5-2014

Development of the Biostatistics and Clinical Epidemiology Skills Assessment for Medical Residents

Patrick Brian Barlow
University of Tennessee - Knoxville, pbarlow1@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Patrick Brian Barlow entitled "Development of the Biostatistics and Clinical Epidemiology Skills Assessment for Medical Residents." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Educational Psychology and Research.

Gary J. Skolits, Major Professor

We have read this dissertation and recommend its acceptance:

Jennifer A. Morrow, William P. Metheny, Shawn L. Spurgeon, Kent Wagoner

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Development of the Biostatistics and Clinical Epidemiology Skills Assessment for Medical Residents

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Patrick Brian Barlow

May 2014

Dedication

I lovingly dedicate this work to my parents, Ken and Theresa, and sisters, Meredith and Caroline, for supporting me in every possible way from pre-k to Ph.D. I love you all very much.

I also dedicate this work to my friend, mentor, and colleague, Dr. Phil Kramer. I could not have even defined assessment, let alone successfully made it into a graduate program and out the other side without your guidance – thank you.

Acknowledgements

There are far too many individuals to possibly thank for their influence in my graduate career but I shall do my best.

I would like to first acknowledge the outstanding support of my committee chair, Gary Skolits, for his steadfast guidance and leadership throughout my graduate career. I would like to thank my committee members, Dr. Jennifer Morrow, Dr. William Metheny, Dr. Shawn Spurgeon, and Dr. Kent Wagoner, for lending their expertise, their support, and most of all their time to what has been an excellent experience.

I also must thank Dr. Eric Heidel and Tiffany Smith, my team at the UT Graduate School of Medicine. I specifically thank Eric for his invaluable assistance with collecting data for this project and reviewing the manuscript. I thank Tiffany for simply being my partner since day one of this adventure.

I also wanted to thank Stefanie Strapko Rieck, Kelly Menard, and all of my amazing friends and colleagues who served as external reviewers, editors, and psychological support these past four years. Likewise, to all of the past and present students of the Evaluation, Statistics, & Measurement program, thank you for the innumerous ways my work was improved due to your individual skills and expertise.

Finally, I need to acknowledge my wonderful participants at the three medical centers who chose to take a statistics test rather than their lunch break. There would be no study without your work.

Abstract

This study developed the Biostatics and Clinical Epidemiology Skills (BACES) assessment, and established its preliminary item characteristics and validity evidence. Unlike previous instruments, the BACES assessment was developed and tested using an item response theory (IRT) approach to measurement to create a new, adaptive biostatistics and clinical epidemiology knowledge assessment for graduate medical professionals. Thirty multiple-choice questions were written to focus on interpreting relevant examples of clinical epidemiology and statistical methods. A four person expert panel reviewed these items for content validity. After this review, the BACES assessment was administered to 147 medical residents across three academic medical centers. Results of the IRT analysis produced a final instrument of 26 items with 13 devoted to statistical methods and 13 to clinical epidemiology, which successfully fit a 2-parameter IRT model. In contrast to previous assessment research, an IRT approach allowed for each BACES item's difficulty, discrimination, and reliability to be estimated *separately* from the sample on which it was tested. As a result, this preliminary study has paved the road for a flexible yet psychometrically rigorous instrument for measuring the biostatistical and clinical epidemiologic knowledge of graduate medical students.

Table of Contents

List of Tables

List of Figures

Chapter One: Statement of Problem

Introduction

This first chapter situates the proposed study within the context of graduate medical education, specifically biostatistics and clinical epidemiology education. The historical context for teaching these topics will be outlined as well as previous attempts to assess graduate medical students' knowledge of them. The proposed research objectives and methods to achieve these objectives will also be discussed in addition to a definition of key terms, assumptions, and limitations of the proposed study.

Statement of Problem

The dominance of Evidence Based Medicine (EBM) in Graduate Medical Education (GME) over the past twenty-five years makes translating medical evidence into clinical decision making an important skill for residents (Hatala & Guyatt, 2002). Although biostatistics and clinical epidemiology are essential components to comprehending the medical evidence (Sahai, 1999), the evidence has shown a consistently low and variable knowledgebase within the GME population (Berwick, Fineberg, & Weinstein, 1981; Novack, Jotkowitz, Knyazer, & Novack, 2006; Weiss & Samet, 1980; Windish, Huot, & Green, 2007). At the same time, there has been an *increase* in the frequency and complexity of statistical methods among the top tier medical journals (Horton & Switzer, 2005; Reed, Salen, & Bagher, 2003; Weiss et al., 1980; Windish et al., 2007).

Many EBM curricula now include content dedicated for biostatistics and/or clinical epidemiologic research methods in order to respond to this problem. However, the length, format, and rigor of these courses is quite variable as are the qualifications of course instructors (e.g.

resident versus faculty led) (M. L. Green, 2001; M. Green, 1999). This environment has made assessment of these skills difficult (Hatala & Guyatt, 2002).

Psychometric properties of previous instruments.

As previously noted, there have been a number of attempts at assessing this challenging population (e.g. Berwick et al., 1981; Enders, 2011; Fritsche, Greenhalgh, Falck-Ytter, Neumayer, & Kunz, 2002), yet the formal psychometric analysis of these instruments has been absent. Moreover, a 2011 review of existing instruments made an explicit call for new and better biostatistics and clinical epidemiologic knowledge (BEK) assessments in this population (Enders, 2011).

The Enders (2011) review noted several content-related shortcomings of existing instruments; however, an examination of the items reveals additional areas for improvement. Specifically, the dominant format for these instruments, and indeed the "gold standard" for medical assessment in general, is the multiple choice question (MCQ) (Brunnquell, Degirmenci, Kreil, Kornhuber, & Weih, 2011). Writing high quality MCQs requires adherence to an extensive list of common item writing practices (Brunnquell et al., 2011; Case & Swanson, 2002; S. M. Downing, 2005). Each instrument possessed a number of "violations" of these common practices, which impacted the validity of the tests. Furthermore, these instruments were developed from a Classical Test Theory (CTT) perspective, which does not allow the test items to be broken-up and reorganized to meet specific educational needs without damaging the instrument's reliability. If new research is to heed the call for new instrumentation for BEK, then a *new* measurement strategy that meets the specific needs of the GME community must be considered. To this end, Item Response Theory (IRT) provides stable estimates of an item's difficulty, discriminative ability, and guessing probability that are *invariant* to changes in sample,

item order, and test conditions. Use of IRT in developing a new, flexible assessment for the unique GME population addresses the salient problem of *how do educators effectively prepare and assess physicians in biostatistics and clinical epidemiology?*

Study Purpose and Objectives

The purpose of the present study is to establish preliminary item characteristics and validity evidence for the Biostatics and Clinical Epidemiology Skills (BACES) assessment. The present study aimed to leverage the power of Item Response Theory (IRT) to create a new, adaptive biostatistics and clinical epidemiology knowledge (BEK) assessment for graduate medical professionals. The following chapter will detail the methodology to meet these three research objectives:

1. Establish content validity evidence of the BACES assessment
2. Examine the model fit of the BACES items to a 1-parameter logistic (1PL)/Rasch, 2-parameter logistic (2PL), and 3-parameter logistic (3PL) IRT model
    a. Test for violations of essential unidimensionality and local independence
    b. Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment
    c. Analyze the quality of item distractors on the BACES assessment
    d. Analyze the total item and test information produced from the BACES instrument
    e. Compare person and item location estimates from IRT models to those of traditional CTT indices.
3. Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.

Importance of the Study

Training in Evidence Based Medicine requires residents be able to read the statistical *evidence* on which their clinical decisions are *based.* The present study aimed to lay the groundwork for improving the BEK assessment in GME by establishing preliminary item parameters and validity evidence for the BACES assessment. There have been no BEK assessments to date that have utilized an IRT measurement approach, which will enable researchers and educators to adapt the BACES items to whichever difficulty (e.g. first-year vs. fourth-year residents) or purpose (e.g. study practice, self-assessment, exam, etc.) without losing the items' reliability, difficulty, or discrimination.

Overview of Methodology

The present study employed a multisite, cross-sectional design using a convenience sample of 147 residents from three large, academic medical centers. Although there are no definitive sample size requirements for the IRT analyses (Edelen & Reeve, 2007), the study used benchmark sizes of between 100 and 500 participants per existing recommendations (Lord, 1983; Drasgow, 1989).

Instrumentation

Instrumentation for the study consisted of the BACES assessment itself along with several demographic items for validity purposes. Content for the BACES items was developed using four sources:

(1) Learning objectives from the biostatistics and clinical epidemiology curriculum taught at the University of Tennessee Graduate School of Medicine;

(2) Commonly used statistics in medical literature as defined by existing reviews (Horton & Switzer, 2005; Reed et al., 2003; Windish et al., 2007);

(3) Common content areas among existing assessment instruments;

(4) Content gaps relating to clinical and translational science and public health core competencies (Enders, 2011).

These items were written in an MCQ format with four response options per question. In accordance with best practices for medical testing (Jozefowicz et al., 2002) and existing BEK assessment measures, the BACES items focused on using clinical or literature-based vignettes to emphasize residents' application of BEK concepts rather than rote memorization. Each item contains a unique case vignette rather than using the same vignette for multiple items so as to avoid interlocking (dependent) items that may violate the IRT assumption of local independence. Once written, these items were reviewed by a panel of five content experts using a standard rubric (Appendix A). Changes in the instrument were made after the panel review, and the final set of items was put into two parallel forms for administration.

Analysis

All responses were collected via group administration, scanned into digital format using Remark OMR 8 (Gravic, Inc.), and transferred into Microsoft Excel 2013 (Microsoft Corporation) for initial recoding. Correct responses to assessment items were keyed as dichotomous "correct" or "incorrect" for use in item analysis, but the original responses were also kept for performing distractor analyses. Demographic responses were also coded appropriately for follow-up validity and group comparison analyses. Omitted responses were given a simulated value in order to better facilitate IRT person-location estimates (de Ayala, 2009). Finally, preliminary construct validity evidence was sought using known-groups validity comparisons between demographic characteristics (Devellis, 2012).

One-parameter Rasch, 2PL, and 3PL models were fit and compared for the best-fitting, most parsimonious model. Overall model fit was assessed via a chi-square goodness of fit index, and by comparing the change in -2 Log Likelihood statistics between models (de Ayala, 2009). Item difficulty, discrimination, and pseudo-guessing parameters were estimated using an expectation-maximization method, and item fit was assessed using standardized residuals and the item characteristic curves. Person location was estimated using an *expected a-posteriori* (EAP) method with a standard normal prior distribution ($M_{(\theta)}$=0.00 $SD_{(\theta)}$=1.00). Also, a standard error of estimate was calculated for each item and used to examine item and total test information (reliability). Further, the quality of each item option was assessed using traditional CTT distractor analysis to compare the frequency of distractor choices between the top and bottom 25% of examinees (Wise, n.d.). Finally, the parameter estimates for the best-fitting IRT model were correlated to their CTT equivalents as a way of further checking the accuracy of the IRT model estimates (Fan, 1998; Hays et al., 2000; Stage, 1998; Xu & Stone, 2011).

All IRT analyses were conducted using Xcalibre v4.2 (Guyer & Thompson, 2012). CTT item analysis and validity analyses were conducted using IBM SPSS v.22 (SPSS Inc., Chicago IL).

Definition of key terms and abbreviations

Biostatistical and Clinical Epidemiologic Knowledge (BEK): Defined in the context of the present study as the ability to correctly identify, interpret, and apply fundamental statistical and epidemiologic theory, commonly used statistical tests, and common epidemiologic research methods relevant to clinical practice. It was derived from the ACGME Core Competencies for Medical Knowledge and Practice-based learning and improvement (ACGME, 2013_a) as well as the body of literature on physician knowledge of biostatistics and clinical epidemiology.

Graduate Medical Education (GME): the period of medical training that follows graduation from medical school; commonly referred to as internship, residency, and fellowship training.

Classical Test Theory (CTT): Also known as *true-score theory* or the *classical measurement model*, CTT views an individual's trait score on the latent variable (i.e. their fixed location on the variable of interest) as a function of their observed score on a measurement scale plus measurement error (de Ayala, 2009). It assumes that these error values are (a) randomly dispersed among the scale's individual items; (b) not related to one another; and (c) not related to the true score on the latent variable.

Item Response Theory (IRT): A measurement approach that contrasts from CTT, Item Response Theory postulates that an individual's response to a test item is a function of their position on a continuous latent trait denoted by the Greek letter "θ" (theta) (DeMars, 2010). IRT is comprised of a system of mathematical models that estimate the probability of a certain response (answering correctly in this study) across different θ levels given the item's difficulty, discrimination, and pseudo-guessing parameters.

Item/Test Characteristic Curve (ICC): A graphical representation of the probability of correctly responding to an item across a continuum of trait levels (θ).

IRT Parameters (de Ayala, 2009 & DeMars, 2010)

> **Latent Trait Distribution (θ, Theta):** The distribution for the latent trait an instrument purports to measure. These trait levels are measured on a continuum along the horizontal axis of an ICC with a mean of 0.0 and standard deviation of 1.0 exactly as a z-score distribution. For example, an individual with an average trait level would be located at θ=0.0, and the majority of individuals will fall between θ= -3.0 and θ = 3.0 (DeMars, 2010).

Item Discrimination ("a" or "α" Alpha): The slope of the ICC line, typically ranges from 0 – 3. The ability for an item to differentiate between individuals at high or low levels of ability (θ, in the case of IRT) (de Ayala, 2009). Also known as the "A" parameter (DeMars, 2010).

Item Difficulty ("b", or "δ" Delta): IRT defines item difficulty as the point of inflection on an ICC. Using the simplest IRT model, difficulty is the location on the latent trait continuum (θ) where a person has a 50% probability of giving the correct answer. Also known as the "B" parameter (DeMars, 2010).

Item Psuedo-Guessing ("c", or "χ" Chi): Represented as the lower-asymptote on an ICC, which is, "The value the function approaches as θ approaches negative infinity" (DeMars, 2010, p. 13). Pseudo-guessing is the probability that someone with a very low level of θ will answer an item correctly given chance alone (de Ayala, 2009). Also known as the "C" parameter (DeMars, 2010).

IRT Models (de Ayala, 2009 & DeMars, 2010)

One-Parameter Logistic (1PL or Rasch) Model: Although slightly different mathematically, the 1PL and Rasch model represent the two simplest IRT models. In each of these models, only the difficulty "b" and theta parameters are estimated while both item discrimination "a" and pseudo-guessing "c" are held constant at 1.0 and 0.0, respectively.

Two-Parameter Logistic (2PL) Model: A slightly more complex model than the 1PL or Rasch approach, which allows for item difficulty "b", discrimination "a", and person location (theta) to be estimated. In this model, only the pseudo-guessing parameter "c" is held constant at 0.0.

Three-Parameter Logistic (3PL) Model: The 3PL model is the most complex IRT model described in this study, and it allows the "a", "b", "c" and theta parameters to all be estimated. Although this is the most complex model, the addition of the pseudo-guessing parameter requires very large sample sizes for it to be accurately estimated.

Item/Test Information: The concept of reliability from an IRT perspective is known as *item* and *test information*, which is the extent the researcher can be certain of a person's location along θ. For each item, the amount of information is proportionate to the standard error of estimate (SEE) for each possible θ location, and smaller SEE indicate more certainty (more information) (de Ayala, 2009). An item provides its highest amount of information near its difficulty value ("b") (DeMars, 2010).

Limitations

Shadish, Cook, and Campbell (2002) outlined a number of statistical conclusion and internal validity threats that applied to the study. With regards to statistical conclusion validity, the most formidable threat was low statistical power brought about by a small sample size. This limitation is particularly visible in the known-groups validity comparisons. Internal validity was also threatened by the convenience sample procedure (sampling bias) and inability to tightly control the testing environment (valid data / self-report, attrition). For example, there was anecdotal evidence that several participants did not sincerely complete the instrument, which may have skewed results.

Although there was no way to eliminate these limitations, steps were taken at every point to minimize their impact. The impact of the statistical power limitation was minimized through using a multisite, multi-specialty resident sample, and by selecting a 2PL IRT model that has

been shown to be appropriate for sample sizes of between 100 and 500. The multisite, heterogeneous sample also sought to minimize the sampling bias introduced with the non-randomized study. To minimize issues of cheating and valid data, all administrations were proctored in person using a paper and pencil, group administration format. Finally, standardized instructions for proctoring the assessment were used for each administration as well as, to the extent possible, a common testing condition (i.e. journal club) in order to minimize extraneous environmental factors.

Organization of the Study

Chapter one has briefly introduced the problem under investigation, its context, three primary study objectives, and the methodological components that the study used to address these objectives. This chapter has also highlighted the assumptions, limitations, and key definitions for the study.

Chapter two will present a complete review of the literature that informs the present study as well as the theoretical framework on which it is based. Chapter three will illustrate details of the study's methodology for developing the BACES assessment as well as administering and analyzing the results. Chapter four will provide the results from collected data, and chapter five will discuss these findings in detail as well as the study's implications and recommendations for future research.

Chapter Two: Literature Review

Section One: Biostatistics and Clinical Epidemiology as Part of Medical Education

The purpose of this chapter to introduce biostatistical and epidemiologic concepts relevant to the teaching of Evidence-Based Medicine (EBM) as well as offer a discussion regarding what is known about physicians' attitudes and knowledge towards these two topics. Physicians' attitude towards and knowledge of biostatistical and epidemiologic concepts is not a new area of inquiry; rather, these topics have been under investigation since the 1980s (e.g. Berwick et al., 1981; Weiss & Samet, 1980). Accordingly, this chapter will also focus on the methodologies used by previous assessments of biostatistical and epidemiologic knowledge (BEK). Finally, the chapter provides an overview of the psychometric approaches to objective test construction including the fundamentals of Item Response Theory (IRT), and how it compares to Classical Test Theory (CTT).

