrimination	0.00

	Percent Chosen
A	6%
B	54%
C	**30%**
D	10%

Q7 A study finds that eating organic food is related to living longer. The study also finds that living longer is related to having a lower income. What are the implications of this study?

FK2: Ability to interpret studies using correlations. Knowing when it is justified to make a claim about causation from results of a statistical analysis.

- Eating organic food increases life expectancy.
  Correlation implies causation.
- Eating organic food is negatively correlated with income.
  If XY and YZ are correlated, then XZ must be correlated.
- Having a higher income leads to living longer.
  Correlation implies causation.
- Income is correlated with life expectancy.
  Correct.

Item statistics	Q7
Mean	0.43
AlphaIfDeleted	0.70
Discrimination	0.20

	Percent Chosen
A	28%
B	26%
C	3%
D	**43%**

Q8 A study found that on average, introverted individuals tend to have fewer interactions with others. Another study found that the number of interactions people have with others is related to how many hours they spend on the computer. What can be inferred from these studies?

FK2: Ability to interpret studies using correlations. Knowing when it is justified to make a claim about causation from results of a statistical analysis.

- Being introverted is positively correlated to the amount of time spent on the computer.
  If XY and YZ are correlated, then XZ must be correlated.
- Spending more time on the computer results in spending less time interacting with others.
  Correlation implies causation.
- The amount of time spent on a computer is correlated to an individual's frequency of interactions with others.
  Correct.
- Being introverted leads to interacting with others less frequently.
  Correlation implies causation.

Item statistics	Q8
Mean	0.40
AlphaIfDeleted	0.71
Discrimination	0.14

	Percent Chosen
A	25%
B	20%
C	**40%**
D	15%

Q9 A company wants to investigate the relationship between coffee consumption and productivity. Which of the following provides evidence that the relationship between the two variables is independent? Assume that there would be a linear relationship.

FK2: Ability to interpret studies using correlations. Knowing when it is justified to make a claim about causation from results of a statistical analysis.

- ❍ If they administer a survey to a random sample and find that the correlation coefficient of coffee consumption and productivity is -1.
  If there is a negative correlation, there is no correlation.
- ❍ If they administer a survey to a random sample and find that the correlation coefficient of coffee consumption and productivity is 0.
  Correct.
- ❍ If they provide coffee to their employees and find that there is no impact on overall productivity.
  Correlation implies causation. (Biased sample)
- ❍ If they find a very low correlation between coffee consumption per person and labor productivity for all countries.
  Correlation implies causation.

Item statistics	Q9
Mean	0.49
AlphaIfDeleted	0.70
Discrimination	0.24

	Percent Chosen
A	14%
B	**49%**
C	24%
D	13%

Q10 There is a 70% chance that a particular bird will appear each day at a designated location during the last 10 days of the year. Which of the following is most likely to happen?

FK3a: Ability to apply probability theory to statistical problems accurately.

- ❍ The bird appeared every single day of the 10 days.
  Outcome orientation—Judging a probability as a yes or no decision rather than a series of events.
- ❍ If the bird appeared on every single day of the first 7 days, the bird would not appear on the last 3 of the 10 days.
  Time axis —Knowing the result of an event will affect a previous outcome.
- ❍ There is a greater chance that the bird appears on the first 7 days if the bird did not appear on the last 3 days.
  Gambler's fallacy—Chance is a self-correcting process.
- ❍ The bird appeared on 6 of the 10 days.
  Correct.

Item statistics	Q10
Mean	0.31
AlphaIfDeleted	0.69
Discrimination	0.36

	Percent Chosen
A	14%
B	22%
C	33%
D	**31%**

Q11 There is a 10% chance that an earthquake will hit a particular city each month. Based on this information, which statement is the most probable?

FK3a: Ability to apply probability theory to statistical problems accurately.

- There will not be an earthquake in the next year.
  Outcome orientation—Judging a probability as a yes or no decision rather than a series of events.
- If there is an earthquake in the first month of the year, then there will not be any earthquakes in the following nine months.
  Gambler's fallacy—Chance is a self-correcting process.
- There is a greater chance that an earthquake will hit a city in California than one in the United States.
  Base rate fallacy –Ignoring base rates and rely on information about personality to determine outcome.
- There will be one earthquake in the next year.
  Correct.

Item statistics	Q11
Mean	0.46
AlphaIfDeleted	0.70
Discrimination	0.22

Percent Chosen	
A	16%
B	29%
C	9%
D	**46%**

Q12 Nora is a 36-year-old woman who loves cats. In 2014, her town had 25 cat breeders and 500 post office employees. Which of the following is the most likely? Assume being a cat breeder and being a post office worker are independent of one another.

FK3a: Ability to apply probability theory to statistical problems accurately.

FK3b: Ability to interpret the relationship between events.

- She works as a cat breeder.
  Base rate fallacy –Ignoring base rates and rely on information about personality to determine outcome.
- She works at the post office.
  Correct.
- She is a cat breeder who supports her passion by working at the post office.
  A and B is more likely than A (Conjunction fallacy).
- It is equally likely that she works at the post office or is a cat breeder.
  Equiprobability bias—Viewing several outcomes of an experiment as equally likely.

Item statistics	Q12
Mean	0.47
AlphaIfDeleted	0.69
Discrimination	0.40

	Percent Chosen
A	10%
B	**47%**
C	10%
D	33%

Q13 An obituary in a local paper reported that Cody Smith, a firefighter, died recently. Which of the following is the most probable cause of his death? Note that heart disease accounts for 45% of firefighter deaths, and fire-related accidents account for 20% of firefighter deaths. Assume that heart disease and fire-related accidents are independent of one another.

FK3a: Ability to apply probability theory to statistical problems accurately.

FK3b: Ability to interpret the relationship between events.

- A fire-related accident.
  Base rate fallacy –Ignoring base rates and rely on information about personality to determine outcome.
- A heart attack.
  Correct.
- A heart attack during a fire-related accident.
  A and B is more likely than A (Conjunction fallacy).
- It is equally likely that he died either by a heart attack or a fire-related accident.
  Equiprobability bias—Viewing several outcomes of an experiment as equally likely.

Item statistics	Q13
Mean	0.69
AlphaIfDeleted	0.69
Discrimination	0.36

	Percent Chosen
A	7%
B	**69%**
C	7%
D	17%

Q14 An engineer programmed a software application to have a 50% chance of crashing each time it is opened. The program was installed in two different computer labs, in which each computer ran on its own independent network. In the large computer lab, 60 students tried to use the program. In the small computer lab, 10 students tried to use the program. The total percentage of times the program crashed in each lab was recorded. In which of the two labs was it more likely that the program crashed 80% of the time?

FK3b: Ability to interpret the relationship between events.

FK4c: Ability to identify qualities of "good" data gathering.

- Small computer lab. The greater the sample size, the more likely the average chance of crashing will be 50%.
  Correct.
- Large computer lab. With a larger sample size, there will be more crashes.
  Misunderstanding how sample size relates to probability.
- Either. The probability that the computer will crash in both cases is the same, so they are both equally likely to have crashed 80% of the time.
  Equiprobability bias—Viewing several outcomes of an experiment as equally likely.
- It is impossible to determine without more information.
  Incorrect.

Item statistics	Q14
Mean	0.37
AlphaIfDeleted	0.70
Discrimination	0.29

Percent Chosen	
A	**37%**
B	15%
C	38%
D	9%

Q15 A data analyst finds an association rule for how customers purchase products in a supermarket. The rule states that 30% of customers who buy cheese will also pick up bread on their way out of the grocery store. What inference can be made based on this statement?

FK3b: Ability to interpret the relationship between events.

- 30% of customers buy bread and cheese together.
  P(A|B) is the same as P(A & B).
- 30% of customers who buy bread also buy cheese.
  P(A|B) is the same as P(B|A).
- Less than 30% of all customers buy bread and cheese together.
  Incorrect.
- More than 30% of all customers who buy cheese do not buy bread.
  Correct.

Item statistics	Q15
Mean	0.31
AlphaIfDeleted	0.70
Discrimination	0.32

	Percent Chosen
A	39%
B	19%
C	11%
D	**31%**

Q16 Students in a science class took two tests. 75% of students who passed the first test passed the second test. What inference can be made based on this statement?

FK3b: Ability to interpret the relationship between events.

- 75% of students passed both the first and second test.
  P(A|B) is the same as P(A & B).
- 75% of students who passed the second test also passed the first test.
  P(A|B) is the same as P(B|A).
- More than 75% of students who passed the second test did not pass the first test.
  Incorrect.
- Less than 75% of students who passed the first test did not pass the second test.
  Correct.

Item statistics	Q16
Mean	0.28
AlphaIfDeleted	0.69
Discrimination	0.36

	Percent Chosen
A	41%
B	27%
C	47%
D	**28%**

Q17 A technician selects 10 oranges from an orchard. The mean weight of this sample is 8 ounces. Based on this information, what can we infer?

FK4a: Ability to identify how "good" samples are created. Ability to differentiate between sample drawn from a population and sampling.

- The population mean should also be 8 ounces, since sampling ensures that the sample will be representative of the population.
  Samples are always representative of the population.
- The sample is normally distributed.
  A sample is normally distributed (confusing with sampling distribution).
- The technician would have to weigh all the oranges to find the population mean.
  Correct.
- If the technician adds one other orange to the sample, the new sample mean will be closer to the population mean.
  Law of large numbers—but for one additional. (Could add an outlier).

Item statistics	Q17
Mean	0.28
AlphaIfDeleted	0.70
Discrimination	0.28

	Percent Chosen
A	33%
B	9%
C	**28%**
D	31%

Q18 An analyst is investigating the number of times different words appear in a document. The analyst samples 20 words and counts the number of times they each appear in the document. If the analyst increases the sample to 40 words, which of the following is most likely to happen?
FK4a: Ability to identify how "good" samples are created. Ability to differentiate between sample drawn from a population and sampling.
FK5: Ability to state the implications of large sample theory on probability and recognize importance in practice of statistical research.

- ❍ The distribution of the new sample will look more like a normal distribution.
  Larger samples converge to a normal distribution.
- ❍ The distribution of the new sample will look more like the population distribution.
  Correct.
- ❍ The distribution of the new sample will look the same as the old sample.
  Misunderstanding law of large numbers.
- ❍ The mean of the new sample will be equal to the mean of the old sample.
  Misunderstanding law of large numbers.

Item statistics	Q18
Mean	0.45
AlphaIfDeleted	0.70
Discrimination	0.27

	Percent Chosen
A	30%
B	**45%**
C	12%
D	13%

Q19 A researcher asks 25 randomly sampled employees from a large company how many years they spent in school, where the majority of employees finished either college or graduate school. The sample is currently not normally distributed. The researcher is concerned that this could bias the results. What would you suggest to the researcher?

FK4b: Ability to recognize the nature of the relationship between sample size and probability.

- ❍ Increase the sample size so that the distribution comes closer to a normal distribution.
  Larger samples converge to a normal distribution.
- ❍ Survey all the employees—otherwise, there is no way to ensure the sample will not be biased.
  To be a representative sample, the sample must represent a large portion of the population.
- ❍ The current sample is adequate since populations are not always normally distributed.
  Correct.
- ❍ Discard the sample and construct a new random sample of 25 employees.
  Incorrect.

Item statistics	Q19
Mean	0.22
AlphaIfDeleted	0.70
Discrimination	0.18

	Percent Chosen
A	51%
B	19%
C	**22%**
D	8%

Q20 A researcher says she randomly sampled 30 students from a school with replacement. This school has exactly the same proportion of males and females. In this sample of 30 students, there were 20 males and 10 females. What can we infer about the sample chosen?