In order to better understand the need for increased resident education in biostatistical and epidemiologic concepts, it is necessary to consider the educational context of residents. Evidence Based Medicine (EBM) has become the dominant medical education paradigm since the Evidence-Based Medicine Working Group (1992) found it to be superior to the pedagogy of the time. Sackett and Rosenburg (1996) defined EBM as, "The conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (p. 71). Green (2000) indicated that the process consisted of four steps or skills:

> "(1) Convert emerging medical information needs into answerable questions; (2) efficiently search for the best information; (3) appraise the evidence for its validity and usefulness; and (4) integrate the evidence into the decision making for an individual patient" (p. 121).

Evidence Based Medicine has become popular among Graduate Medical Education educators; one report found 37% of United States and Canadian internal medicine residencies had dedicated time for EBM (Hatala & Guyatt, 2002). The Accreditation Council for Graduate Medical Education (ACGME) adopted a series of core program requirements, which mandate that residents, "Apply knowledge of statistical methods to the appraisal of clinical studies" (Morreale, Balon, & Arfken, 2012; ACGME, 2012).

Teaching biostatistics and clinical epidemiology to medical students and residents

Before entering an academic medical center, the majority of residents will have had *some* exposure to biostatistics and/or epidemiology in their undergraduate medical school (Looney, Grady, & Steiner, 1998). Biostatistics and epidemiology have been present in the medical school curricula for the greater half of a century (Sahai, 1999), and a number of studies have looked at both the content and structure of these topics over the years (Looney et al., 1998; Sahai, 1999). Looney and his colleagues conducted a cross-sectional survey of all 125 medical schools in the United States in 1993 to update the knowledgebase of what and how biostatistics and epidemiology are taught in medical schools. A biostatistical course was required in 89% of the 100 medical schools that responded to the survey (p. 92). The course was primarily taken in the first and second year among those schools that required it (55% and 32%, respectively), and very few schools had courses that continued for more than a single academic year (5%) (p. 93). Although the vast majority of medical school instructors surveyed felt that they had sufficient time to cover necessary biostatistical and epidemiologic topics, the median number of instructional hours was as low as 20 hours per course (range of 2 to 48) to cover an average of 25 topics (pgs. 93-94). The authors concluded, "The amount of instructional time in the required

courses was rather limited especially when one considers that 25 topics were covered in at least 75% of the courses" (p. 94).

The next year, Sahai (1999) offered a critique of teaching methods for biostatistics and epidemiology in medical school under the claim that, "Undoubtedly, physicians and other medical professionals are becoming increasingly aware of their need for biostatistical principles and methods, and a basic knowledge of biostatistics is considered to be of prime relevance to every medical professional…" (p. 188). He joined Looney et al. (1998) in highlighting the variant levels of exposure medical students across institutions receive in these topics; however, Sahai claimed that the inability for medical students to see the relevancy in statistics education was the key factor in making biostatistics so difficult to teach. He addressed this concern by suggesting, "If a biostatistics instructor fails to use practical problems and the proper method of handling and communicating solutions, his or her expositions, even when correct and intelligible, and may become a source of confusion rather than illumination" (p. 193).

As previously described, teaching EBM in Graduate Medical Education (GME) relates to the fundamental standards put forth by the Accreditation Council for Graduate Medical Education (ACGME) (Hatala & Guyatt, 2002). Green (2000) reviewed a substantial number of medical residency programs across the U.S. to describe the ways in which EBM was being addressed in school curricula. He discovered that the most commonly used approach to teaching EBM through journal clubs aimed at, "Improving residents[sic] critical appraisal skills" (p. 123). The typical format for these clubs was a group of residents who meet to engage in a series of critical discussions on articles relevant to their practice or a particular lesson. Six of the fourteen EBM-focused journal club curricula Green reviewed in his study included some reference to research methodology, biostatistics/statistical concepts, and/or epidemiology in their objectives;

however, four of these six were exclusively resident-directed (i.e. limited or no faculty involvement), and another provided only two didactic sessions (i.e. lectures) on the topics. Similarly, Cheatham (2000) investigated the prevalence of statistics education in journal clubs through a survey of 77 (62 responded) southeastern general surgery programs. He found that although 81% of those who responded had a resident journal club, only 33% of them indicated that statistics was part of their post-graduate medical education curriculum.

The second avenue by which medical residents receive their EBM, biostatistical, and clinical epidemiologic training is through freestanding EBM curricula (M. L. Green, 2001). Green defined such curricula in his review as, "self-contained learning sessions that occur during dedicated curricular time" (p. 126). Of these freestanding curricula, 35% of the 99 reviewed included references to research methodology, biostatistics/statistical concepts, and/or epidemiology in their objectives. Several of these freestanding EBM curricula will be described in detail in a forthcoming section.

Undergraduate medical students typically encounter biostatistics and/or clinical epidemiologic concepts through their coursework (Looney et al., 1998) while graduate medical students use either freestanding EBM curricula, or journal clubs to teach these methods (M. L. Green, 2001). Different pedagogical strategies notwithstanding, the emphasis on problem-based learning and application of knowledge is clear in both educational environments (Sahai, 1999)

Most commonly taught statistics in medical schools and found in the medical literature

Now that both undergraduate and graduate medical education teaching strategies have been outlined, the next step is to discern which BEKs are the most important for students at each level to understand. Previous research has synthesized both frequently taught topics in medical schools and frequently used statistics in medical research in response to this issue.

The frequency at which certain statistical concepts are addressed in medical school education was also addressed by both Looney et al. (1999) and Salhai (1999). Looney and colleagues reviewed 74 *required* biostatistics courses, and found interpreting *p*-values (95%), hypothesis testing (93%), interpreting confidence intervals (93%), descriptive statistics (92%), and *t*-tests (92%) as the top five topics taught to medical students. Similarly, Salhai (1999) found *p*-values (94.8%), interpreting confidence intervals (93.1%), hypothesis testing (89.7%), frequency distributions (86.2%), and *t*-tests (86.2%) to be the top five statistical concepts. Both studies found epidemiologic research topics taught with similar frequencies with case-control studies, cohort studies, and randomized control trials taught in roughly 91%, 91%, and 88%, respectively.

The first source of BEK topics comes from reviews of the medical literature. It has been shown that journal clubs' primary objective is usually improving critical appraisal of the medical evidence (Green, 2001; Cheatham, 2000). This focus raises the question, *how are biostatistical and epidemiologic concepts found in the medical evidence?* Numerous reviews have been conducted in a number of major specialty and general medicine journals over the last 40 years to answer this question. The majority of these reviews build upon the work of Emerson and Coldiz (1983) who developed a typology of statistical concepts used in the New England Journal of Medicine. The authors developed a hierarchy of increasing statistical sophistication while trying to gauge which statistical concepts a physician must know in order to read published evidence. Since its publication, a number of other researchers have used this hierarchy to update Emerson's review of the *NEJM* (Emerson & Coldiz, 1992; Horton & Switzer, 2005; Switzer & Horton, 2007), review statistical methods multiple journals (Windish, Hout, & Green, 2007), specialty journals (Hellems, Gurka, & Hayden, 2007), and international journals (Wang & Zhang, 1998;

Rigby, Armstrong, Campbell, & Summerton, 2004; Karan, Goyal, & Bhardwaj, 2009). All of the aforementioned studies draw the similar conclusion that the use of statistical methods in published literature is steadily growing in both frequency and sophistication; however, the top three to five most commonly used statistics in medical research have continuously been descriptive statistics, t-tests, and contingency tables, regardless of specialty or country of publication. Table 2.1 below provides a summary the Switzer & Horwitz 2007 review, which was chosen because the *New England Journal of Medicine* has the broadest audience of practicing physicians when compared to some of the other reviews.

Table 2.1.

Comparison of Fifteen Commonly Used Statistics in The New England Journal of Medicine Between 1989 and 2004 – 2005 (adapted from Horwitz & Switzer 2005; & Switzer & Horwitz, 2007)

Statistical Procedure	Articles Containing the Procedure	
	1989 Review (*N*=115)	2004 – 2005 Review (*N*=311)
t-tests	39%	26%
Contingency tables	36%	53%
Survival methods (including logistic regression)	32%	61%
Epidemiological Statistics (risk, measures of association, sensitivity and specificity)	22%	35%
Nonparametric tests	21%	27%
Analysis of variance (ANOVA)	20%	16%
Pearson correlation	19%	3%
Multiple Regression	14%	51%
No statistics or descriptive statistics only	12%	13%
Multiple comparisons (post hoc analysis)	9%	23%
Simple linear regression (single predictor, single dependent)	9%	6%
Power analysis	3%	39%
Repeated measures analysis	-	12%
Noninferiority / Equivalence trials	-	4%
Receiver Operating Characteristic (ROC) Curve	-	2%

Note. Total number of procedures used = 297 in 1989 and 1271 in 2004 – 2005. The average methods per article in 1989 and 2004 – 2005 was 2.9 and 4.7, respectively.

Section Two: Previous Assessment of Biostatistical and Epidemiologic Concepts

The following section focuses on previous assessment of physicians' biostatistical and epidemiologic knowledge. The first step will be to revisit the results of previous assessments discussed in Chapter One to expand upon what is known about physicians' knowledge in the two areas. The second step will address the assessment instruments themselves to compare the psychometric properties and methodological vulnerabilities of each. Finally, the section

concludes with a comparison of two strategies for item development: literature review and core discipline competencies.

Assessment of physicians' biostatistical and epidemiologic concepts.

Physician numeracy, "The ability to understand the quantitative aspects of clinical medicine, original research, quality improvement, and financial matters" (Rao, 2008, p. 355) has been acknowledged as an essential component to effective practice for many years. Unfortunately, previous studies have shown a consistent lack of physician confidence in these areas. A 1987 survey found that 85% of graduating medical residents saw statistical methods as vital to effectively use medical literature, yet two-thirds of those surveyed admitted to having limited or no knowledge of these areas (Reznick, Dawson-Sanders, & Folse, 1987). Not surprisingly, Swift et al. (2009) found that while 79% of physicians surveyed (N=130) agreed that knowledge of probability and statistics were important, 63% stated there were activities they could do better if they knew more about the topics. A similar survey of physician attitudes toward biostatistics reported only 17.6% of the 301 respondents felt that their statistical training was adequate to meeting their needs while only 14.6% felt they could conduct their own analysis (West & Ficalora, 2007). Moreover, only 21.6% and 38.6% of academic clinicians and academic researchers, respectively, agreed (or strongly agreed) that they were able to tell when a correct statistical test had been applied in a study. Windish et al. (2007) found in their multi-institutional survey that while 75% of the 367 respondents reported that they did not understand all of the statistics they encountered in the literature, 58% of them indicated they used statistical results when making decisions for patient care. Confidence in specific statistical concepts was no better. When respondents were asked to rank themselves from 1-5 on their confidence with key

statistical concepts, they reported a mean confidence rating of 11.4 (SD=2.7) of a possible 20 (i.e. all "complete confidence").

Studies across the last few decades have consistently found a low yet widely variant knowledge in biostatistics and clinical epidemiology. Weiss and Samet (1980) used the most common statistical concepts in the medical literature that they found in their own multi-journal review to create a questionnaire of residents' biostatistical knowledge. They created a 10-item exam on statistical concepts that was administered to 141 practicing physicians, and found the mean score to be 74% correct with higher scores attributed to participants' previous biostatistics or epidemiology training.

Similar results were found in a similar study by Berwick, et al. (1981) who concluded a broad lack of knowledge in physicians who completed their 36-item statistics Self-Assessment Questionnaire (SAQ). This instrument included questions on five areas deemed important by the authors' views of statistics in the medical literature including: 10 items on definitions; six items related to knowledge of basic properties of statistical data (e.g. Bayesian Theory); 10 items on limiting inferences to those shown by the data, and five items related to interpretation from data. The five final items focused on what the authors referred to as "expected value calculations" (p. 993), which was described as, "The ability to combine utilities (i.e. the values attached to outcomes) with probabilistic information according to the rules of decision theory so as to maximize utilities" (p. 993). The 281 participants scored an average of 63% on the SAQ with medical students and academic physicians scoring significantly better than practicing physicians (72% and 55%, respectively). The SAQ study set the stage for a series of similar studies in the next 30 years, which have continued to show inadequate training biostatistics and epidemiology among medical professionals.

Among these similar studies was a 2002 assessment by several German physicians who were studying the impact of short, intensive courses on EBM (Fritsche, et al., 2002). They developed a 15-item assessment instrument in which course participants were given a series of clinical research scenarios linked to published studies. Fritsche and colleagues saw a significant improvement in *overall* EBM knowledge from pretest to posttest of roughly 3.6 points; however, posttest mean scores still did not exceed 60% (9.9 of 15 correct).

More recent studies of physician numeracy have also found similar results. Both L. Novak et al. (2006) and Windish et al. (2007) reported disappointing performance on their own assessment instruments. These more recent studies showed average scores of 40% (4/10 items correct) and 41.4% (8.3/20 items correct), respectively. Ahmadi-Abhari, Soltani, and Hosseinpanah (2008) found similar results with their small BEK assessment (6 items) averaging only 50% in a sample of 104 residents and sub-specialty fellows. The consistent stream of evidence over the past thirty years has left little controversy in concluding a low and variable mean knowledge of biostatistics and epidemiology among medical professionals. Rather, the salient question becomes *how do educators effectively prepare and assess physicians in these areas?*

Best practices for writing multiple-choice questions.

Effectively preparing and assessing physicians' BEK will require new assessment instruments (Enders, 2011). A recent review found at least three formal item writing flaws in all 40 of the continuing medical education items published in the *New England Journal of Medicine* (Stagnaro-Green & Downing, 2006), so reviewing standards for properly written items is a logical first step in developing a new assessment. The following section will outline some of the most common recommendations and guidelines for writing effective assessment items in both a

broad context and medical education, specifically. Examples from existing BEK assessments are used whenever possible to illustrate these guidelines.

The multiple-choice question (MCQ) is still considered the gold standard for objective test development, particularly in high-stakes testing (Brunnquell et al., 2011); in fact, all of the existing BEK assessments use MCQs exclusively. These MCQs generally take two formats in medical assessments: 1) true/false (TF) and 2) one-best-answer (OBA) (Case & Swanson, 2002). As shown in the first example in Table 2.2, the TF format presents the examinee with a response set which includes a single, "true" answer. If options are not absolutely true or false, then the examinee must use their own definitions of the concept or resort to guessing as to what the test writer thought was true (Case & Swanson, 2002). These items are considered to have content and psychometric shortcomings, and are no longer used by high-stakes tests such as medical licensure exams (Case & Swanson, 2002). By contrast, response options for an OBA item can be qualitatively ranked from least to most correct; therefore, they give the instructor greater information on where the examinee went wrong in their thinking (Case & Swanson, 2002). This format is preferable to the TF method because the blurred (or situational-dependent) line between "true" and "false" is the focus of the MCQ rather than an unintended, unmeasured consequence. In practice, "many item writers believe the true/false items are easier to write than one-best-answer items" (Case & Swanson, 2002, p. 18); however, the authors conclude that using the TF format is *not* recommended. OBA items have been advocated as a better option than TF items (Brunnquell,et al., 2011; Case & Swanson, 2002; Downing & Baranowski, 1995).

Table 2.2.
Example True/False and Single-Best-Answer MCQ Items

Stem (Source)	Response Options
Any systematic error in the design, conduct, or analysis of a study that results in a mistaken estimate of an exposures' effect on the risk of disease is called: (Windish et al., 2007, p. 1019)	a. Confounding **b. Bias** c. Interaction d. Stratification
A study investigating an effect of a new drug for decreasing blood pressure should be a study of type: (Novak et al., 2006)	a. Retrospective cohort study b. Prospective case-control study **c. Double-blind placebo-controlled study** d. Cross-sectional study

The most prominent item flaws on existing BEK assessments involve item dependencies, or "interlocking items" (Suskie, 2009, p. 170). Essentially, the answer to one item should not be given in the stem of another and vice-versa. The items become dependent because one item directly influences the response to a subsequent item for a reason other than knowledge (DeMars, 2010). In truth, this error ought to be rightfully called a "time-saver" rather than a consequence of naive test construction because avoiding dependencies logically implies that one cannot use a single figure/case/vignette/etc. for multiple questions (Case & Swanson, 2002; DeMars, 2010; Suskie, 2009). One example of item dependency can be seen in an example from Enders' instrument, the REsearch on Global Regression Expectations in StatisticS (REGRESS) assessment (Retrieved April 2013 from: http://bit.ly/ZOgFHe).

"A group of investigators gathered vital statistics taken during annual checkup visits. Their goal was to create models to help identify typical values for on vital statistic based on another. Their results are presented in questions 1-7.