FK4b: Ability to recognize the nature of the relationship between sample size and probability.

- If the sampling process had been truly random, the sample would have been representative of the population.
  To be a representative sample, the sample must represent a large portion of the population. / Disregards sampling error.
- If the researcher samples 10 additional students randomly from the school population, there is a greater chance that a female will be chosen.
  Gambler's fallacy.
- The more students the researcher includes in the survey, the more likely that there will be exactly the same number of males and females in the sample.
  Disregards sampling error.
- This difference in sample and population demographics can result from random sampling.
  Correct.

Item statistics	Q20
Mean	0.35
AlphaIfDeleted	0.71
Discrimination	0.10

Percent Chosen	
A	14%
B	19%
C	32%
D	**35%**

Q21 Two different researchers measure the weight of a dog. One researcher records the weight after one reading. The other researcher takes 10 measurements and records the average. How would you expect these two measurements to compare?

FK4b: Ability to recognize the nature of the relationship between sample size and probability.

- The researcher who recorded one reading would have the more accurate measure.
  Incorrect.
- The researcher who recorded the average of 10 readings would have the more accurate measure.
  Correct.
- Both researchers would have equally accurate measures.
  Equiprobability bias.
- It is impossible to predict which researcher had the more accurate measurement without knowing the population parameter.
  Incorrect.

Item statistics	Q21
Mean	0.64
AlphaIfDeleted	0.70
Discrimination	0.22

	Percent Chosen
A	6%
B	**64%**
C	17%
D	12%

Q22 Your employer wants you to determine an adequate sample size for a study. The employer wants the sample to be cost-efficient, yet a sufficient size to make inferences about the population. What would you suggest?

FK4c: Ability to identify qualities of "good" data gathering.

- Any sample size should be adequate.
  Law of small numbers—Any sample drawn from the population will be highly representative of the population.
- The sample size should equal to the population size.
  To be a representative sample, the sample must represent a large portion of the population.
- The sample size should be 30 participants since this is the general rule of thumb for minimum sample sizes.
  A normal distribution will necessarily result from a sample size of 30 / will result from empirical data.
- The sample size should be based on the amount of error the employer is willing to accept.
  Correct.

Item statistics	Q22
Mean	0.64
AlphaIfDeleted	0.69
Discrimination	0.38

Percent Chosen	
A	3%
B	13%
C	20%
D	**64%**

Q23 According to a report, in a certain pond, 70% of the fish are blue and 30% are red. A biologist catches one fish from the pond, notes its color, and then puts it back in the pond. She does this 10 times. The biologist finds that of the fish caught, 50% were blue fish and 50% were red fish. Based on this new information, what can the biologist infer?

FK4b: Ability to recognize the nature of the relationship between sample size and probability.

FK4c: Ability to identify qualities of "good" data gathering.

- The sampling process was likely biased-- otherwise, the biologist would have found 70% blue fish and 30% red fish.
  To be a representative sample, the sample must represent a large portion of the population.
- There is more than a 70% chance that next fish the biologist catches will be blue, since the biologist's sample underrepresented the blue fish.
  Gambler's fallacy.
- The more fish the biologist includes in the sample, the more likely there will be exactly 70% blue fish and 30% red fish.
  Disregards sampling error.
- Differences between the biologist's sample and the report can result from random sampling.
  Correct.

Item statistics	Q23
Mean	0.40
AlphaIfDeleted	0.71
Discrimination	0.16

Percent Chosen	
A	12%
B	9%
C	39%
D	**40%**

Q24 A researcher assumes a significance level of 0.05 in an experiment. What is the most accurate way to interpret this number?

FK6a: Understanding the practical significance of p-values and alpha.

- There is a 5% probability that the alternative hypothesis is true.
  Alpha is the probability that the null hypotheses is true.
- There is a 5% probability the null hypothesis is true, assuming it is rejected.
  Alpha is the probability that the null hypothesis is rejected when the null hypothesis is wrong.
- There is a 5% probability the null hypothesis is true, assuming it is rejected.
  The probability of rejecting the null hypothesis is 1-alpha.
- There is a 5% probability of rejecting the null hypothesis, assuming the null hypothesis is true.
  Correct

Item statistics	Q24
Mean	0.42
AlphaIfDeleted	0.71
Discrimination	0.18

	Percent Chosen
A	24%
B	16%
C	17%
D	**42%**

Q25 How would you interpret a p-value of 0.96 given a 0.05 significance level?

FK6a: Understanding the practical significance of p-values and alpha.

- ❍ Reject the null hypothesis.
  Incorrect decision.
- ❍ Fail to reject the null hypothesis.
  Correct.
- ❍ The result is highly statistically significant.
  A result's significance level is dependent on its p-value.
- ❍ There is a 96% probability the null hypothesis is true.
  A p-value is deterministic—the null hypothesis is true or false.

Item statistics	Q25
Mean	0.23
AlphaIfDeleted	0.70
Discrimination	0.21

	Percent Chosen
A	19%
B	**23%**
C	25%
D	34%

Q26 What does a p-value of 0.01 signify?

FK6a: Understanding the practical significance of p-values and alpha.

- There is a 1% chance of making an error when rejecting the null hypothesis.
  Incorrect.
- The probability of the null hypothesis is 1%.
  The p-value is the probability of the null (or alternative) hypothesis.
- There is a 1% probability the null hypothesis is true, assuming the same or more extreme data.
  A p-value is deterministic—the null hypothesis is true or false.
- There is a 1% probability of obtaining the same or more extreme data, assuming the null hypothesis is true.
  Correct.

Item statistics	Q26
Mean	0.24
AlphaIfDeleted	0.71
Discrimination	0.05

	Percent Chosen
A	28%
B	24%
C	24%
D	**24%**

Q27 An engineer investigates whether an energy drink contains an average of 100 mg of caffeine per serving. The null hypothesis is that the average caffeine content is 100 mg per serving. The alternative hypothesis is that the average caffeine content is more than 100 mg per serving. The engineer assumes a significance level of 0.05. After analyzing a sample, she finds that the p-value is 0.17 and the mean is 111 mg per serving. How can she interpret these results?

FK6a: Understanding the practical significance of p-values and alpha.

- There is a 17% chance that the average caffeine content per serving is not 100 mg and the alternative hypothesis is correct.
  A p-value is the probability that the event happened by chance (incomplete because it does not specify the nature of the conditional).
- There is a 17% chance of getting 111 mg or more per serving if the mean caffeine content is 100 mg per serving.
  Correct
- There is a 17% chance that the null hypothesis is true.
  A p-value is deterministic—the null hypothesis is true or false.
- There is a 17% probability of rejecting the null hypothesis.
  The p-value is the probability of the null (or alternative) hypothesis.

Item statistics	Q27
Mean	0.34
AlphaIfDeleted	0.72
Discrimination	0.00

	Percent Chosen
A	31%
B	**34%**
C	20%
D	15%

Q28 In an experiment, a psychologist sets the significance level to 0.05 and finds a p-value of 0.03. Which of the following statements is the most accurate interpretation?

FK6a: Understanding the practical significance of p-values and alpha.

- The null hypothesis is false.
  A p-value is deterministic—the null hypothesis is true or false.
- There is a 3% probability that the event happened by chance.
  A p-value is the probability that the event happened by chance (incomplete because it does not specify the nature of the conditional).
- There is a 3% chance of making an error when rejecting the null hypothesis.
  M6a.5: A p-value is the probability of making an error when rejecting the null hypothesis.
- Given that the null hypothesis is true, there is a 3% probability of getting a statistic at least as extreme as the one observed.
  Correct

Item statistics	Q28
Mean	0.34
AlphaIfDeleted	0.71
Discrimination	0.05

Percent Chosen	
A	15%
B	18%
C	33%
D	**34%**

Q29 A researcher finds that the sample mean of a subgroup is greater than the population mean. Is the difference statistically significant?

FK6b: Understanding the practical significance of hypothesis testing.

- Yes, if the values are different then the result is statistically significant.
  Hypothesis testing is not related to the decision making process.
- Only if the subgroup has 30 or more participants.
  A normal distribution will necessarily result from a sample size of 30 / will result from empirical data.
- Only if she runs the test again and gets the same results.
  Incorrect.
- More information is needed to conclude that the difference is statistically significant.
  Correct.

Item statistics	Q29
Mean	0.60
AlphaIfDeleted	0.69
Discrimination	0.36

	Percent Chosen
A	14%
B	13%
C	13%
D	**60%**

Q30 A teacher wants to support the claim that an intervention has a statistically significant impact on student test scores. She calculates the means of a pre- and post- test for a random sample and finds there is a difference between the two means. Does this support the claim?
FK6b: Understanding the practical significance of hypothesis testing.

- ❍ Yes, if there is a difference in means, the intervention had an impact on test scores.
  Hypothesis testing is not related to the decision making process.
- ❍ No, she would need to run the test again and get the same results to verify any difference.
  Incorrect.
- ❍ Yes, any intervention will lead to a statistically significant difference.
  Hypothesis testing is not related to the decision making process.
- ❍ No, she would also need to determine if the difference in means of a pre- and post- test administration was due to sampling error.
  Correct.

Item statistics	Q30
Mean	0.35
AlphaIfDeleted	0.68
Discrimination	0.48

Percent Chosen	
A	38%
B	19%
C	8%
D	**35%**

Q31 [ATTENTION CHECK] Linda is 31 years old, single, outspoken, and very bright. She is training to become a nurse. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which of the following is true?

- ❍ Linda is a bank teller.
- ❍ Linda is a bank teller and is active in the feminist movement.
- ❍ Linda is a training to become a nurse.

Appendix G: Demographics Questionnaire

What is your gender?
[ ] Female
[ ] Male

What is the highest level of education that you have completed?
[ ] Some college
[ ] Bachelor's degree
[ ] Graduate Degree/professional
[ ] Other. Please indicate: _____

Please specify your ethnicity.
[ ] American Indian or Alaskan Native
[ ] Asian/Pacific Islander
[ ] Black
[ ] Hispanic/ Latino
[ ] White/ Caucasian
[ ] Would rather not say
[ ] Other. Please indicate: _____

How many statistics classes have you taken?
[ ] 0 *[NOTE: This is to ensure that participants did not accidently fill out the eligibility requirements incorrectly.]*
[ ] 1
[ ] 2
[ ] 3 or more

How many years has it been since you have taken a statistics course?
[ ] Less than 1
[ ] 1
[ ] 2
[ ] 3 or more

Thank you for participating in our assessment!

Appendix H: Demographics of Participants (on using MTurk)

Amazon Mechanical Turk (MTurk) is an online labor market that connects researchers and businesses with human workers to complete pre-specified tasks such as image tagging, transcribing, filling out surveys, and writing. Ever since the site was launched publicly in November, 2005, scientists have been using MTurk for social and behavioral science. However, this resource has not been widely used for educational research. As of May 2016, there was only one published study using MTurk to validate concept inventory items.[4] Morgan, Baggett, and Rus (2014) used MTurk to gather open-response data to 22 questions on Newton's 3rd Law for the Force Concept Inventory from 30 workers. They mapped Turker responses to learning progressions; these learning progressions were refined according to these student responses. Overall, this seemed like an effective means to gather a large amount of open-response data to inform the development of their framework.

There are several reasons for using MTurk participants (also called "Turkers") over college students. First, MTurk provides access to a larger sample size. A larger sample size is needed to make more robust inferences about student misunderstandings. Second, the assessment could be administered anytime—not just a particular date during the semester. Third, some college professors who were contacted for the study were worried that an assessment in conceptual understanding might hurt students' self-efficacy in the subject.