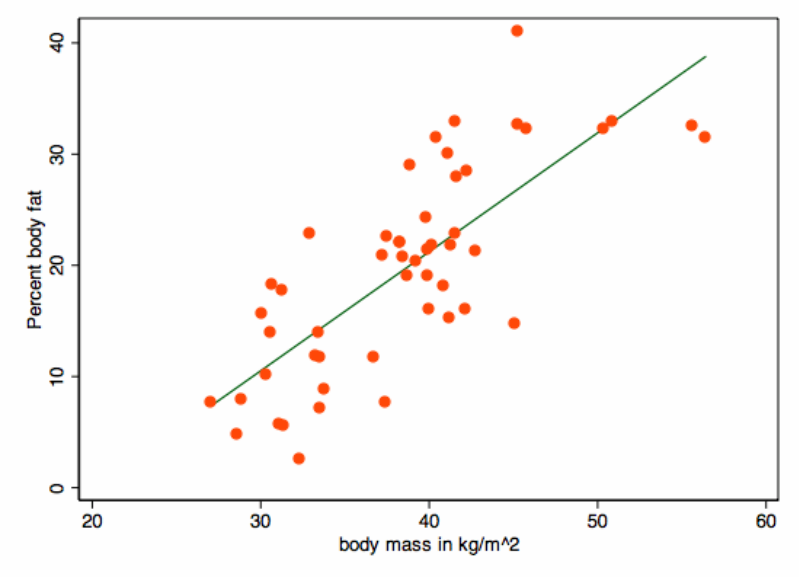


Figure 2.0-1. Example REGRESS MCQ Item

Items

1) What is a reasonable value for the slope in the graph above?

    a. 0

    b. 1

    c. 20

    d. -20

    e. I don't know

2) What is a reasonable value for the Y-intercept in the graph above?

    a. 0

    b. 1

    c. 20

    d. -20

    e. I don't know" (Enders, 2013, http://bit.ly/ZOgFHe)

These two items, and the five that followed, exhibit at least two opportunities for item dependencies to occur. First, the same graph is being used for the first seven questions, which

means that if a student is not comfortable with (or "I don't know" is selected) for question one, then they are at a disadvantage when answer the remaining items on that graph. Second, the identical response options for these two items provide an easy process of elimination opportunity for the examinee. If "(c) 20" was the correct answer to the slope of the graph (item one), then anyone familiar with a linear function will know that the slope and Y-intercept (item 2) will not be identical; therefore, "(c) 20" will not be the answer to item two.

Writing an unfocused stem is also a very common item-writing flaw. An unfocused stem fails to give sufficient information to the examinee for them to answer the question correctly (De Champlain, 2010). Put another way, "The student shouldn't have to read the options to discern the question" (Suskie, 2009, p. 171). Assessment at the undergraduate medical education level has shown unfocused stems to contribute to reduced correct response rate (Brunnquell et al., 2011). A number of items in the BEK assessments reviewed thus far contain one or more items with an unfocused stem. For example:

1) "In a research study, the age of the participants was 26 years +/- 5 years (mean +/- standard deviation). Which of the following statements is the most correct?
    a. It is 95% certain that the true mean lies within the interval of 16-36 years.
    b. Most of the patients were aged 26 years; the remainder were aged between 21 and 31 years.
    c. Approximately 95% of the patients were aged between 16 and 36 years.
    d. No patients were younger than age 16 or older than age 36." (Windish et al., 2007, p. 1019).

It would be very unlikely that an examinee would know for what statement the stem is asking without having to read through each of the options, which puts an additional time constraint on the examinee.

A negative stem, although not common in the reviewed studies, warrants brief discussion. Use of words such as *not* and *all except* in an item stem can be quickly overlooked by an examinee (Suskie, 2009) as well as damage both readability and difficulty (Brunnquell et al, 2011, Case & Swanson, 2002). Similar advice is given about using phrases such as "Which of the following…" (Suskie, 2009).

Construct irrelevant difficulty is another key area where item-writing errors are made. Test questions should be difficult because of the concepts being tested, and not because they were poorly written. To this end, Suskie (2009) offers two key precepts to follow: "Remove all the barriers that will keep a knowledgeable student from answering the item correctly...[and] Remove all clues that will help a less-than-knowledgeable student answer the item correctly" (p. 170). In medical testing, the highest quality MCQs incorporate a clinical vignette within the usual format of stem and responses model (Jozefowicz et al., 2002). The challenge for the item writer is to create these vignettes that present a sufficient challenge, yet maintain appropriate focus on the concept being tested rather than, say, medical knowledge. One approach to this hazard has been to construct vignettes from typical clinical situations or broad areas of medicine such as internal medicine, family medicine, or general practice examples (Fritsche, Greenhalgh, et al., 2002; Windish et al., 2007;). Other more general risks for irrelevant difficulty changes include:

- failing to write a concise stem/vignette;

- grammatical clues in the stem (i.e. a/an is/are) which rule out grammatically incorrect responses; or
- "trick" questions that are written around an insignificant detail rather than meaningful fact.

Other common issues include inconsistent response length, similarities between the response, use of "none of the above" (NOTA) or "all of the above" (AOTA) options, and irrelevant distractors. These poor response choices in MCQs can lead to giving unnecessary cues to testwise (i.e., those who can correctly answer based on finding flaws or hints in the test items) students, which in turn reduces the accuracy of the assessment (Downing, 2005; Suskie, 2009; Case & Swanson, 2002). For example, inconsistent response length can be associated with the *long correct answer* error where the longest and most complex answer is usually the correct one (Case & Swanson, 2002). Similarly, NOTA or AOTA options and other absolute options like "always" or "never" provide clues to students. According to Case & Swanson (1998), "Use of 'none of the above' essentially turns the item into a true/false item…" (p. 25). Other clues such as grammatical links or word/phrase repetition between the stem and correct answer can artificially reduce test accuracy due to testwiseness.

Content gaps in previous assessments

In a 2011 systematic review of existing biostatistics assessment among medical researchers, Dr. Felicity Enders concluded, "This analysis shows a need for a new instrument to assess biostatistical competencies for medical researchers" (Enders, 2011, p. 4). Most notably, the instruments lacked sufficient validity evidence, and did not include some of the core competencies in public health and translational medicine. She also concluded that previous instruments failed to ask questions about certain common statistical techniques of which the

exclusion of repeated-measures was, “Perhaps most egregious of these omissions…” (p. 4). Furthermore, she claimed that instruments aimed at the practicing physician population (Berwick et al., 1981, Windish et al., 2007, & Novak et al., 2006, among others) did not focus enough on, “Whether the appropriate method has been used or…[on] interpreting statistical results” (p. 4). The Enders review suggested filling these content gaps by including both clinical and translational science (CTS) and public health (PH) as additional sources for BEK assessment topics because they are the, “two primary disciplines which train medical researchers” (p. 1). The majority of existing BEK assessments are home-grown instruments (Windish et al., 2007), and they are also usually constructed from the commonly used statistics in medical journals. Use of core competencies for developing BEK assessments has, until recently, extended only as far as stating the topics’ relationships with the ACGME core competencies (e.g. Green, 2001; Morreale et al., 2012; Rao, 2008). Since the ACGME competencies are vaguely written when discussing BEK, the CTS and PH core competencies allow prospective assessment writers with key BEK concepts and skills advocated by closely related disciplines.

Although Enders (2011) provided compelling evidence and solutions for the existing BEK instruments’ content gaps, she did not address the psychometric or item construction properties of the assessments. However, she agreed that those who constructed the instruments were talented statisticians but neither psychometricians nor measurement experts (F. Enders, personal communication, April 30, 2013).

Section Three: Review of Objective Test Development and Item Response Theory

The following section presents a brief introduction to Item Response Theory (IRT) from which the present study derives its conceptual framework and methodology. It seeks to briefly highlight its definition and use as well as its strengths and weaknesses compared to Classical

Test Theory (CTT). First, CTT will be briefly reviewed including its definitions and assumptions, models for reliability and validity, and item analysis for objective tests. Second, the fundamentals of IRT are described, which will include definitions and assumptions, IRT item parameters and the Item Characteristic Curve (ICC), Item and Test Information Functions, and how three common IRT models differ in analyzing dichotomous test data. This section concludes with an empirical comparison of CTT and IRT as well as the theoretical benefits of the latter approach. This section offers conclusions regarding the literature that underlie the focus of this research study.

Introduction to measurement: classical test theory approach

Several fundamentals of the measurement process must be reviewed prior to describing Classical Test Theory (CTT) and Item Response Theory (IRT) in detail. Measurement, broadly, involves assigning numeric values to objects or events in an effort to make meaning and understanding of a particular variable (de Ayala, 2009). In educational and psychological testing, a number of individual measurement items are combined to create a single, composite instrument, which is referred to as a *scale* (DeVellis, 2012). Responses to individual items on these scales are combined to create a single score meant to measure theoretical or *latent* variables or *traits*. A latent trait is one that cannot be easily observed directly, and is therefore estimated by an individual's observed score on the scale. These traits can be personality or psychological such as anxiety and depression or knowledge and achievement traits like BEK. Both IRT and CTT view these latent traits as continuous, which means an individual's trait score on it could be anywhere from zero to infinity (Devellis, 2012). For example, the physicians who took the BEK assessments described thus far were not considered either "knowledgeable" or "not knowledgeable"; rather, they were graded on a continuum from "very little knowledge" to "a

great deal of knowledge." How IRT and CTT differ in their approach to placing individuals on these continuums will be the focus of the upcoming section.

Any measurement, regardless of using CTT or IRT, faces concerns over reliability, validity, and generalizability. Broadly speaking, reliability refers to an instrument's *consistency* at estimating someone's score on the latent variable of interest (DeVellis, 2012). A highly reliable instrument will produce very similar scores over multiple administrations whereas scores could vary considerably across multiple administrations on an instrument with poor reliability. If an individual takes Scale A, for example, and receives a raw score of a 10, then they should score similarly on repeated administrations of the instrument. Likewise, the score of a 10 should consistently reflect the same magnitude of the latent variable the scale is meant to measure. Finally, each item within the scale ought to be an independent manifest of said latent variable. CTT and IRT differ in how reliability is assessed, but its impact on researchers' confidence in a particular measurement cannot be overstated (de Ayala, 2009).

Reliability is considered to be necessary but not sufficient for attaining validity, which is the degree to which the instrument *accurately* measures the latent variable of interest (DeVellis, 2012). For instance, a reliable instrument will consistently estimate an individual's latent variable score; however, it could consistently estimate the *wrong* latent variable score if it lacked validity. Although there are many types of validity evidence, the three most relevant to the current study are content validity, construct validity, and concurrent validity.

Content validity concerns how closely the content covered in a test match the content that *should* be included in the test (Furr & Bacharach, 2008). In other words, the test should include content relevant to all major facets of the latent variable it was built to measure. This type of validity evidence is typically gathered *before* the test is administered for the first time. In

particular, evidence for content validity is usually attained through consultation with experts on the construct of interest. For example, the researchers creating a BEK assessment may use clinicians, biostatisticians, and/or epidemiologists to critique the tests prior to administration.

Finally, construct validity concerns the extent to which the items on a scale behave the way they ought to if they were measuring the intended construct (Devellis, 2012). Evidence for construct validity can be assessed through multiple means; however, the present study used known-groups validity (Devellis, 2012) as the primary indicator of construct validity. Known-groups validity evaluates the sensitivity of the new instrument in differentiating among groups of individuals known to differ on the latent trait being measured. The present study used comparisons among demographic groups to assess this component of validity evidence.

Formally, CTT is the approach to measurement that is also known as *true-score theory* or the *classical measurement model*. It views an individual's trait score on the latent variable (i.e. their fixed location on the variable of interest) as a function of their observed score on a measurement scale plus measurement error (de Ayala, 2009). CTT assumes that these error values are (a) randomly dispersed among the scale's individual items; (b) not related to one another; and (c) not related to the true score on the latent variable. Essentially, the measurement error associated with any single item must not be dependent upon either the error of another item or the latent trait of interest. The unit of analysis when CTT is used is the *scale* rather than the *items*, which means that the respondent's observed score on the entire instrument is the focus (de Ayala, 2009). Consequently, this focus has implications for how reliability is dealt with according to CTT.

Reliability, according to CTT, is the degree to which differences in respondents' observed scores are consistent with those on their true or trait scores (Furr & Bacharach, 2008). In

objective testing, reliability is usually calculated through multiple administrations to the same group of examinees (test-retest reliability), or by how close responses to similar items relate to one another (internal consistency reliability) (DeVellis, 2012). Test-retest reliability involves giving the same instrument to the same group of students at two different time points. A reliable instrument will result in scores that are highly correlated to one another across time points while an instrument with low validity will not show such a relationship (Furr & Bacharach, 2008).

The second common method for gathering reliability evidence is through internal consistency reliability. In fact, one review of in *Psychological Assessment* stated, "The single most widely used method for item selection in scale development is some form of internal consistency analysis" (Clark & Watson, 1995, p. 313). Internal consistency generally refers to the degree to which the items on a scale relate to one another as well as the scale altogether. Statistician Lee Cronbach created the Cronbach coefficient alpha in 1951 as an indicator for estimating internal consistency of an instrument (as cited in de Ayala, 2008). Cronbach's alpha ranges from 0 to 1, and a higher alpha indicates a more reliable instrument. This approach to reliability was used by Fritsche et al. (2002) as well as Windish et al. (2007) in the instruments they developed.

Although only two of the current BEK assessments performed formal reliability or validity analyses on their instruments, each conducted some degree of item analysis. Item analysis is most applicable to knowledge and achievement tests because it uses a dichotomous (i.e. correct or incorrect) model to describe the response patterns of each test item. According to CTT, item analysis involves two essential calculations: item difficulty and item discrimination (Academic Technology Services, 2009).

To calculate item difficulty, the researcher looks at each item and calculates the proportion of students who got that particular item correct. Ideally, a well-tuned test item will have a difficulty between 0.3 and 0.5 (medium difficulty) while those with fewer than 20% or greater than 80% are considered too hard or too easy, respectively.

Item discrimination, according to CTT, is calculated by correlating the responses to each item with the total score on the test (Academic Testing Services, 2009). A poorly discriminating item will show a correlation near 0.0, which indicates correct responses to that item have no relationship with someone's overall test score. Moreover, an item may show a *negative* correlation with overall test score, which indicates a correct answer is inversely related to a high overall score. Conversely, an adequately discriminating item will have a statistically significant, positive relationship with overall test score, which means that correct responses to that item is associated with higher overall test scores.

A third possible approach to item analysis is called a distractor analysis. Each of the incorrect answers (i.e. distractors) for an item are analyzed for the frequency at which they are chosen by both the top 25% and bottom 25% of examinees (Wise, n.d). Distractor analysis is used to identify weak or poorly performing response options that may be impacting an item's overall difficulty or discrimination. This approach is usually used in conjunction with difficulty and discrimination indexes in order to weed out potentially problematic items; however, none of the existing BEK assessments' authors made explicit reference to distractor analysis in their studies.

The basics of item response theory

Item Response Theory is hardly a new concept. Indeed, IRT splintered from the more common Classical Test Theory (CTT) back in the 1950s as described in Frederick Lord's 1952

monograph *A Theory of Test Scores* (Lord, 1952), and more firmly established by Lord and Novick (1968). Since that time, IRT has been called, "The way of thinking of test construction as the way of the future" (Wainer, 1989, p. 191). Nearly twenty-five years after that statement was made IRT continues to be the dominant and preferred method for test construction due to its appealing advantages over CTT and traditional item analysis (De Champlain, 2010; Stage, 1998; Waller, Ostini, Marlow, McCaffery, & Zimet, 2013). Despite its widespread use, IRT remains a mystery to many as series editor David A. Kenny wrote in the opening editorial of R.J. de Ayala's *The Theory and Practice of Item Response Theory* (2009), "One could make a case that item response theory (IRT) is the most important statistical method about which most of us know little or nothing" (as cited in de Ayala, 2009, p. vi). The following section aims to explore the fundamentals of IRT, and why it is seen as advantageous versus CTT.

At its core, Item Response Theory postulates that an individual's response to a test item is a function of their position on a continuous latent trait denoted by the Greek letter "θ" (theta) (DeMars, 2010). The models used in IRT view this relationship in terms of probability (i.e. probability of answering correctly) using a similar procedure as logistic regression analysis (de Ayala, 2009). The word "theory" is sometimes misunderstood, but de Ayala (2009) clarified the term stating:

> "IRT is, in effect, a system of models that defines one way of establishing the correspondence between latent variables and their manifestations. It is not a theory in the traditional sense because it does not explain why a person provides a particular response to an item or how the person decides what to answer (Falmagne, 1989)" (as cited in de Ayala, 2009, p. 4).

For simplicity, this discussion will only focus on three major IRT methods for modeling dichotomous (i.e. correct or incorrect) items although more complicated models have been created to examine polytomous items (i.e. multiple categories) such as Likert-type scales (DeMars, 2010).

Assumptions of item response theory and its item parameters

Item Response Theory has two primary tenets: (1) the test data must contain a single dimension, and (2) the data must be locally independent (Waller et al., 2013). Unidimensionality refers to the requirement that only a single latent trait, θ, is measured by the items on a test (Hays, Morales, & Reise, 2000). Although this assumption is considered key to the three logistic IRT models presented in this chapter, the literature notes it is likely that there will inevitably be some degree of violation in any given test environment (de Ayala, 2009; DeMars, 2010). de Ayala equated it to the homogeneity of variance assumption to which an analysis of variance is robust when minor violations are committed. Furthermore, two related content areas (e.g. biostatistics and epidemiology), when included in equal proportions on an exam, can easily be mathematically unidimensional thereby not violating this key assumption (DeMars, 2010). The single dimension simply becomes a hybrid of the two content areas. Hays and colleagues (2000) asserted that "essential unidimensionality" (p. 9) is recognized as acceptably satisfying the assumption. However, caution must be taken to avoid serious violations of this assumption since, "Violating this assumption could bias several item and ability parameter estimations" (Yu, Popp, Digangi, & Jannasch-pennell, 2007, p. 1).

When the assumption of local independence is met, any two items will be unrelated with one another after controlling for θ (DeMars, 2010). Local independence is usually met if a test is unidimensional (Hays et al., 2000; Waller et al., 2013); however, DeMars (2010) claimed:

> "Two items that violate local independence may not be enough to form another dimension…Local dependency may be a concern when one item builds on the answer to a previous item, or when items are grouped around reading a passage or a common scenario…" (p. 49).

To be sure to avoid violating local independence, it is advised that test writers create a separate passage, example, etc. for each item.

IRT uses three item parameters within its system of models, namely, (1) item discrimination, (2) item location (henceforth called "difficulty"), and (3) pseudo-guessing, which are denoted by the letters "a", "b", and "c", respectively (de Ayala, 2009). Each of these parameters define separate characteristics of the Item Characteristic Curve (ICC) (Stage, 1998). An ICC is a graphical representation of the probability of correctly responding to an item across an array of θ levels (Figure 2.2). These trait levels (θ) are measured on a continuum along the horizontal axis of an ICC with a mean of 0.0 and standard deviation of 1.0 exactly as a z-score distribution. For example, an individual with an average trait level would be located at θ=0.0, and the majority of individuals will fall between θ= -3.0 and θ = 3.0 (DeMars, 2010). Item difficulty ("b") represents the point of inflection on the ICC curve (where the slope changes direction). de Ayala (2009) uses the term *location* when referring to difficulty because the parameter is written in terms of a specific location on θ. Using the simplest IRT model, difficulty is the location where a person has a 50% probability of giving the correct answer (DeMars, 2010). That is, an item with a difficulty of 0.0, for example, indicates an individual with an average trait level (θ=0.0) would have a 50% probability of correctly answering that item. A difficulty between "b" = -2.0 and "b" = 2.0 is considered to be the acceptable range for items that will be neither too easy nor too difficult (de Ayala, 2009). The relationship between item

difficulty and θ is illustrated in the figure below, which shows three hypothetical items with difficulties of "b" = -1.0, 0.0, and 1.0. An item with average difficulty will be located at or near 0.0 (Item One) while a more difficult or less difficult item will be located above 0.0 or below 0.0, respectively.

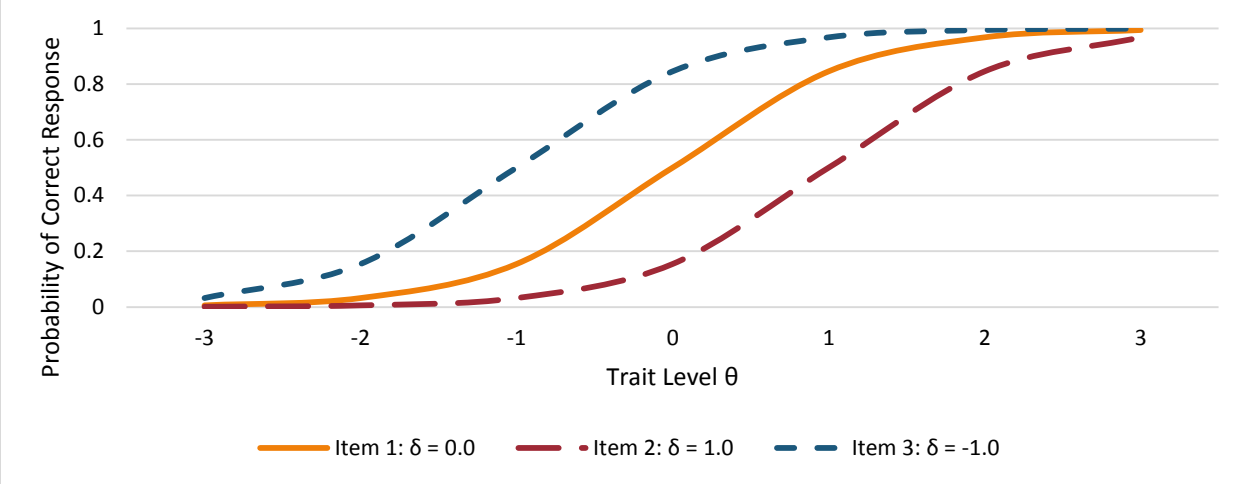


Figure 2.2. Example Item Response Function for Three Hypothetical Items with Difficulties of "b" or "δ" = -1.0, 0.0, and 1.0

Item discrimination ("a") is also vital to an ICC as it forms the *slope* of the line. Just like with CTT item analysis, the IRT perspective defines discrimination as the ability for an item to differentiate between individuals at high or low levels of ability (θ, in the case of IRT) (de Ayala, 2009). DeMars (2010) advises that an item discrimination typically falls between 0 and 3 although the value can theoretically range between -∞ and ∞. Also, a negative discrimination parameter, just as in CTT, indicates an item on which individuals with a higher ability level have a *lower* probability of answering correctly; therefore, these items ought to be removed for poor performance. Figure 2.3 below provides three hypothetical items with an equal level of difficulty ("b" = 0.0), but each has a different discrimination ability of 1.0 (Item One), 1.5 (Item Two), and 0.5 (Item Three). An item will discriminate most accurately at the point where the slope is the steepest (DeMars, 2010). On the figure, item one is of average discrimination whereas the

steeper slope on item two indicates *higher* discrimination. The flat, gradual slope of item three is characteristic of an item that discriminates poorly because the probability of a correct response remains relatively unchanged across a wide range of θ (de Ayala, 2009).

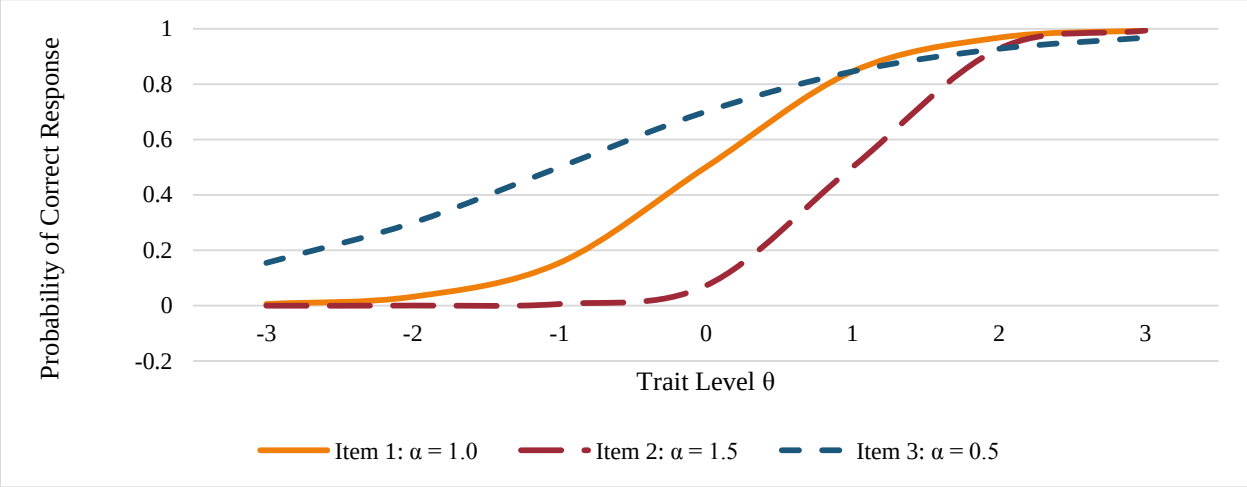


Figure 2.3. Example Item Response Function for Three Hypothetical Items with Discriminations of "a" or "α" = 1.0, 1.5, and 0.5

The third parameter that IRT estimates is known as the pseudo-guessing parameter, or ““c”.” Pseudo-guessing is the probability that someone with a very low level of θ will answer an item correctly given chance alone (de Ayala, 2009). The parameter got its name because well-written distractors are apt to pull individuals with lower ability levels towards selecting them; therefore, the realistic “c” value is usually lower than what would be expected by random chance (e.g. 25% for a four-option MCQ) (de Ayala, 2009). The pseudo-guessing parameter is represented as the lower-asymptote on an ICC, which is, “The value the function approaches as θ approaches negative infinity” (DeMars, 2010, p. 13). One noteworthy drawback of including this third parameter is its adverse effects on estimating an individual’s ability levels. Specifically, Wainer (1983) found that nonzero pseudo-guessing parameters *lower* the estimates of person location (as cited in de Ayala, 2009). Figure 2.4 illustrates this effect using three items with “b” = 0.0, “a” = 1.0, and “c” = 0.0, 0.20, and 0.30, respectively. As the guessing chance increases the

probability for a correct response at “b” increases considerably (from 50% when “c” = 0.0 to 65% when “c” = 0.30) despite keeping “b” set at a constant 0.0. Although a nonzero pseudo-guessing parameter may be appropriate for a given testing situation, steps must be taken to reduce this value as much as possible, which can be done primarily through well-written items and distractor options (de Ayala, 2009).

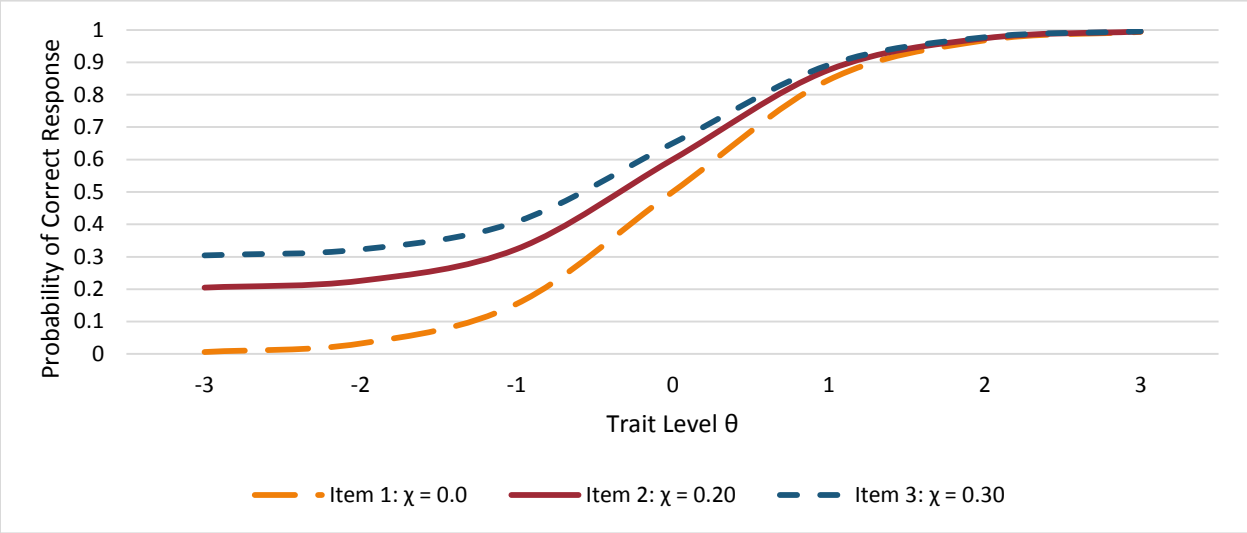


Figure 2.4: Example Item Response Function for Three Hypothetical Items with Pseudo-Guessing Parameters of "c" or "χ" = 0.0, 0.20, and 0.30

Three common IRT models for dichotomous data

The number of parameters an IRT analysis estimates is based on the type of model one chooses to apply. Three of the most common IRT models for dichotomous data, in order of complexity, are the Rasch model, 2-parameter logistic (2PL) model, and 3-parameter logistic (3PL) model (DeMars, 2009). The Rasch model is the simplest in that it only models an item’s difficulty value while keeping both discrimination and pseudo-guessing constant at 1.0 and 0.0, respectively (Figure 2.5a). This leads to all examinees with the same number of correct responses having the same θ just as the total correct score is used to estimate person ability in CTT (DeMars, 2010). Similarly, the CTT approach to item difficulty (i.e. proportion correct) is also

sufficient for item difficulty using a Rasch model. These two properties are unique to the Rasch model, and do not hold for either the 2PL or 3PL models. Additionally, fitting a Rasch model requires a marginally lower sample size than either a 2PL or 3PL model since it is only estimating a single parameter. de Ayala cites Lord (1983) who found that this model provided more stable parameter estimates (i.e. less error) than the 2PL and 3PL with samples sizes of 200 or fewer. The Rasch model is the considered a practical approach for most testing conditions due to these simpler estimation procedures and lower sample size requirements.

The 2PL model estimates both difficulty and discrimination while keeping only pseudo-guessing constant at 0.0 (Figure 2.5b). The primary benefit of the 2PL model versus a Rasch model is that the discrimination can vary among the items, so the researcher does not need to assume every item is equally discriminatory. The 2PL model is generally the most economical of the three models when considering the trade-off between required sample size and accuracy of parameter estimation. Simulation studies found a 2PL model produced nearly as stable parameter estimates as a 3PL model, yet only required between 200 and 500 subjects to fit the model (Drasgow, 1989; Yen, 1981; Stone, 1992 as cited in de Ayala, 2009).

Finally, the 3PL model adds an estimation for pseudo-guessing in addition to difficulty and discrimination (Figure 2.5c). The benefit of a 3PL model is that it takes into account non-random guessing, which is generally a more accurate representation of a real-world testing situation. DeMars (2010) stated, “Among the dichotomous models, the 3PL model is the most common choice for multiple choice items because it seems reasonable to assume that low-ability examinees have some non-zero probability of choosing the correct answer” (p. 29). Freeing all three parameters to change rather than being held constant also lets researchers use the 3PL model to estimate items to be different be easier or harder dependent on ability level while also

accounting for non-zero guessing chance. As appealing as this benefit appears, a major weakness with the 3PL model is that it is the most complex of the three discussed thus far; consequently, it also requires substantially larger samples to estimate properly. de Ayala (2009) recommends samples sizes of at least 1000 to generate stable estimates of the pseudo-guessing parameter. Moreover, DeMars wrote that a 2PL model will likely be useful when high quality distractors are used to minimalize the chances for guessing. Overall, the 3PL model is the ideal representation for MCQ items, but it may not always be feasible to use in practice.

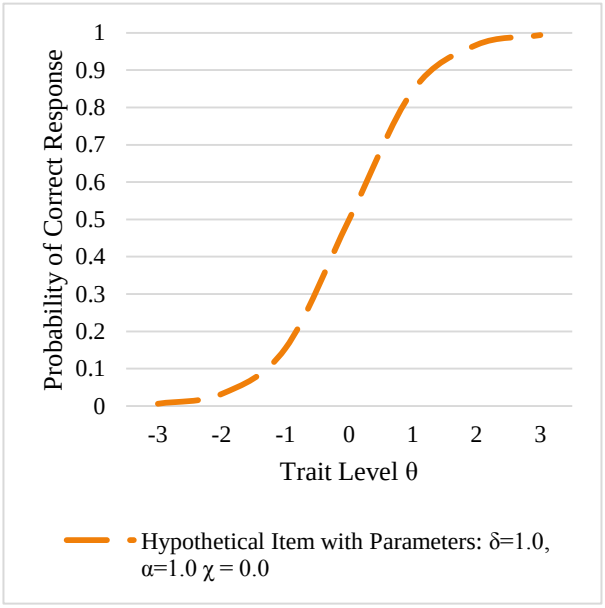


Figure 2.5a. Example ICC Using a Rasch Model

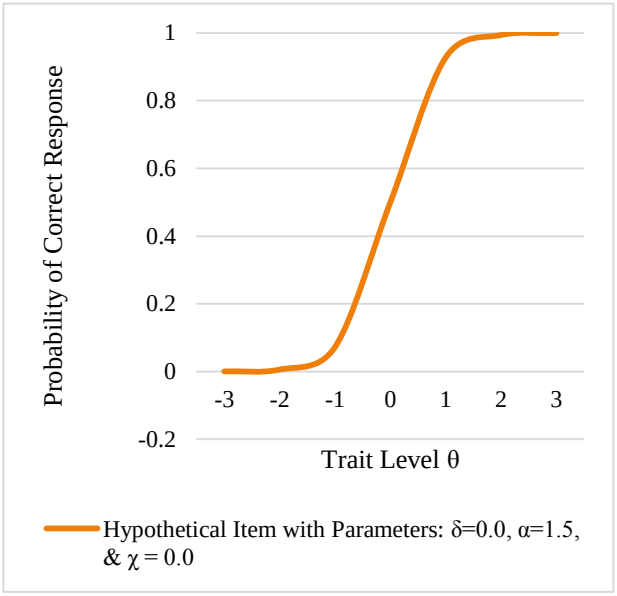


Figure 2.5b. Example ICC Using a 2PL IRT Model

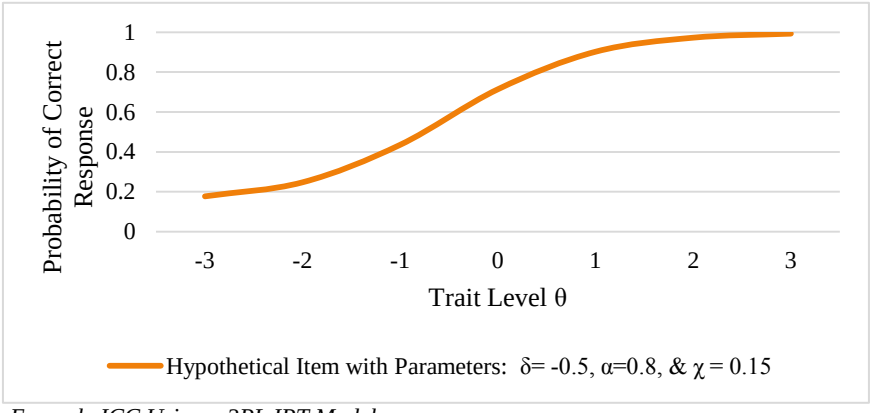


Figure 2.5c. Example ICC Using a 3PL IRT Model

Choice of model depends on a number of practical testing issues such as sample size, instrument characteristics such as length and administration, likelihood of meeting assumptions, and other external forces (de Ayala, 2009). The Rasch model has been shown to produce fairly stable, robust estimates of item parameters even when assumptions are violated; however, the restrictiveness of the model make it less appealing to some. On the other hand, the 3PL model, while the most thorough, requires sample sizes that are many times unreasonable for applied research. Adding the pseudo-guessing parameter can have a detrimental impact on the study if

not handled properly (de Ayala, 2009). For this practical reason, the 2PL model provides sufficiently accurate estimation as de Ayala wrote:

> "It is the validity of the person location estimates that is paramount…if convincing validity evidence can be accrued for person location estimates using a particular model…then it would seem that the above arguments [on which model to use], although interesting in their own right, are somewhat irrelevant" (p. 154).