Reliability of responses from Turkers is one concern for this study. Several studies indicate that Turkers versus those recruited in online and offline settings are comparable in terms of demographics and study results (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, &

[4] Philip Sadler has used Amazon Turk to validate his Astronomy Concept Inventory; however, his research is still under review as of May 2016.

Ipeirotis, 2010; Suri & Watts, 2011). In regards to demographics, one survey found that 75% of workers were U.S. residents, 40% were younger than 30, 70% had Bachelor's degrees, 33% were students, and 60% had an annual income over $25,000 a year (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). Ipeirotis (2009) found that the demographics of Turkers were similar to U.S. internet population users, except that the former tended to be younger, have smaller families, be mostly female, and have lower incomes. Berinsky, Huber, and Lenz (2012) found that U.S. Turkers are often more representative of the U.S. population than convenience samples, but less representative than internet-based panels. Comparative differences between a benchmark survey and an MTurk survey showed that demographical differences decreased when the data was subset to younger participants (Huff & Tingley, 2014). This study was geared to political scientists, so it is difficult to extrapolate the degree to which the Turker undergraduate population will be comparable to the U.S. undergraduate population.

It should be noted that these demographics have been somewhat variable since MTurk's launch in 2005. For example, the survey in February, 2010 (n = 984) found that 60% of workers were female (Ross et al., 2010). A 2015 meta-analysis examined 75 surveys since 2013 (n = 15,324) found that 47% of Turkers were female, statistically smaller than the number of male workers (TurkPrime, 2015).

Regarding comparability of response patterns, researchers found that Turkers were just as attentive and able to think through highly cognitive problems as a community sample (Goodman, Cryder, & Cheema, 2012). Based on their findings, Goodman et al. "highly recommend" MTurk for behavior decision-making research, with the caveat that questionnaires include an attention check to ensure participants are paying close attention. Moreover, participants are rewarded for accuracy on assignments and surveys—those that provide consistent results can obtain an "elite" status that offers better pay for responses to surveys.

It is possible that participants may try to screen for eligibility in order to be paid for the survey or try to answer the survey as quickly as possible without regard to the complexity of the items. As a result, reliability checks were an important part of data processing. The researcher embedded two attention checks in the assessment and rejected the work of those participants who answered these basic items incorrectly or inconsistently.

Appendix I: Diagnostic Scoring Report

Diagnostic Scoring Report Student Name: Dexter

Answers

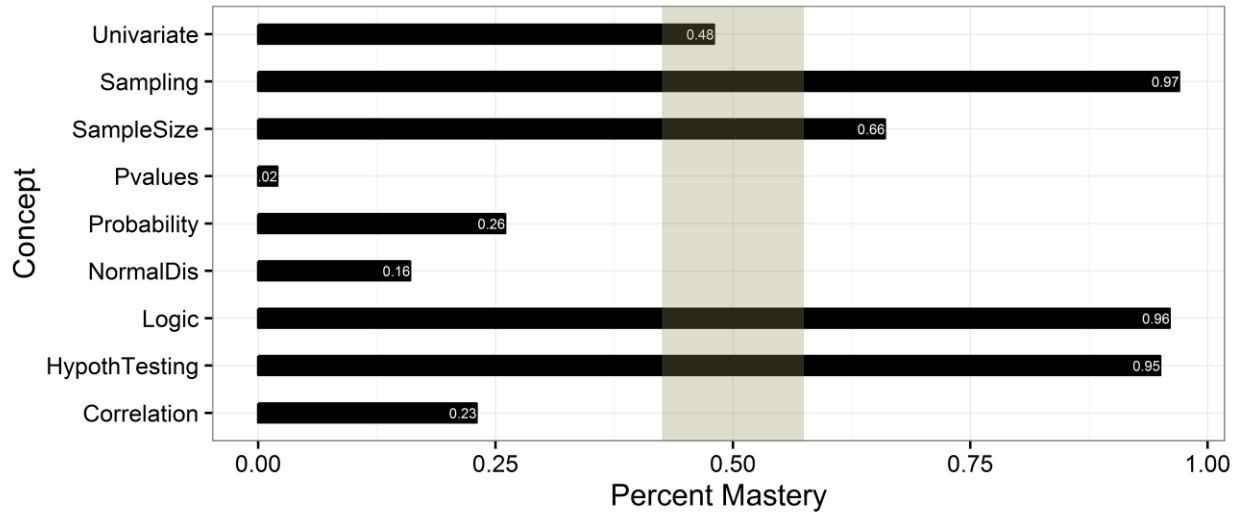
Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Chosen	B	✓	✓	B	B	✓	C	✓	✓	C	C	✓	✓	✓	A
Correct	C	C	D	A	A	C	D	C	B	D	D	B	B	A	D

Question	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Chosen	✓	✓	A	A	C	✓	✓	C	✓	C	B	A	B	✓	✓
Correct	D	C	B	C	D	B	D	D	D	B	D	B	D	D	D

Score

Dexter correctly answered 15 out of 30 questions.

Mastery



Mastered: Sampling, Sample Size, Logic, Hypothesis Testing
Unsure of Mastery: Univariate
Not Mastered: P-values, Probability, Normal Distributions, Correlation

Misconceptions

Chosen twice:
The smaller the standard deviation, the more normally distributed the sample.
Larger samples converge to a normal distribution.
Disregards sampling error.
A p-value is the probability that the event happened by chance (incomplete because it does not specify the nature of the conditional).

Chosen once:
The mean is always the most appropriate measure of central tendency.
Correlation implies causation.
Gambler's fallacy—Chance is a self-correcting process.
Base rate fallacy –Ignoring base rates and rely on information about personality to determine outcome.
P(A|B) is the same as P(A & B).
A result's significance level is dependent on its p-value.