In other words, researchers ought to be mindful of *Occam's razor*, which suggests the simplest explanation is usually the correct one. The goal is to find the model that best fits the data and allows for the most accurate estimate of trait scores, and it is not to always fit the most complex model.

Reliability in IRT: item and test information functions

Recall that, according to CTT, reliability was the consistency of the observed score and the true score on the latent variable of interest. Also, a single reliability coefficient is calculated at the *test* level for each sample, and this reliability value cannot be separated from the individuals (Stage, 1998). From an IRT perspective, the concept of reliability is known as *item* and *test information* or, the degree to which the researcher can be certain of a person's location along θ. For each item, the amount of information is proportionate to the standard error of estimate (SEE) for each possible θ location (de Ayala, 2009). A smaller SEE indicates a stronger certainty in the estimate of θ and therefore more information about individuals with that particular θ value. By rule, an item provides its highest amount of information near its difficulty value ("b") because there is the least amount of variability (error) near this value (DeMars, 2010). Similarly, an item with a high discrimination value will provide a large amount of information over a short range of θ whereas the flatter line of a poorly discriminating item will provide less

information over a lengthier range of θ values (DeMars, 2010). This relationship is graphically displayed on the example item information function in Figure 2.6. Both items one and two have a difficulty of 0.0; however, item two has a steeper slope than item one ("a"=1.5), thus, it has substantially higher information at 0.0 than item one.

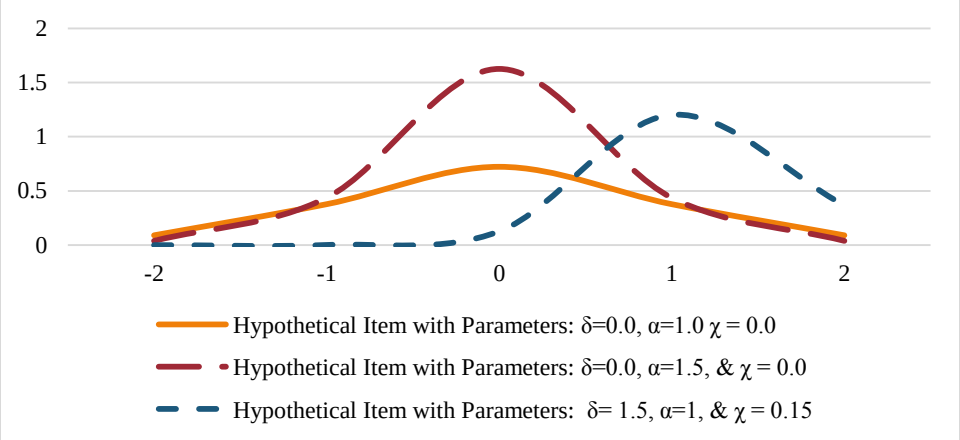


Figure 2.6. Example Item Information Curve for Three Hypothetical Items

The most important function of item information versus CTT understanding of reliability is that item information is independent of both other items *and* the sample from which it is calculated (Stage, 1998; DeMars, 2010; de Ayala, 2009, Furr & Bacharach, 2008). This property allows for items to be broken apart, rearranged, and reassembled into new test versions without losing their accuracy or consistency in estimating person trait levels. Moreover, total test information is calculated by summing the item information values for each item in a test; therefore, test information can be easily recalculated according to the items chosen for a particular form. This leads to a test being having stronger psychometric properties for some individuals and weaker ones for others (Furr & Bacharach, 2008). For example, a test geared towards novice students may want to be more accurate at determining differences among θ levels that are between -1 and 1, so specific items can be chosen which provide the most information

along this range. On the other hand, a high-stakes test among a group of high ability students such as a very selective scholarship opportunity, would want to have more information on the *upper* end of the θ spectrum to better choose the most qualified student. Following this process, IRT allows for educators to pull items that will give them the most accurate and reliable measures of student performance at a predetermined range of ability. Figure 2.7 shows the test information function for same three items from Figure 2.6. This figure suggests that the three-item test from Figure 6 provides the most information at ability levels that are slightly above average through about θ = 1.0. The test provides relatively little information for ability levels either below θ = -1.0 or above θ = 1.0.

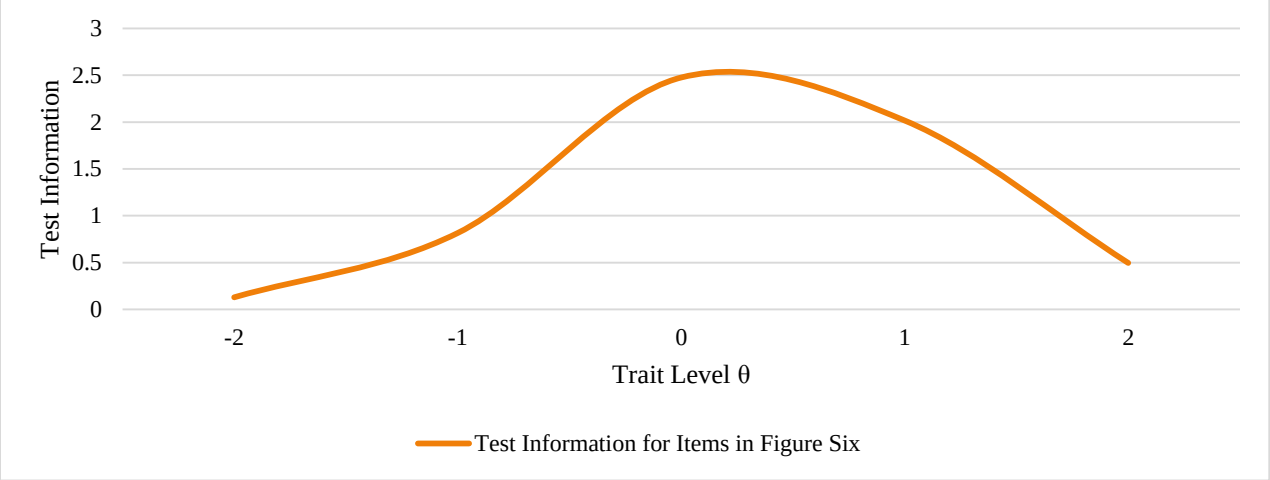


Figure 2.7. Test Information Function for Items in Figure 2.6

Primary strengths of IRT versus CTT in test construction

The single *most important* distinction between IRT and CTT is that IRT item parameters carry the property of sample invariance (de Ayala, 2009; DeMars, 2010; Stage, 1998; Waller et al., 2013). Invariance means that the parameters estimated through IRT may be taken independently of the sample or population from which they were derived. By comparison,

traditional CTT item analysis may show the same item to be far too difficult when given to a group of low-ability students but far too easy when administered to high-ability students (DeMars, 2010). Invariance allows the difficulties from one population to be placed on the same metric as those from another population within a linear transformation, which DeMars describes as, "the *b*'s [difficulties] from one population are multiplied/divided by a constant and another constant is added or subtracted" (p. 8). The same property is also true for the IRT discrimination parameter whereas the correlation used for calculating CTT discrimination index values is dependent upon the item's difficulty in the sample population.

It has already been shown that IRT places both trait parameters and item parameters on the same metric, which eases the interpretation for estimating an individual's item response probability given their ability level (Hays et al., 2000). Hays and colleagues provided the example that if an individual's trait level exceeds the item's difficulty level, then that person is very likely to answer the item correctly. The same interpretation cannot be made using CTT methods because both item and person characteristics are wrapped into a single difficulty value.

The CTT true-score model is assumed to be true, yet it cannot be tested or disproven because both trait scores and error scores are unknown quantities (de Ayala, 2009). In contrast, IRT provides researchers with a chance to assess model fit. This process compares both item and person characteristics that are predicted by the model to those observed in the dataset. Of course, this is only advantageous if the IRT model actually fits the data in question, otherwise using IRT provides no measurable benefit (Xu & Stone, 2011).

However advantageous IRT appears over CTT in theory, the empirical evidence consistently shows estimates from both methods to be quite similar to one another. Xu and Stone

(2011) conducted a simulation study to compare summated scores (CTT) and trait estimates (IRT), and concluded that the type of score had no meaningful effect on their predicted outcome.

Fan (1998) tested the difference between IRT and CTT using twenty random samples of 1,000 11^{th} graders' Texas Assessment of Academic Skills test. He looked both at the person statistics and item statistics using both approaches. He used standardized *T* scores, a common measure in CTT that transforms summated scores into a standard distribution with a mean of 50 and standard deviation of 10, and compared them to IRT θ scores. Fan found these values to correlate to one another greater than $r = 0.96$ for all comparisons, which indicates a very high level of agreement between the two measures of person ability. Further, when traditional CTT difficulty values were compared to IRT item location ("b"), his study showed the two approaches to be a near perfect match when comparing a Rasch model to the CTT statistics. The comparison of item discrimination estimates, Fan concluded, "May yield noticeable discrepancies with regard to which items have more discrimination power…" (Fan, 1998, p. 373). These results were again supported that same year by Stage (1998) who found the two approaches yielded similar results in 13 of the 20 items tested in the study.

At first glance, the empirical evidence comparing IRT to CTT appears at odds with the substantial theoretical differences illustrated in the preceding section. Indeed, there has been consistent literature to support a close relationship between the two approaches in both person and item estimates (e.g. Fan, 1998; Hays et al., 2000; Stage, 1998; Xu & Stone, 2011). On the other hand, as Fan (1998) put it, "…as the cornerstone of IRT, the importance of the invariance property of IRT model parameters cannot be overstated, because, without this crucial property, the complexity of IRT models can hardly be justified on either theoretical or practical grounds" (p. 360). The key distinction is that CTT and IRT offer comparable item and person estimates on

a per-sample basis; however, the theoretical framework from which each operates prevents the estimates made based on CTT's true-score theory from translating to new samples without losing their psychometric integrity. The invariance characteristic of IRT along with the item-specific reliability information make it an undoubtedly attractive choice for researchers in need of a highly adaptive instrument.

Chapter Summary

Assessment of biostatistical and clinical epidemiologic knowledge (BEK) among graduate medical professionals is an area in need of new assessment tools (F. Enders, personal communication, April 30, 2013). Previous research on medical residents' BEK extending back to Weiss and Samet (1980) consistently concludes that there is a generally low and variable level of knowledgebase within this population despite an equally-evidenced increase in the use of statistics over the same time period (e.g. Horton & Switzer, 2005; Reed er al., 2003; Windish, Huot, & Green, 2007). It has been hypothesized that this gap in knowledge begins with inadequate instructional hours devoted to these topics in medical schools (Looney et al., 1998; Sahai, 1999). This trend continues into graduate medical education where these topics are usually delivered in journal clubs run by the students, or through stand-alone courses on Evidence Based Medicine (EBM) (Fritsche et al., 2002; M. L. Green, 2001). Although BEK have been seen as essential to practice for many years (Rao, 2008), and are established among the Accreditation Council for Graduate Medical Education's (ACGME) core program standards (ACMGE, 2012), the prevalence of these topics among established programs has been as little as 30% in some areas of the country (Cheatham, 2000).

A number of attempts to develop assessments for this challenging population have been developed over the past few decades (e.g. Berwick et al., 1981; Enders, 2011; Fritsche et al.,

2002; Windish, 2011), yet there has been little formal psychometric treatment for these instruments outside of their initial development. Additionally, each of these assessments contains gaps in relevant content per Enders' (2011) systematic review as well as item common construction flaws such as item dependencies and unfocused stems. Further, none of these previous assessments were developed from an Item Response Theory (IRT) perspective, which had been experiencing an exponential growth in popularity across the same time period up through present day (Clark & Watson, 1995; De Champlain, 2010; Stage, 1998).

Unlike Classical Test Theory (CTT) on which all of the existing BEK assessments were based, IRT produces estimates of person ability (denoted by "θ") as well as item difficulty and discrimination that are invariant across samples. Moreover, IRT offers test-writers and educators item-specific measures of reliability to create custom test forms aimed at accurately measuring the ability level most desirable for their needs. It is upon this theoretical framework that the current study rests.

Specifically, the present study sought to develop the Biostatistics and Clinical Epidemiology Skills assessment (BACES) by leveraging the power of Item Response Theory to create a new, dynamic biostatistics and clinical epidemiology knowledge assessment for graduate medical professionals. The study aimed to address the following research objectives:

1. Establish content validity evidence of the BACES assessment
2. Examine the model fit of the BACES items to a 1PL/Rasch, 2PL, and 3PL IRT model
    a. Test for violations of essential unidimensionality and local independence
    b. Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment

c. Analyze the quality of item distractors on the BACES assessment
d. Analyze the total item and test information produced from the BACES instrument

3. Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.

The next chapter will introduce the methodology used to address these objectives. Specifically, the sampling methodology, test construction process, test administration, and statistical analyses will be presented.

Chapter Three: Methodology

Review of the Problem

Chapter Two illustrated the fundamental problem on which the current study is based. The dominance of Evidence Based Medicine (EBM) in Graduate Medical Education (GME) over the past twenty-five years has placed a premium importance on educating residents to be adept at translating medical evidence into clinical decision making (Hatala & Guyatt, 2002). Knowledge of biostatistics and clinical epidemiology is an essential component to comprehending the medical evidence (Sahai, 1999), yet studies from the past several decades have shown a consistently low, variable knowledgebase among graduate medical students (Berwick et al., 1981; Novack et al., 2006; Weiss & Samet, 1980; Windish et al., 2007). Conversely, reviews of top tier medical journals over the same time period have shown a steady *increase* in the frequency and complexity of statistical methods (Horton & Switzer, 2005; Reed, Salen, & Bagher, 2003; Weiss et al., 1980; Windish et al., 2007).

To address the growing need for adequate training, many EBM curricula now include content dedicated for biostatistics and/or clinical epidemiologic research methods. Unfortunately, the type, rigor, and length of these courses differ significantly among curricula as do the qualifications of course instructors (e.g. resident versus faculty led) (M. L. Green, 2001; M. Green, 1999). The GME learning environment and variability in baseline training has made assessment of these skills difficult (Hatala & Guyatt, 2002).

Berwick et al., (1981), Fritsche et al. (2002), and Windish et al. (2007), among others have all created instruments to assess the GME population. Unfortunately, the formal psychometric treatment for these instruments remains scarce outside of their initial development (Enders, 2011). Each instrument also carries several content gaps (Enders, 2011), and common

item writing flaws such as item dependencies and unfocused stems. Moreover, each of these instruments was developed from a Classical Test Theory (CTT) perspective, which does not allow the test items to be broken-up and reorganized to meet specific educational needs without damaging the instrument's reliability. Item Response Theory (IRT), by contrast, offers educators item and person ability parameters that are independent of the sample from which they are estimated. The invariance trait of IRT gives GME educators the freedom to choose specific, relevant biostatistics and clinical epidemiology assessment topics, and administer them to their own residents while maintaining the item's difficulty, discrimination, and ability estimates. Use of IRT in developing a new, flexible assessment for the unique GME population addresses the salient problem of *how do educators effectively prepare and assess physicians in biostatistics and clinical epidemiology?*

Study Purpose and Objectives

The purpose of the present study is to establish preliminary item characteristics and validity evidence for the Biostatics and Clinical Epidemiology Skills (BACES) assessment. The study aimed to leverage the power of Item Response Theory (IRT) to create a new, adaptive biostatistics and clinical epidemiology knowledge (BEK) assessment for graduate medical professionals. The following chapter will detail the methodology employed to meet these three research objectives:

1. Establish content validity evidence of the BACES assessment
2. Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model
    a. Test for violations of essential unidimensionality and local independence
    b. Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment

c. Analyze the quality of item distractors on the BACES assessment
d. Analyze the total item and test information produced from the BACES instrument
e. Compare person and item location estimates from IRT models to those of traditional CTT indices.

3. Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.

The first section of this chapter will outline both the population of interest as well as the sampling procedures for the current study. Next, the *Instrumentation* section will detail the process by which the BACES assessment was constructed, and how validity evidence was collected in the study. The remaining portion of the chapter will detail the study *Procedure*. A description of the data collection procedures will include detail on the collaborating sample sites, the mode(s) of administration, incentives for participation, and software used for data collection and analysis. Lastly, the statistical methods will be explained per each study objective.

Participants

Ethical considerations

Ethical approval was obtained from both the University of Tennessee Institutional Review Board (IRB) and the University of Tennessee Graduate School of Medicine IRB. Written consent from each participating site was submitted with the IRB documents, and copies of IRB approval were sent to each site prior to administering the BACES instrument. An informed consent document was also included for each study participant.

Study population and inclusion criteria

The BACES instrument was developed for the medical residents and sub-specialty fellows GME population. A cross-sectional study by Brotherton and Etzel (2012) catalogued the demographic characteristics of 8,712 of the 9,111 training programs across the United States. They found that this population is approximately 46% female and predominantly White (59%) with Asian and Black representing 28% and 6% of current residents, respectively (Brotherton & Etzel, 2012). The current study was conducted in the East South Central Region (Alabama, Kentucky, Mississippi, and Tennessee), which reported 28 residents per 100,000 population. In comparison, the highest concentrations of trainees (65 residents per 100,000 population) are located in New England (i.e. Connecticut, Massachusetts, New Hampshire, Rhode Island, and Vermont), but the Middle Atlantic States (New Jersey, New York, and Pennsylvania) reported similar concentrations (64 residents per 100,000 population). Finally, while the majority of residents were U.S. citizens (37%), 26.8% and 21.7% were either Non-U.S. citizens or of unknown citizenship, respectively (Brotherton & Etzel, 2012). There were no specific age statistics for the population. All participants must be both over 18, and English language proficient.

Sampling procedure

The study used a multi-institutional, convenience sample procedure of the resident population within the University of Tennessee System (Total resident population of N=1033) (ACGME, 2013_b). Although the sample was non-randomized, the intent was to sample as broadly and heterogeneously as possible in order to obtain a representative mix of the resident ability level. Colleagues from three academic medical centers across the state of Tennessee (resident populations of n=683, 178, and 172) were asked to grant the researcher access to their

residents. Those who agreed were given a Memorandum of Understanding to sign for IRB approval, which details their institution's role as well as the use and ownership of the data.

Statistical Power and Sample Size Considerations

Edelen and Reeve (2007) remarked, "Although there are no definitive answers regarding sample size requirements, there are some general statements and guidelines…" (p. 8). Indeed, previous studies have noted minimum sample sizes that yielded stable parameter estimates for the Rasch, 2PL, and 3PL IRT models (Lord, 1983; Drasgow, 1989; Yen, 1981; Stone, 1992). The maximum sample size required for these three models is between 1000 and 2000 to reliably estimate the parameters of a 3PL model (de Ayala, 2009); however, the more feasible sample size of between 100 and 500 can be used to estimate the simpler Rasch or 2PL models (Lord, 1983; Drasgow, 1989). Regardless, the study attempted to sample as many participants as the data collection period allowed because both parameter and person estimates for all models have smaller standard error terms as sample size increases (Orlando & Reeve, 2007).

Instrumentation

The instrumentation section addresses three key components of the BACES assessment. First, biostatistics and clinical epidemiologic knowledge is operationalized in the context of the study as well as the sources from which the BACES content is selected. Second, each content area is appropriated a percent of BACES items using a test blueprint process (Suskie, 2009). The final component details the way in which individual items are constructed per best practices in item writing (i.e. Case & Swanson, 2002; Suskie, 2009).

BACES item construction

Content Selection and test blueprint

Biostatistics and clinical epidemiologic knowledge was defined in the context of the present study as the ability to correctly identify, interpret, and apply fundamental statistical and epidemiologic theory, commonly used statistical tests, and common epidemiologic research methods relevant to clinical practice. This operational definition was derived from the ACGME Core Competencies for Medical Knowledge and Practice-based learning and improvement (ACGME, 2013_a) as well as the body of literature on physician knowledge of biostatistics and clinical epidemiology reviewed in Chapter Two. Item content based on this definition was selected from four sources:

(1) Learning objectives from the biostatistics and clinical epidemiology curriculum taught at a southeastern regional academic medical center.

(2) Commonly used statistics in medical literature as cited in literature reviews (i.e. Horton & Switzer, 2005; Reed, Salen, & Bagher, 2003; Windish, Huot, & Green, 2007). Finally,

(3) Common content areas among existing assessment instruments

(4) Content gaps relating to clinical and translational science (CTS) and public health (PH) core competencies (Enders, 2011).

These four sources were used to generate a test blueprint (Suskie, 2009) for the BACES assessment, which ensured, to the extent possible, that the assessment covered the full domain of knowledge it intended to cover (Table 3.1). From the four sources above, five learning goals were established for the final BACES assessment:

(1) Apply the epidemiologic research design that will yield the strongest evidence for a given research scenario.

(2) Evaluate research findings for correct statistical methodology

(3) Critique research findings in terms of biases, reliability, and validity

(4) Use research findings to generate common measures of association in epidemiologic and medical research

(5) Integrate basic statistical concepts such as hypothesis testing, statistical power, confidence intervals, and scales of measurement into a medical research scenario.

The goals were written using a forward assessment approach (Fink, 2013), which focuses on skills or knowledge the student will use *after* the teaching and learning activities rather than *during* them. As most of the teaching of BEK in GME occurs within journal clubs or evidence-based medicine courses (Green, 2001), the learning goals were developed with attention towards critical appraisal of existing medical literature. Also, focusing the goals on interpreting the medical literature better aligned the BACES assessment with the evidence-based practice and medical knowledge ACGME core competencies.

Within each of these five learning goals, a number of individual topics (i.e. possible items) were generated using the four sources of content. The goals were weighted according to number of topics and allotted both a number of items and percent of final BACES instrument. Table 3.1 provides a description of (a) the learning goals on which the BACES assessment is developed, (b) the individual concepts or topics within each goal that were developed from the four sources listed above; and (c) the estimated number of items and percent of the instrument for each goal.

Table 3.1
Proposed Test Blueprint for BACES Assessment by Learning Goal

Assessment Learning Goal	Concepts/Skills	Number (%) of Items in Proposed Instrument
Apply the epidemiologic research design that will yield the strongest evidence for a given research scenario.	Case Control Cross-Sectional Clinical trial (Factorial and cross-over) Cohort Equivalency/Non-Inferiority	7 (23%)
Evaluate research findings for correct statistical methodology	*t*-test (independent and dependent) ANOVA (one-way and factorial) Chi-square Correlation Linear Regression Non-parametric test	8 (26%)
Critique research findings in terms of biases, reliability, and validity	Selection bias Information bias Threats to Internal validity Threats to External validity	4 (13%)
Use research findings to generate common measures of association in epidemiologic and medical research	AR/ARR* Odds Ratio /Relative Risk Ratio Logistic Regression Survival analysis NNT/NNH** Diagnostic testing*** Incidence Rates	8 (26%)
Integrate basic statistical concepts such as hypothesis testing, statistical power, confidence intervals, and scales of measurement into a medical research scenario.	95% confidence interval Power Type I/II error Scales of measurement	4 (13%)

*Attributable risk / absolute risk reduction
**Number needed to treat, number needed to harm
***Includes sensitivity, specificity, and predictive values

Methods for Pretesting the BACES Assessment

The remaining BACES assessment items were pretested with three methods. First, a group of sub-specialty fellows completed the assessment during an annual week-long educational workshop. Second, a group of residents completed the assessment during scheduled educational modules on biostatistics and epidemiologic research methods. These first two methods provided quantitative feedback on item performance to identify any substantial errors such as unclear stems or examples, poor distractors, or unclear instructions.

The third method for pretesting the BACES assessment was through a small-group, discussion-based feedback session with examinees. The examinees completed a number of BACES items in small groups during their scheduled education period, and then each group shared their group's response to the item as well as their reasoning for choosing that response. The intent of these discussions was to get more detailed information on *why* examinees were choosing response options, but also to allow for the researcher to ask follow-up questions regarding item clarity or errors. A small-group discussion approach gave an opportunity for the researcher (and assessment) to benefit from qualitative feedback on the items without taking any more of the residents' and fellows' time. Additional modifications based on these three pretesting methods were made before sending the BACES assessment to the content reviewers.

Methods for Writing BACES Items

BACES items were written using a multiple choice question (MCQ) format with four response options per question. It has been shown that the best MCQs in medical testing incorporate a clinical vignette within the item stem (Jozefowicz et al., 2002). In accordance with these findings and existing BEK assessment measures, the BACES items used clinical or literature-based vignettes to emphasize residents' *application* of BEK concepts rather than their simple memorization. The final BACES assessment contained 30 items (Appendix B).

The two key assumptions for Item Response Theory (IRT) introduced a number of special considerations. According to the *strict* unidimensionality assumption of IRT, items must measure only a single latent trait, θ (DeMars, 2010). This assumption necessitated that the BACES items avoid measuring unintended knowledge areas as much as possible. Using the vignette model, an item may unintentionally measure respondents' medical knowledge if the vignette is too specific such as including a condition few residents learn about. As an approach to minimize this effect, an upper-level surgery resident was consulted for a list of five broad topics that *all* residents learn during their first year of residency. These broad topic areas steered the researcher to write items that were relevant to the broadest audience possible while still being able to focus on using real, clinical examples.

Further complicating the item writing process was the conditional independence assumption. To satisfy the conditional independence assumption, items may not be linked to one another via content clues, shared examples, or other interdependencies (DeMars, 2010). Avoiding violations of conditional independence required that a *unique* vignette or example was used for each item, and that the vignette did not give a clue to the answer for another question. Finally, response sets were varied to reduce chances that respondents could guess the correct answer through process of elimination. Although each BACES item was given a unique vignette, these issues cannot be feasibly avoided altogether.

In addition to meeting assumptions, BACES items followed established guidelines for high quality MCQ items both in a broad context (Suskie, 2009), and in the health sciences, specifically (Brunnquell et al., 2011; Case & Swanson, 2002; S. Downing & Baranowski, 1995; S. M. Downing, 2005). The key guidelines from several of these authors are briefly summarized in Table 3.2 below.

Table 3.2
Summary of Best Practices for MCQ Writing

Author(s)	Area of Best Practice (Commonalities are in **Bold**)		
	General	Stems	Responses (Distractors)
Suskie (2009)	• Be concise • Define all terms • Avoid unnecessarily complex vocabulary • Avoid "interlocking" items	• Ask a complete question. • Avoid "which of the following" items. • Avoid common knowledge questions • Avoid negative stems • Avoid grammatical clues to the correct answer	• Not all questions need the same number of options.* • Order responses logically • Use vertical responses rather than horizontal • Make all options similar length. • Avoid "None of the above" and "All of the above" • The best distractors identify where students' thinking went wrong, and should be intrinsically possible or true statements
Brunnquell et al. (2011)		• Avoid negative stems • Avoid unfocused or vague stems • Avoid verbal associations between stem and answer	• Avoid "cues" such as "always," "never," "usually," etc. • Avoid "None of the above" and "All of the above" • Make all options similar length.
Case & Swanson (1998)	• Items should focus on important concepts only • Avoid trick questions • Assess application of knowledge rather than recall of facts • Avoid clues for testwise students	• Be clear and concise • Avoid "which of the following" or "Each of the following…except" items. • Avoid "hinging" (i.e. interlocking) items	• Distractors should be homogeneous. • Avoid options with two parts • Order responses logically • Make all options similar length • Distractors should be intrinsically possible or true statements

Methods for Establishing Validity Evidence for the BACES Assessment

Content validity of the BACES items

Furr and Bacharach (2008) note two specific threats to content validity: construct irrelevant content, and construct underrepresentation. Construct irrelevant content is introduced when the test includes items that are not relevant to the latent construct of interest. Construct underrepresentation, by contrast, occurs when the test does not cover a sufficiently broad range of the latent construct. One method for addressing construct irrelevant content was discussed in a previous section (i.e. choosing vignettes based on *common* conditions) although the primary strategy for addressing content validity was through expert review.

Expert review is the same approach taken by previous researchers in establishing content validity for their BEK assessments (Enders, 2011). The four-person group of experts included areas of expertise relevant to both the content and educational context of the BACES assessment. The item content was reviewed by a clinical pharmacist faculty member who is also a member of the exam-writing committee for the University of Tennessee College of Pharmacy, a senior general surgery resident, and an MD/DPh faculty member in the Public Health Department for their feedback on the medical applicability of the BACES items as well as the relevance of the concepts being tested to their residents' education. Secondarily, the assessment director for a private liberal arts college in central Minnesota reviewed the BACES items for their fidelity to quality assessment practice.

Each reviewer was given four documents: 1) a copy of the BACES items, 2) a detailed answer key with answer descriptions and continuum of option "correctness" (Appendix C), 3) a brief overview of the study, its purpose, objectives, and methods (Appendix D); and finally 4) a copy of the item review and scoring rubric described in a previous section (Appendix A). Although the appraisal of the items varied depending on the content specialty of the reviewer, there were no items that were candidates for removal. For clarification, two of the four reviewers were informally interviewed regarding their suggestions for improving the instrument, which led to a number of other improvements.

Construct Validity Evidence

Known-groups validity (Devellis, 2010) was the primary source for preliminary construct validity evidence. This method of validity involves, "Demonstrating that some scale can differentiate members of one group from one another based on their scale scores" (Devellis, 2012, p. 65). With regards to BEK, Windish et al. (2007) found male residents, those holding an advanced degree, and residents with past training in biostatistics were significantly associated with higher knowledge scores while successive years after medical school were associated with a significant *decline* in scores. Novack et al., (2006) found a similar relationship between years since medical school and BEK scores, but they also saw reading the methods section of a journal article and number of publications as significant predictors of higher scores. Since the BACES assessment targets only residents, known-groups validity comparisons were conducted using sex, degree, year of residency, and prior exposure to biostatistics or epidemiology as possible predictors.

Study Procedures

The following section describes the study's data collection and statistical procedures. The data collection subsection will describe the administration of the BACES assessment as well as the software used to collect and analyze responses. The majority of this section is devoted to describing the statistical procedures for the study both in general and by specific statistical procedures organized by research objective. A summary of the methods for each study objective can be found in Table 3.2.

Data collection procedures

The study utilized a multisite, cross-sectional survey approach to developing the BACES assessment. To recruit each sample site, the Designated Institutional Official (DIO), the individual responsible for GME administration at the institutional level, from each site was given a summary of the present study as well as several example BACES assessment items. Those who agreed to participate were able to grant access to individual residency programs to whom the BACES assessment was administered. Data collection occurred during either (a) journal club meetings, or (b) other scheduled didactic session using a paper-pencil, group administration format. Data collection took place over the course of approximately 30 days with a total of 10 different residency departments visited across the three sites.

Prior to each administration, residents received an informed consent page where they were given more information about the study including its voluntary nature and the risks and benefits of participation (Appendix E). They were given the BACES assessment as well as a brief set of demographic items similar to those used in the Windish et al.

(2007) study including sex, age, years of training, location of training, and any previous training in biostatistics, epidemiology, or evidence-based medicine.

The 30-item BACES assessment consistently took between twenty and approximately thirty minutes to complete across the 10 administrations. After the assessment was completed, the researcher used the remaining journal club or didactic session time to review the answers with the group. Each resident received a copy of the answers for each BACES item as well as the brief description for those answers in order to provide them immediate feedback on their performance. These answers were the same as the descriptions given to the expert reviewers, and it also included a scannable link to the researcher's series of online lecture materials as a small incentive for participation.

Software used for data collection and analysis

All responses were transcribed electronically using Remark OMR Software (Gravic, Inc.) and downloaded into Microsoft Excel 2013 (Microsoft Corporation, 2013) for initial recoding. Once recoded, all IRT CTT analyses were conducted using Xcalibre v4.2 (Guyer & Thompson, 2012). Distractor analyses, descriptive statistics, and validity analyses were conducted using IBM SPSS v.22 (SPSS Inc., Chicago IL, 2013). Additional details about these programs will be shared as needed in the upcoming subsections.

General statistical methodology: outliers, missing data, and demographic comparisons

General Methodology for Demographic and Perceived Knowledge Data

Descriptive statistics including cross-tabulations, frequency distributions, skewness and kurtosis statistics, measures of central tendency, and measures of variation were used to screen data for coding errors, missing data, and outlying values. The considerations for these data concerns were handled differently for the assessment items than for the demographic items.

For demographic and perceived knowledge items, responses to items with greater than 10% missing data were excluded from any inferential comparisons. Outlying values were defined as those that exceed a standardized *z*-score of the absolute value of $z = 3.29$ (Tabachnick & Fidell, 2013). Coding errors were judged on a case-by-case basis to determine, to whatever extent is possible, whether it was a client-side (participant) or researcher-side mistake. Coding errors were adjusted in the case of the latter or when the true response was clearly marked, but the value was otherwise set to missing.

General Methodology for Objective Assessment Data

Data were handled differently with the objective assessment items. Estimating IRT parameters is sensitive to missing data (de Ayala, 2009), which could be caused by omitted responses or speededness. Speededness refers to the inability for participants to reach items near the end due to time constraints (de Ayala, 2009). The data collection methods were specifically designed to mitigate this effect through randomly presenting the participants with one of two test forms – form "A" and form "B." These two parallel forms contained the same items; however, the first and last 15 items were swapped on

form "B," so that there would not be a systematically low response on the second half of the test. Although participants were not timed during the test, speededness was still guarded against because it was common for residents to unexpectedly stop taking the exam due to medical emergencies, being on-call, or being paged to a patient. With respect to omitted responses, previous BEK assessments, developed using CTT approaches, coded omitted (e.g. skipped) responses as incorrect responses (Fritsche et al., 2002; Novack et al., 2006; Windish et al., 2007); however, research indicates that this approach is not optimal when using an IRT approach. Specifically, de Ayala (2009) states, "Omits should not be treated as incorrect nor should they be ignored…However, using a fractional value of 0.5 in place of omitted values leads to improved person location estimation…" (p. 150). Following this recommendation, omitted responses were automatically imputed with simulated data as part of the calibration process.

Statistical Methods by Study Objective

The following subsection will describe the statistical methods for each study objective (summarized in Table 3.3). Objective one, "Establish content validity evidence of the BACES assessment," has already been addressed in a previous section. This subsection will begin with methods for checking statistical assumptions for IRT, and then continue with objective two, "Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model," and objective three, "Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons."

Table 3.3
Summary of Methods by Study Objectives

Study Objective	Data Source	Primary Methods
Establish content validity evidence of the BACES assessment	Four-person expert review panel and senior medical residents.	An expert in assessment, epidemiology, medicine, and clinical pharmacy reviewed the BACES items and provided feedback.
Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model	BACES assessment item responses.	Each model was fit and compared for the best-fitting model, which was assessed via chi-square goodness of fit index, and by comparing change in -2 Log Likelihood statistics between models (de Ayala, 2009)
2a Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment	BACES assessment item responses.	Item parameters were estimated using an expectation-maximization method for each model. Item fit was assessed using standardized residuals and ICCs.
2b Identify the Person-Location (Theta) Distribution for the BACES assessment	BACES assessment item responses.	Person locations were estimated using an *expected a-posteriori* (EAP) approach with a standard normal prior distribution ($M_{(\theta)}$=0.00 $SD_{(\theta)}$=1.00)
2c Analyze the total item and test information produced from the BACES instrument	BACES assessment item responses.	Standard error of estimation (SEE) were calculated for each item, and item information functions were reviewed. Item information was summed to examine the total test information.
2d Analyze the quality of item distractors on the BACES assessment	BACES assessment item responses.	Distractor analysis compared choices between the top and bottom 25% of examinees as well as the biserial correlations between each option, total-correct score, and theta estimates. Distractors that perform poorly were flagged for review.
2e Compare person and item location estimates from IRT models to those of traditional CTT indices.	BACES assessment item responses and IRT parameter estimates.	Pearson correlations were used to compare (1) IRT theta estimates with CTT total-correct scores, (2) IRT “b” parameters to CTT difficulty index, and (3) IRT “a” parameters to CTT discrimination index.