Appendix J: IRB Approval for Research

UNIVERSITY OF ILLINOIS
AT CHICAGO

Office for the Protection of Research Subjects (OPRS)
Office of the Vice Chancellor for Research (MC 672)
203 Administrative Office Building
1737 West Polk Street
Chicago, Illinois 60612-7227

Approval Notice
Continuing Review

September 29, 2015

Natalie Jorion, BA,MA
Learning Sciences Research Institute
857 N Marshfield Ave, Unit 1
Chicago, IL 60612
Phone: (949) 246-4304

RE: **Protocol # 2014-0917**
"Assessing Student Misconceptions in Statistics"

Dear Ms. Jorion:

Your Continuing Review application was reviewed and approved by the Expedited review process on September 24, 2015. You may now continue your research.

Please note the following information about your approved research protocol:

Protocol Approval Period: October 13, 2015 - October 12, 2016
Approved Subject Enrollment #: 306 (0 subjects enrolled)
Additional Determinations for Research Involving Minors: These determinations have not been made for this study since it has not been approved for enrollment of minors.
Performance Site: UIC
Sponsor: None
Research Protocol:

a) Research Proposal for IRB Review; Version 1; 09/26/2014

Recruitment Materials:

a) Eligibility Form; Version 1; 10/10/2014
b) Recruitment Document; Version 1; 10/10/2014
c) Subject Information Sheet (telephone screener); Version 1; 06/18/2015
d) Recruitment Material for Verbal Protocol; Version 2; 09/18/2015

APPENDIX J (continued)

Informed Consents:

a) Debriefing Form; Version 1; 09/26/2014
b) Consent Form; Version 3; 10/22/2014
c) Debriefing Form; Version 1; 06/01/2015
d) Consent Form for Protocol Study; Version 2; 06/18/2015

e) A waiver of documentation of informed consent has been granted under 45 CFR 46.117 and an alteration of consent has been granted under 45 CFR 46.116(d) for this online research; minimal risk; electronic consent will be obtained.

Your research continues to meet the criteria for expedited review as defined in 45 CFR 46.110(b) under the following specific category:

(7) Research on individual or group characteristics or behavior (including but not limited to research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

Please note the Review History of this submission:

Receipt Date	Submission Type	Review Process	Review Date	Review Action
09/16/2015	Continuing Review	Expedited	09/24/2015	Approved

Please remember to:

→ Use your **research protocol number** (2014-0917) on any documents or correspondence with the IRB concerning your research protocol.

→ Review and comply with all requirements on the OPRS website under:
"UIC Investigator Responsibilities, Protection of Human Research Subjects"
(http://tigger.uic.edu/depts/ovcr/research/protocolreview/irb/policies/0924.pdf)

Please note that the UIC IRB has the prerogative and authority to ask further questions, seek additional information, require further modifications, or monitor the conduct of your research and the consent process.

Please be aware that if the scope of work in the grant/project changes, the protocol must be amended and approved by the UIC IRB before the initiation of the change.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact OPRS at (312) 996-1711 or me at (312) 996-2014. Please send any correspondence about this protocol to OPRS at 203 AOB, M/C 672.

Sincerely,

Sandra Costello
Assistant Director, IRB # 2
Office for the Protection of Research Subjects

Enclosures:

1. **Informed Consent Documents:**
   a) Debriefing Form; Version 1; 09/26/2014
   b) Consent Form; Version 3; 10/22/2014
   c) Debriefing Form; Version 1; 06/01/2015
   d) Consent Form for Protocol Study; Version 2; 06/18/2015
2. **Recruiting Materials:**
   a) Eligibility Form; Version 1; 10/10/2014
   b) Recruitment Document; Version 1; 10/10/2014
   c) Subject Information Sheet (telephone screener); Version 1; 06/18/2015
   d) Recruitment Material for Verbal Protocol; Version 2; 09/18/2015

cc: Susan Goldman, Learning Sciences Research Institute, M/C 285
James Pellegrino (faculty advisor), Learning Sciences Research Institute, M/C 057

APPENDIX J (continued)

UNIVERSITY OF ILLINOIS
AT CHICAGO

Office for the Protection of Research Subjects (OPRS)
Office of the Vice Chancellor for Research (MC 672)
203 Administrative Office Building
1737 West Polk Street
Chicago, Illinois 60612-7227

Approval Notice
Amendment to Research Protocol – Expedited Review
UIC Amendment # 2

September 29, 2015

Natalie Jorion, BA,MA
Learning Sciences Research Institute
857 N Marshfield Ave, Unit 1
Chicago, IL 60612
Phone: (949) 246-4304

RE: **Protocol # 2014-0917**
"Assessing Student Misconceptions in Statistics"

Dear Ms. Jorion:

Members of Institutional Review Board (IRB) #2 have reviewed this amendment to your research under expedited procedures for minor changes to previously approved research allowed by Federal regulations [45 CFR 46.110(b)(2)]. The amendment to your research was determined to be acceptable and may now be implemented.

Please note the following information about your approved amendment:

Amendment Approval Date: September 24, 2015
Amendment:

Summary: UIC Amendment #2, dated September 15, 2015 and received September 16, 2015 is an investigator-initiated amendment about the following:

1) Revising the survey instrument (Statistics Concept Inventory, Version 2, 9/15/2015).

Approved Subject Enrollment #: 306
Performance Site: UIC
Sponsor: None

APPENDIX J (continued)

Please note the Review History of this submission:

Receipt Date	Submission Type	Review Process	Review Date	Review Action
09/16/2015	Amendment	Expedited	09/24/2015	Approved

Please be sure to:

→ Use your research protocol number (2014-0917) on any documents or correspondence with the IRB concerning your research protocol.

→ Review and comply with all requirements on the OPRS website under:
"**UIC Investigator Responsibilities, Protection of Human Research Subjects**"
(http://tigger.uic.edu/depts/ovcr/research/protocolreview/irb/policies/0924.pdf)

Please note that the UIC IRB #2 has the right to ask further questions, seek additional information, or monitor the conduct of your research and the consent process.

Please be aware that if the scope of work in the grant/project changes, the protocol must be amended and approved by the UIC IRB before the initiation of the change.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact the OPRS at (312) 996-1711 or me at (312) 996-2014. Please send any correspondence about this protocol to OPRS at 203 AOB, M/C 672.

Sincerely,

Sandra Costello
Assistant Director, IRB # 2
Office for the Protection of Research Subjects

cc: James Pellegrino (faculty advisor), Learning Sciences Research Institute, M/C 057
Susan Goldman, Learning Sciences Research Institute, M/C 285

Appendix K: Permission to Reprint Copyrighted Material

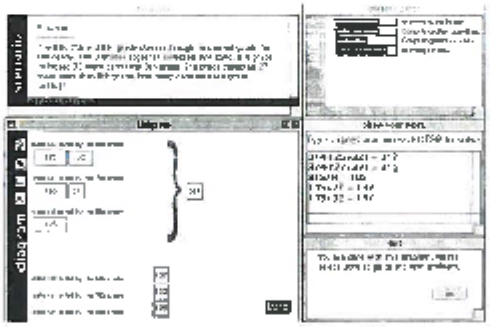
University of Illinois at Chicago
Learning Sciences Research Institute

January 4, 2016.

Kenneth Koedinger
3601 Newell-Simon Hall
Pittsburgh Science of Learning Center
Carnegie Mellon University
Pittsburgh, PA 15213-3891

Dear Dr. Koedinger,

I am writing to request permission to use the following figure from your publication *Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6*, p.12, 2002, in my thesis. This material will appear as originally published. Unless you request otherwise, I will use the conventional style of the Graduate College of the University of Illinois at Chicago as acknowledgment.



A copy of this letter is included for your records. Thank you for your kind consideration of this request.

Sincerely,

Natalie Jorion
857 N Marshfield Ave. Unit 1
Chicago, IL 60622

The above request is approved.

Approved by: Kenneth Koedinger Date: 1/20/16

Stella DeVries
Sr. Copyright Administrator
660 Rosedale Road
Princeton, NJ 08541
Phone: 609-683-2825
Email: sdevries@ets.org

January 21, 2016

Natalie Jorion
University of Illinois, Chicago
857 N Marshfield Ave Unit 1
Chicago, IL 60622

Dear Natalie Jorion:

This Letter Agreement is in response to your request to use copyrighted Educational Testing Service (ETS) material specified in the attached Appendix as part of your dissertation entitled Designing an Evidence-based Assessment of Concept.

ETS is pleased to grant royalty-free, nonexclusive, nontransferable permission to reproduce the materials listed above. The following terms apply to this permission:

1. The material is to be used only for the research purposes described in your request and is not to be distributed, published, or used in any other manner without written permission from ETS.
2. You are permitted to use these materials in print and electronic formats.
3. The following credit line will be printed on the page where you use the material and/or appropriate locations:

   Copyright © 2010 Educational Testing Service. www.ets.org

4. This agreement shall be considered null and void if not signed and returned within 30 days of the date of this letter.

The signing of this license shall constitute agreement of the above terms and shall be considered a binding contract once a countersigned copy is returned to you.

Natalie Jorion
Signature

Natalie Jorion
Print Name

University of Illinois, Chicago
Institution Name

1/22/16
Date

Stella DeVries
Stella DeVries
Copyright Licensing & Permissions Group
Educational Testing Service

1/28/16
Date

Request # ETS159

University of Illinois at Chicago
Learning Sciences Research Institute

March 22, 2016.

Jim Minstrell
The Hunt Lab
Department of Psychology
University of Washington
Box 351525
Seattle, WA 98195-1525

Facet Innovations
1314 NE 43rd St.
Suite 207
Seattle, WA 98105

Dear Dr. Minstrell,

I am writing to request permission to use the following figure from your publication *Evaluating the Diagnostic Validity of a Facet-based Formative Assessment System* (p.7) by A. DeBarger, L. DiBello, J. Minstrell, M. Feng, W. Stout, J. Pellegrino, G. Haertel, C. Harris, & L. Ructinger 2011, in my thesis (see attached page for specific figure). This material will appear as originally published. Unless you request otherwise, I will use the conventional style of the Graduate College of the University of Illinois at Chicago as acknowledgment.

A copy of this letter is included for your records. Thank you for your kind consideration of this request.

Sincerely,

Natalie Jorion
857 N Marshfield Ave. Unit 1
Chicago, IL 60622

The above request is approved.

Approved by: Jim Minstrell Date: 3/22/16

VITA

Natalie Jorion

857 N Marshfield Unit 1
Chicago, Il 60622

Phone: (949)246-4304
E-mail: njorio2@uic.edu

EDUCATION

Ph.D. Candidate in Learning Sciences
University of Illinois at Chicago, Chicago, IL; 2010-2016
Concentration: Measurement, Evaluation, Statistics, and Assessment
GPA: 3.9/4.0

M.A. In Learning Sciences
Northwestern University, Evanston, IL; 2008-2009
GPA: 3.8/4.00

B.A. In French Literature and Creative Writing (Honors), Math Minor
University of California San Diego, La Jolla, CA; 2001-2006
GPA: 3.6/4.0

SKILLS

Languages

- Speaking, reading, and writing fluency in French
- Reading fluency in Spanish

Technical Proficiency

- SPSS, R (programming language), SAS, STATA
- Amos, Bilog-MG, HLM6, WinSteps, FACETS
- Python, C, Java

TEACHING EXPERIENCE

Psychology Department, University of Illinois at Chicago; 2014-2016
Statistical Methods in Behavioral Science

- Teach undergraduate students introductory statistics; grade exams and homework

Northwestern Educational Center, Wheeling, IL; 2008-2015
Private Tutor

- Tutor students standardized assessments, math, and English

Global Youth Leadership Council, Washington, DC; 2008
Faculty Advisor

- Facilitate educational diplomacy simulation with 28 international high school students