Table 3.3 Continued
Summary of Methods by Study Objectives

Study Objective	Data Source	Primary Methods
Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.	BACES item responses and demographic. variables	A combination of descriptive analysis and independent *t*-tests were used to compare total-correct scores and theta estimates across different participant groups.

Assess the IRT assumptions for essential unidimensionality and local independence.

Attaining essential unidimensionality (EU) is indicated by defining, "The dimensionality of item response data in terms of the minimum number of traits necessary to achieve LI (local independence)" (Abswoude, van der Ark, & Sijtsma, 2004, p. 5). Similarly, the items that measure each of these dimensions may be independently calibrated as testlets rather than being required to use a complex multidimensional model (Abswoude et al., 2004, de Ayala, 2009). As opposed to strictly unidimensional IRT approaches, EU allows for the researcher to relax these otherwise strict assumptions.

To assess these assumptions, the DIMTEST procedure (Nandakumar & Stout, 1993, and Stout, Froelich, & Gao, 2001) was used to assess the degree that the BACES items departed from unidimensionality. A DETECT procedure, was then used to cluster the items into their respective dimensions (Kim, 1994; Zhang, 1996; Zhang & Stout, 1999). The DETECT procedure utilizes a combination of the items' covariance directionality and a genetic algorithm to produce the most parsimonious set of test dimensions that simultaneously maximize the DETECT statistic. The process is analogous to traditional factor analysis in which the procedure finds the combination of factors that maximize the amount of variance the model takes into account.

Objective two: examine the model fit of the BACES items to a 1PL Rasch, 2PL, and 3PL IRT model.

The BACES item response data were coded into the Xcalibre v4.2 (Guyer & Thompson, 2012) for all IRT and CTT item analyses. All three models were fit to the data to compare among the item parameter estimates each yielded for both test dimensions and the test overall (Yen, 1981). Two methods for assessing model fit were used to compare which of the models fit the

data best. The first of these methods was comparing a chi-square goodness of fit test over the overall fit of each model. As with traditional goodness of fit statistics, a model was considered a poor fit for the data if the chi-square value is statistically significant at $p<.05$ (de Ayala, 2009). Comparing among the three models was done by observing the change in the -2 Log Likelihood (-2LL) statistic per de Ayala's (2009) guidelines. The change in -2LL between two models, he suggests, is a practical approach to comparing model fit that is analogous to comparing changes in R^2 values for linear regression models. The choice of best-fitting model depended on these statistics, but parsimony was also taken into consideration.

Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment.

Item statistics were estimated for each model using an expectation-maximization procedure, which iteratively estimates item parameters until the IRT model converges (default criteria of 0.001), similar to traditional factor analysis techniques. These parameters were produced at the full test, the dimension, and at the item level for each model. A priori estimates of item parameters or "priors" were set at a Mean (SD) of 1.0 (0.25), 0.0 (1.00), and 0.25 (0.25) for item discrimination, difficulty, and pseudo-guessing, respectively as suggested by DeMars (2010) to reduce the chances for "Very odd sets of item parameters" (p. 67). Once estimated, item parameters were tested for fit using the size of their standardized residual similar to a chi-square statistic (de Ayala, 2009) in addition to visual appraisal of each item characteristic curve (ICC). A significant residual value was flagged as a poorly fitting item, and it was reviewed for possible removal.

Identify the person-location (theta) distribution for the BACES assessment.

Theta (θ) estimates for BACES respondents were estimated using the Expected a-Posteriori (EAP) procedure. EAP is a Bayesian estimation method that uses a prior distribution

(default to $M_{(\theta)}$=0.00 $SD_{(\theta)}$=1.00), and estimates θ by using the mean of the posterior distribution. De Ayala (2009) recommends EAP over other estimation procedures because of the lower overall error and ability to estimate accurate θs for individuals who get all answers correct or incorrect. Also, the exploratory nature of this study made EAP an appropriate step to establishing initial item parameters that have not yet been tested.

Analyze the total item and test information produced from the BACES instrument.

Item information functions for each BACES assessment item was computed for reliability analysis. First, these information functions were reviewed to look for θ values for which the BACES assessment items provide the most accurate estimates. Second, the item information functions were summed to generate a total dimension and test information functions for the BACES assessment.

Analyze the quality of item distractors on the BACES assessment.

The quality of item distractors were analyzed using the same methods outlined by CTT analysis (Wise, n.d.). The top 25% and bottom 25% of examinees on each test dimension were compared on the item response patterns for the BACES items for that dimension. Distractor choices were compared between the high and low ability level individuals, and those that perform poorly were flagged for review. As an additional measure, the biserial correlations of each item response option to both the total-correct score and theta estimates were calculated to ensure that the keyed (i.e. correct) response was most strongly associated with a higher ability level. Poorly performing item was defined as any distractor choice that was (1) never chosen, (2) chosen significantly more or less than others, or (3) possessed a higher correlation with either the total-correct score or theta estimates than the keyed answer.

Compare person and item location estimates from IRT models to those of traditional CTT indices.

The Xcalibre program produces CTT item analysis indices such as item-total correlation (discrimination), proportion correct values (difficulty) and total-correct scores. These values were compared to those estimated using IRT to observe differences between the two measurement approaches using Pearson correlations. Consistent with previous research, it was hypothesized that CTT item difficulty would show a significant, negative association with IRT item location (B) parameters, CTT discrimination would be strongly associated with IRT slope (A) parameters, and CTT total-correct scores would be strongly associated with person location (θ) (Fan, 1998; Hays et al., 2000; Stage, 1998; Xu & Stone, 2011).

Objective three: Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.

Known-groups validity evidence was assessed using a combination of independent samples *t*-tests and descriptive comparisons as sample size allowed (DeVellis, 2012). Strength of the validity evidence was determined by both a statistically significant relationship among the three scores and the magnitude of the effect size for each association. Specifically, differences between males and females, degree status, year of residency, and prior exposure to biostatistics or epidemiology were used as points of comparison.

Chapter Summary

Chapter three detailed each specific element of the methods for developing the BACES assessment. To review, this cross-section study was guided by three primary objectives: (1) establish content validity evidence of the BACES assessment; (2) examine the model fit of the BACES items to a 1PL Rasch, 2PL, and 3PL IRT model; and (3) gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons. BACES items were written in a multiple choice, one-best answer format (Case & Swanson, 2002), and given to a four-person review committee to evaluate the instruments' content validity evidence. Then, a group administration format was used to administer the BACES assessment to 10 departments at three academic medical centers within the University of Tennessee system. After scanning the data into digital format, preliminary analyses included comparisons of demographic characteristics and data cleaning to assess statistical assumptions and validity of responses. Primary data analyses included fitting a 1PL Rasch, 2PL, and 3PL IRT model to the dataset to observe item fit and item parameters as well as comparing the IRT parameter estimates to the traditional CTT item indices such as difficulty, discrimination, and total-correct score. Item distractors were also assessed using comparisons between the top and bottom 25% of participants, and through evaluating the strength of each option's biserial correlation with total-correct scores and theta estimates. Finally, validity evidence was gathered through a known-groups validity comparison of demographic variables.

Chapter Four: Results

The following chapter presents the results of the data collection and analysis processes carried out according to design and procedures introduced in Chapter Three. The chapter begins by describing the procedures for data entry as well as processes used to clean the data prior to quantitative analysis. Similar to the previous chapter, the analyses and their results addressed in this chapter will be presented organized by study objective.

Data entry and cleaning

One-hundred and fifty completed assessment forms were gathered through the course of the study. Two forms of the assessment were used, and respondents were asked to identify the particular form they received. Several participants failed to indicate their particular form; however, all but three of these forms were successfully identified as either "A" or "B" when the number of each form administered per department and overall answer patterns were reviewed. Each of the remaining 147 bubble sheets were hand-reviewed for mismarking, ambiguous marking, or multiple responses. When multiple responses were found, the intended answer (if easily discerned) would be marked by hand, so that the electronic review of the error would be more accurate. An example of this situation would be a participant who marked both "A" and "B" for an item but hand-wrote some indication that option "A" was their final answer either by an arrow, words, or scribbling out their other response. All sheets were then scanned into Remark OMR 8 (Gravic Inc.), where the aforementioned errors were manually corrected on a case-by-case basis before the data was transferred for analysis.

The data sheet was next loaded into Microsoft Excel 2013 (Microsoft Corporation) where the two forms were filtered apart for scoring before being merged back into a single dataset. Scoring was done using an excel formula to check the participant's answer choice against the

answer key for each test form. Correct answers were coded as “1” and incorrect were coded as “0.” Finally, the questions from both forms were manipulated in the datasheet so that all of the questions mimicked the order for form “A.” This process had to be completed so that the responses for all participants were correctly included for each item.

Next, the data were transferred to IBM SPSS v.21 (SPSS Inc., Chicago IL) for further analysis. Based on initial frequency statistics, two variables were recoded. The variables for degree-attainment were combined to create an additional category for “Multiple degrees,” and a similar procedure was used to classify medical school location as either “United States” or “International.” Upon completion of data cleaning, the total-correct score showed a near perfect normal distribution (skewness = 0.15, kurtosis = -0.07).

Initial Analysis, Participant characteristics

A total of 147 instruments (77 form “A” and 70 form “B”) were gathered from ten separate academic medical departments among the three different locations (Table 4.1). Descriptive comparisons showed that the average raw score varied slightly with form “A” having a slightly higher average score (M = 14.38, SD = 3.44) compared to form “B” (M = 12.93, SD = 3.60). Although the scores differed, there was no evidence that the amount of missing data was significantly different depending on the form. The site locations were all similar in terms of their average score; however, site “A” (M = 14.33, SD = 3.28) scored the highest of the three sites.

Table 4.1.

Administration Descriptive Statistics

Administration Characteristic	Frequency (Valid %)	Exam Performance Mean (SD)
Test Form		
Form A (1 – 30)	77 (52.4%)	14.38 (3.44)
Form B (16 – 30, 1 – 14)	70 (47.6%)	12.93 (3.60)
Site		
A (*n* = 61)	61 (41.5%)	14.33 (3.28)
B (*n* = 50)	50 (34.01%)	12.96 (3.93)
C (*n* = 36)	36 (24.5%)	13.61 (3.44)

As depicted in Table 4.2, the sample was predominantly male (80, 59.3%), which approximates the national numbers found by Brotherton and Etzel (2012). The majority of participants were in their first year of residency (53, 36.1%), and trained in the United States (97, 80.8%). The most common advanced degree attainment was an MD only (102, 76.7%)), but eight participants (2.6%) reported attaining multiple advanced degrees, most commonly an MD and MA or MS. Finally, only 51 (37.8%), 58 (43.3%), and 44 (33.3%) participants had completed a class in epidemiology, biostatistics, or EBM, respectively.

Table 4.2.

Background Characteristics of Examinees and Raw Score Exam Performance

Background Variable	Frequency (Valid %)	Exam Performance Mean (SD)
Postgraduate Year		
PGY1	53 (36.1%)	14.38 (3.9)
PGY2	33 (22.4%)	12.42 (3.29)
PGY3	22 (15.0%)	13.64 (3)
PGY4	15 (10.2%)	13.87 (3.02)
PGY5	3 (2.0%)	12 (3.61)
PGY6	1 (0.7%)	13 (0)
PGY7	2 (1.4%)	16.5 (2.12)
Degree(s)		
MD	102 (76.7%)	13.38 (3.54)
MD & PhD	2 (1.5%)	17 (5.66)
MD & MPH	1 (0.8%)	19 (0.00)
MD & MS / MA	3 (2.3%)	15 (4.00)
DO	17 (12.8%)	14 (3.46)
MPH	1 (0.8%)	14 (0)
MS / MA	2 (1.5%)	18 (0)
Other	5 (3.8%)	16 (2.12)
Sex*		
Male	80 (59.3%)	13.79 (3.6)
Female	54 (40.0%)	13.65 (3.49)
Training Location		
U.S.	97 (80.8%)	14.12 (3.60)
International	23 (20.2%)	12.61 (3.16)
Epidemiology		
No	84 (62.2%)	13.63 (3.35)
Yes	51 (37.8%)	14.1 (3.77)
Biostatistics		
No	76 (56.7%)	13.49 (3.12)
Yes	58 (43.3%)	14.09 (4.03)
EBM		
No	88 (66.7%)	13.89 (3.82)
Yes	44 (33.3%)	13.5 (2.87)

*Note. One additional participant selected “prefer not to answer.”

Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model

Test for violations of essential unidimensionality and local independence

The first step in examining the model fit of the BACES items is to examine the assessment data for violations of essential unidemensionality (EU) and local independence (LI). The idea of EU comes from the notion that the strictness of the unidimensionality assumption is oftentimes hard to meet in real-life data. In this case, EU refers to defining, "…the dimensionality of item response data in terms of the minimum number of traits necessary to achieve LI (local independence)" (Abswoude, van der Ark, & Sijtsma, 2004, p. 5). As opposed to strictly unidimensional IRT approaches, EU allows the researcher to relax the otherwise strict assumptions because individual dimensions of a single large test can be independently analyzed as a smaller "testlet" rather than forcing the researcher to use a complex multidimensional model (Abswoude et al., 2004, de Ayala, 2009). For example, a chemistry test that had 15 questions on forming basic compounds and 15 questions on balancing chemical equations would likely fail the *strict* unidimensionality assumption, but both topics could be analyzed separately due to EU.

To assess EU and, by extension, LI, the DIMTEST procedure (Nandakumar & Stout, 1993, and Stout, Froelich, & Gao, 2001) was used to assess the degree that the BACES items departed from strict unidimensionality. The DIMTEST procedure tests the hypothesis that the set of items is made up of only one dimension, and if this hypothesis is rejected (i.e. results are statistically significant), then the conclusion is that the items measure multiple dimensions. The procedure found the 30 BACES items to be significantly multidimensional ($T = 3.018$, $p = 0.0013$); therefore, a DETECT procedure was used to cluster the items into their respective dimensions (Kim, 1994; Zhang, 1996; Zhang & Stout, 1999). The DETECT procedure looks for the simplest structure (number of dimensions) within the 30 items by looking at the relationships

among the item response for those items. The results confirmed two 15-item dimensions, which split the test content between clinical epidemiology and statistics. Table 4.3 provides a list of the item numbers associated with each of the two dimensions. For example, items related to research design (e.g. 1, 2, 9, and 25) or common epidemiology concepts (e.g. 17, 23, and 27) all clustered into the first dimension. On the other hand, items that dealt with interpreting statistical tests or concepts such as 3, 7, 11, and 16 were clustered together in dimension two. These results appeared intuitive based on the test blueprint for the exam, but follow-up DETECT procedures were done to look at how the results changed based on removing overly difficult or poorly discriminating items. These follow-up analyses failed to produce a more parsimonious and/or theoretically plausible structure, so the original two-dimensional structure was chosen for IRT parameter calibration. In other words, the IRT parameters “a,” “b,” “c,” and theta would be calculated for the fifteen items in each dimension as if they were individual tests or, “testlets,” where one measured clinical epidemiology and the other measured statistics.

Table 4.3

Two Dimension Solution for DETECT Procedure

Dimension One: Clinical Epidemiology		Dimension Two: Statistics	
Item Number	Topic	Item Number	Topic
1	Retrospective Cohort	3	Equivalence Testing
2	Case-Control	4	Covariates
6	Measurement of Variables	5	Odds Ratio
9	Cross-sectional	7	Statistical Power
10	Measurement of Variables	8	Type I Error
13	Se & SP	11	Effect Size
14	SE & SP	12	Central Tendency
15	Reliability / validity	16	Independent *t*-test
17	RR	18	Non-inferiority testing
19	SE & SP	20	2x2 Factorial Design
23	Rates / Person-time	21	Linear Regression
25	Bias	22	95% Confidence Intervals
27	NNT	24	Cox Regression
29	Reliability / validity	26	Within-Subjects ANOVA
30	Hypothesis testing	28	Internal Validity

Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment

Once the dimensionality of the BACES items had been finalized, the next step in the analysis process was to run an IRT analysis to obtain the initial parameter estimates. Item response data was then entered into XCalibre v4.2 (Guyer & Thompson, 2012) for parameter calibration along with a pre-specified control file that allocated each item to its respective dimension according to the DETECT results. A 1PL, 2PL, and 3PL model was tested for each of the two dimensions as well as the complete set of 30-items. Since the dimensions were identified in the control file, XCalibre produced a single output file for the two dimensions and overall test each time a model was tested. Initially, the 2PL model provided the best appropriation of model

fit, which was decided based on (1) the goodness-of-fit statistics for each dimension and overall test, and (2) the size of the -2Log Likelihood (-2LL) statistic for each model. The IRT software was instructed to "flag" or mark any item that had an unusually high parameter estimate because those items would need to be reviewed and possibly removed before continuing the analysis. Table 4.4 shows that items 3 and 6 were flagged as being overly difficult (i.e. "b" parameters > 3.5), and items 2 and 20 were flagged as being poorly discriminating (i.e. "a" parameters < 0.40). These items were removed, and the model was rerun similar to traditional factor analysis (de Ayala, 2009). Table 4.5 displays the comparisons for overall model fit among the three different IRT models using the same goodness of fit test and -2LL indices. The table shows that both of the testlets had adequate model fit after deleting the four overly difficult items. Model fit, in other words, the difference between the item response patterns predicted by the model and those that were observed in the real data were not significantly different for either the clinical epidemiology dimension ($p = 0.06$) or the statistics dimension ($p = 0.07$). When a model fits the data, that model may be used to describe the same items in future studies (De Ayala, 2009), which is one of the strongest elements of the IRT approach. Although the 2PL model fit the data, five additional models were run to test the impact of removing additional items and/or changing the specified IRT model.