YMCA, Burlington, WI; 2008
Wilderness Instructor

- Design ecology curricula and instruct students on biology and team-building

TEACHING EXPERIENCE (continued)

Cambridge Educational Services, Waukegan, IL; 2007
Teacher

- Teach third grade low SES students remedial math for federally mandated NCLB program

RESEARCH EXPERIENCE

NSF Data Consortium Fellow, Columbia University, New York; May 2016

- Develop data visualizations for large-scale educational data sets

PearsonVUE, Chicago, IL; 2016- Current
Psychometrician

- Develop innovative assessment items and measurement models for nurse certification exam

Learning Science Research Institute, University of Illinois at Chicago; 2010-2015
Research Assistant

- Analyze engineering assessments for validity and write articles for conferences and journals
- Perform psychometric tests on data sets (reliability analysis, factor analysis, IRT modeling)

Office of Campus Learning Environments, University of Illinois at Chicago; 2014-2016
Statistical Consultant

- Analyze data, write reports, and assist in design of beta study

College of Nursing, University of Illinois at Chicago; 2013-2014
Research Aide

- Manage large-scale surveys and datasets using SPSS and Excel for medical research

Richard Day Research, Evanston, IL; 2010-2012
Research Associate

- Manage large-scale surveys and datasets for Fidelity Insurance
- Create quarterly deliverables for clients using SPSS, Excel, WinCross

The Learning Partnership, Chicago, IL; 2009-2010
Consultant

- Perform content analysis on multiple choice assessments to determine validity and reliability
- Run regression analyses on student demographics to predict future student performance

Chicago Public Schools, Chicago, IL; March-July 2009
Intern

- Conduct interviews & qualitative research on fidelity of implementation of curriculum
- Inform curriculum developers for future curriculum design and assessment

PUBLICATIONS

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. (2015). An Analytic Framework for Evaluating the Validity of Concept Inventory Claims. J*ournal of Engineering Education, 104*(4), 454-496.

PUBLICATIONS (continued)

Jorion, N., Gane, B. D., DiBello, L. V., & Pellegrino, J. (2015, June). Developing and validating a concept inventory. Presented at the *2015 American Society for Engineering Education Annual Conference*, Seattle, WA.

Gane, B. D., Denick, D., DiBello, L. V., Pellegrino, J, & Jorion, N. (2015, June). Continuous improvement of a concept inventory: Using Evidence Centered Design to refine the Thermal and Transport Concept Inventory. Paper presented at the *2015 American Society for Engineering Education Annual Conference*, Seattle, WA.

Denick, D., Gane, B. D., Jorion, N., Miller, R. L., Streveler, R. A., DiBello, L. V., & Pellegrino, J. W. (2015, April). Continuing refinement of a concept inventory: Developing and selecting items for an expanded domain model. Poster presented at the *American Educational Research Association annual meeting*, Chicago IL.

Jorion, N., James, K., DiBello, L., & Pellegrino, J. (2014). Statistical analyses of performance on the CATS and the TTCI: Foundations of inventory validity and utility. In J. Pellegrino (Chair), Evaluating and improving concept inventories as assessment resources in STEM teaching and learning. Symposium conducted at the meeting of the *American Educational Research Association*, Philadelphia, PA.

Jorion, N., James, K., DiBello, L., & Pellegrino, J. (2014). Quantitative analyses of student performance on concept inventories. In J. Pellegrino (Chair), Evaluating and improving concept inventories as assessment resources in STEM teaching and learning. Symposium conducted at the meeting of the *American Educational Research Association*, Philadelphia, PA.

Pellegrino, J., DiBello, L., Miller, R., Streveler, R., Jorion, N., James, K., Schroeder, L., & Stout, W. (2013). An analytical framework for investigating concept inventories. In J. Pellegrino (Chair), The conceptual underpinnings of concept inventories. Symposium conducted at the meeting of the *American Educational Research Association*, San Francisco, CA.

Jorion, N., James, K., Schroeder, L., & DiBello, L. (2013). Statistical and diagnostic analyses of student performance on concept inventories. In J. Pellegrino (Chair), The conceptual underpinnings of concept inventories. Symposium conducted at the meeting of the *American Educational Research Association*, San Francisco, CA.

Jorion, N., Self, B., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2013). Classical Test Theory analysis of the Dynamics Concept Inventory. Proceedings of the *2013 American Society for Engineering Education Annual Conference*, Riverside, CA.

Pellegrino, J. W., DiBello, L.V., James, K., Jorion, N., & Schroeder, L. (2011). Concept inventories as aids for instruction: A validity framework with examples of application. In Proceedings of *2011 International Research in Engineering Education Symposium* (pp. 698-706). Madrid, Spain.

PUBLICATIONS (continued)

Pellegrino, J. W., DiBello, L.V., James, K., Jorion, N., & Schroeder, L. (2011). Concept inventories as aids for instruction: A validity framework with examples of application. Presented at the *Proceedings of Research in Engineering Education Symposium*, Madrid, Spain.

McGee, S., Witers, J., & Jorion, N. (2011). A*ssessing understanding of launch commit criteria using NASA's Kennedy Launch Academy Simulation System* (KLASS).

Jorion, N. (2010). Aligning classroom rigor to high stakes outcomes. Presented at *American Education Research Association Conference*, Chicago, IL.

WORKSHOPS

Play Data Conference, 1-Day Workshop, Chicago, IL; 2015
Learning Analytics Research, 3-Day Workshop, Boston, MA; 2014
Structural Equation Modeling, 2-Day Seminar, Miami, FL; 2011
Assessment in K-12 Mathematics Conference, Atlanta, GA; 2011

VOLUNTEER WORK

Interdisciplinary Undergraduate Research Journal, Chicago, IL; 2015
Peer Reviewer