Table. 4.4.
Classical Test Theory Statistics and Item Parameter Estimates for Initial 2PL Model

Item	CTT Statistics		IRT Parameters	
	R	P	a	b
1	0.17	0.65	0.44	-1.52
2	0.14	0.43	0.38*	0.68
3	-0.04	0.14	0.49	3.69**
4	0.05	0.32	0.45	1.71
5	0.19	0.46	0.55	0.30
6	-0.01	0.05	0.79	3.89**
7	0.16	0.81	0.66	-2.32
8	0.10	0.65	0.51	-1.26
9	0.31	0.66	0.75	-1.00
10	0.08	0.50	0.50	0.03
11	0.13	0.37	0.53	1.00
12	0.21	0.45	0.58	0.38
13	-0.06	0.18	0.53	2.83
14	0.10	0.48	0.53	0.19
15	0.08	0.44	0.51	0.47
16	0.10	0.12	0.70	3.05
17	0.32	0.33	0.77	1.02
18	0.38	0.21	0.93	1.69
19	0.48	0.44	1.07	0.31
20	-0.12	0.39	0.40*	1.12
21	0.22	0.29	0.69	1.41
22	0.47	0.33	1.03	0.85
23	0.27	0.46	0.69	0.24
24	0.03	0.16	0.59	2.81
25	0.44	0.42	1.02	0.39
26	0.39	0.27	0.87	1.33
27	0.58	0.49	1.39	0.04
28	0.16	0.35	0.59	1.08
29	0.21	0.42	0.63	0.60
30	0.21	0.44	0.58	0.43

Note. P = “Difficulty Probability,” R = “Biserial Correlation”
*Parameter estimate was poorly discriminating (a < 0.40)
**Parameter estimate was overly difficulty (b > 3.50)

Table 4.5.

Overall IRT Model Fit Statistics for Best Fitting Model

Model	Model Fit Statistics			Δ-2Ln*L*
	Chi-Square (df)	*p*-value	-2Ln*L*	
Rasch (1PL)				
Clinical Epidemiology	256.47 (182)	< 0.001	2273	
Statistics	281.62 (182)	< 0.001	2056	
Full Test	538.08 (364)	< 0.001	4329	
2PL				
Clinical Epidemiology	197.73 (169)	0.06	2190	-83
Statistics	198.19 (169)	0.07	2041	-15
Full Test	395.92 (338)	0.02	4232	-97
3PL				
Clinical Epidemiology	230.72 (156)	< 0.001	2229	39
Statistics	228.05 (156)	< 0.001	2123	82
Full Test	458.77 (312)	< 0.001	4352	120

Note. *p*-value > 0.05 indicates adequate model fit

Note. *p*-value > 0.05 indicates model fit is not significantly affected by the more complete model

Note.Δ-2LnL is calculated by comparing change in -2LnL from a less complete to more complete model.

Once the best model was chosen, the CTT and IRT estimates for the 2PL model could be assessed. The estimates for each item are shown in Table 4.6. On average, the estimated item discrimination (a-parameter) ranged between 0.42 and 1.51 with a mean of (0.75, SD = 0.31) for the clinical epidemiology dimension and between 0.35 and 1.07 with a mean of (0.68, SD = 0.20) for the statistics dimension. This parameter refers to the *slope* of the item characteristic curve (ICC) across a range of ability levels (theta), so a higher "a" parameter indicates a greater ability for an item to discriminate among different ability levels. For example, item 27 has the highest discrimination ability of any item on the test, which means that the probability of correctly answering this item rises sharply across a short span of ability (Figure 4.1). Compare

this item to the relatively flat slope of item 1 where the probability of answering correctly changes very slightly across a wide range of ability levels.

The "b" parameter defines the difficulty estimates for the items, which can be directly compared to the proportion correct ("P" column) to show how items located at a higher level of ability (i.e. a higher "b") translated to a smaller proportion of correct responses. Overall, these ranged from -1.61 to 2.73 (M = 0.29, SD = 1.0) for clinical epidemiology dimension and -2.30 to 2.90 (M = 0.91, SD = 1.45) for statistics. The final test characteristic curves for both testlets and the overall test are shown in figure 4.2, which summarizes the average difficulty and discrimination into a single curve. The ability level at which one is 50% likely to answer a question correctly is considered to be the difficulty or "location" of that particular set of items. Intuitively, the statistics dimension was more difficult, so its location is near $\theta = 1$, or, above average ability. On the other hand, the epidemiology dimension is somewhat easier, so its location is just beyond $\theta = 0$. The location for the overall test fell directly between the two dimensions, which was very close to $\theta = 0$. Additional ICCs for each individual item are located in Appendix F.

Table 4.6

Classical Test Theory Statistics and Item Parameter Estimates for Best Fit 2PL Model

Item	CTT Statistics		IRT Parameters	
	R	P	a	b
1	0.15	0.65	0.42	-1.61
4	0.04	0.32	0.35*	1.96
5	0.19	0.46	0.47	0.33
7	0.17	0.81	0.68	-2.29
8	0.10	0.65	0.52	-1.26
9	0.33	0.66	0.78	-0.97
10	0.06	0.50	0.51	0.03
11	0.12	0.37	0.55	0.97
12	0.21	0.45	0.60	0.36
13	-0.07	0.18	0.55	2.73
14	0.12	0.48	0.56	0.18
15	0.09	0.44	0.53	0.45
16	0.13	0.12	0.75	2.90
17	0.32	0.33	0.81	0.99
18	0.36	0.21	0.95	1.66
19	0.50	0.44	1.12	0.30
21	0.23	0.29	0.72	1.35
22	0.48	0.33	1.07	0.82
23	0.28	0.46	0.73	0.23
24	0.03	0.16	0.62	2.69
25	0.45	0.42	1.06	0.38
26	0.40	0.27	0.92	1.28
27	0.62	0.49	1.51	0.03
28	0.17	0.35	0.62	1.04
29	0.22	0.42	0.65	0.57
30	0.21	0.44	0.60	0.41

Note. P = “Difficulty Probability,” R = “Biserial Correlation”

*Parameter estimate was poorly discriminating (a < 0.40)

Table 4.7

Final Two Dimensions of BACES Assessment After Removing Poor Items

Item Number*	Topic	Item Number*	Topic
1	Retrospective Cohort	4	Covariates
9	Cross-sectional	5	Odds Ratio
10	Measurement of Variables	7	Statistical Power
13	Sensitivity & Specificity	8	Hypothesis testing
14	Sensitivity & Specificity	11	Effect Size
15	Reliability / validity	12	Central Tendency
17	Relative Risk	16	Independent *t*-test
19	Sensitivity & Specificity	18	Non-inferiority testing
23	Rates / Person-time	21	Linear Regression
25	Bias	22	95% Confidence Intervals
27	Number Needed to Treat	24	Cox Regression
29	Reliability / validity	26	Within-Subjects ANOVA
30	Hypothesis testing	28	Reliability / validity

Note. Items 2, 3, 6, and 20 were removed.

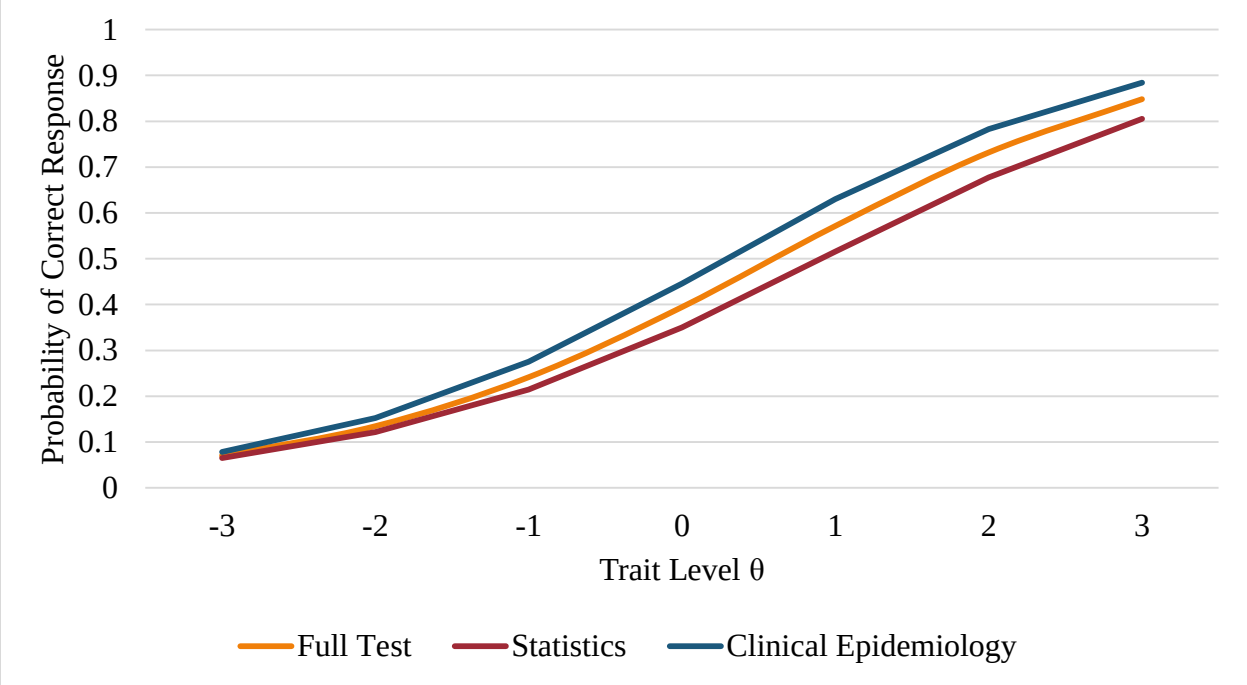


Figure 4.1: Test Characteristic Curve for Clinical Epidemiology Dimension, Statistics Dimension, and Full Test

Person Location Estimates (Theta)

One of the strengths of IRT over CTT is the ability to estimate both item and person parameters on the same scale, so the difficulty or discrimination of a certain item can be discussed in terms of the ability level they are most suited for measuring. Figure 4.2 displays the frequency distribution for the person location (i.e. theta) estimates for the full test, research methods dimension, and statistics dimension. The highest frequency of person location estimates fell between -0.8 and -0.4 for both research and statistics dimensions while the full test was somewhat more spread out with 87 estimates falling between -0.8 and 0.4. When the normal distribution of raw test scores, the prior distribution used to assist with estimating theta ($M = 0$, $SD = 1$), it is logical that the theta estimates would cluster near $\theta = 0$, or, "average" ability. To put this another way, the raw scores were very closely clustered near the average score, and only a couple participants scored far beyond that average. When these scores were translated into theta estimates that are on a scale that has an average of 0.0 and standard deviation of 1.0, it made sense to see so many of the participants' thetas very close to 0.00.

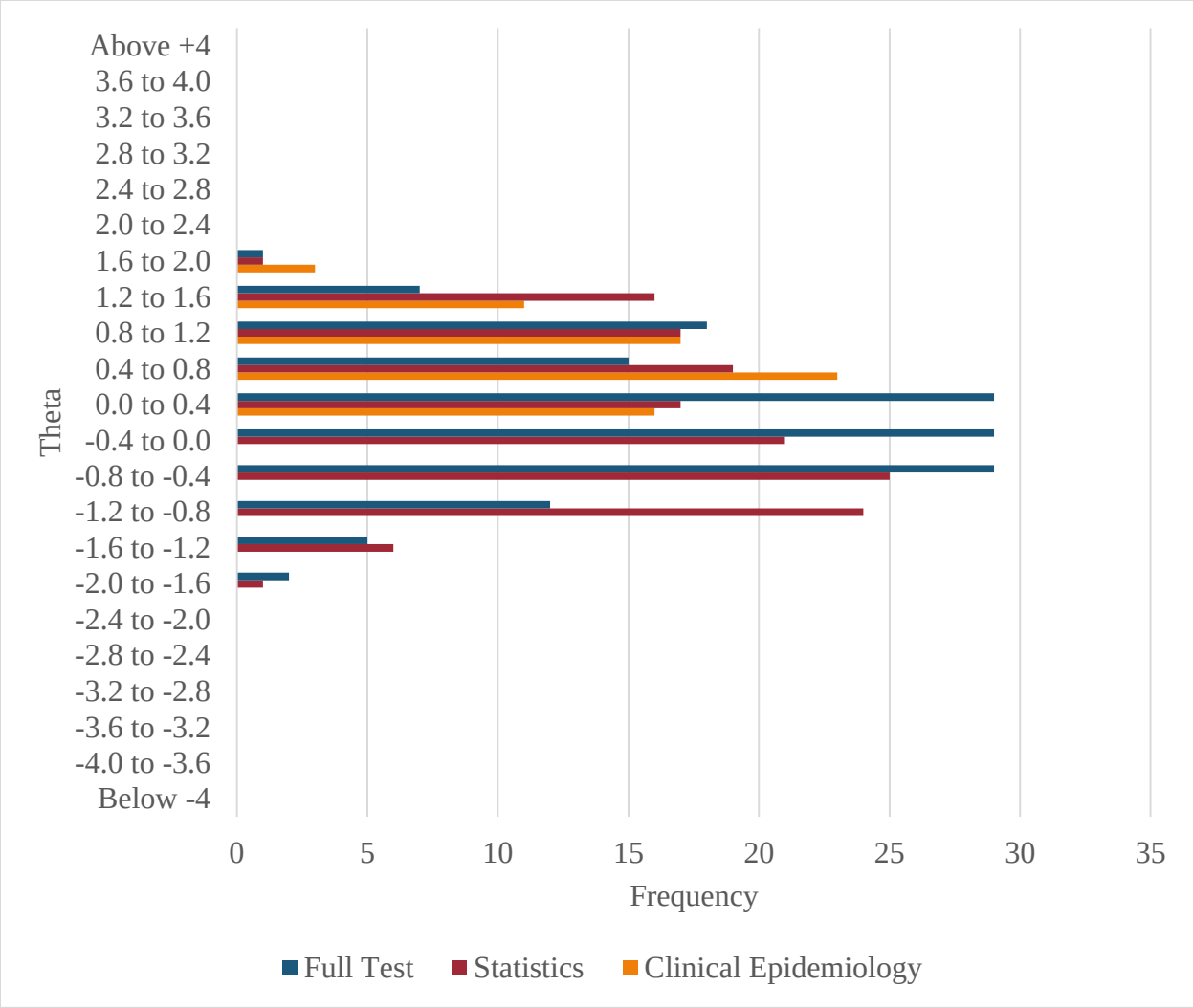


Figure 4.2. Distribution of Theta Estimates for Best Fit Model

Analyze the quality of item distractors on the BACES assessment

Once person and item estimates had been completed, the analysis examined the response options for each item. To accomplish this task, a distractor analysis was completed using two different approaches. One common method for considering quality distractors is by comparing the answering patterns for the top and bottom 25% of each dimension's total-correct score (Tables 4.8 and 4.9) (Wise, n.d.). Table 4.8 shows that the top 75% did not choose 11 of the distractors in the epidemiology dimension and eight in the statistics dimension. On the other

hand, the lowest 25% selected all but two of the possible distractors on the epidemiology dimension and every distractor on the statistics dimension. A distractor would be considered poor if it was not chosen by *either* quartile, which did not occur in the BACES data; however, many distractors were chosen by only one or two individuals such as items 1, 10, and 14 "A" or 7 "D." These options were sparsely chosen, so they were flagged for review, but overall these frequencies indicate the item distractors performed correctly in misleading those with relatively little knowledge while not "tricking" the higher performing examinees.

Table 4.8
Response Option Information for the Top and Bottom 25% of Participants for the Both BACES Dimensions

Dimension Item (Key)	Frequency of Responses to Each Option							
	A		B		C		D	
	25%	75%	25%	75%	25%	75%	25%	75%
Clinical Epidemiology								
1 (D)	2	0	12	5	7	1	13	25
9 (D)	5	1	8	2	8	0	13	28
10 (B)	1	0	18	7	6	20	7	0
13 (C)	2	0	7	8	15	11	4	7
14 (B)	0	1	7	24	8	4	17	1
15 (C)	3	10	16	4	9	17	6	0
17 (B)	13	5	3	23	16	2	2	1
19 (B)	9	0	2	30	10	0	13	1
23 (C)	11	3	12	4	6	22	5	2
25 (B)	4	1	2	27	14	1	14	2
27 (B)	11	0	2	31	3	0	18	0
29 (A)	10	23	7	3	17	2	0	3
30 (A)	9	21	9	1	10	5	6	4
Statistics								
4 (A)	9	17	3	1	16	6	15	8
5 (A)	9	23	11	2	10	2	13	5
7 (C)	6	0	10	2	26	30	1	0
8 (A)	13	26	8	0	19	6	3	0
11 (B)	1	0	13	0	10	23	14	7
12 (A)	1	0	9	24	19	2	9	3
16 (D)	11	1	21	12	7	9	3	9
18 (A)	2	16	18	0	16	15	7	1
21 (A)	3	17	4	5	18	5	18	5
22 (D)	7	5	12	1	20	5	4	21
24 (C)	11	5	21	15	3	9	8	3
26 (D)	3	1	30	5	8	5	2	21
28 (D)	23	1	7	5	7	3	6	23

Note. Quartiles were calculated based off of each participant's raw score on the item's dimension.

The second approach to assessing the strength of each response option was through additional biserial correlations between each item's response options and both the total-correct score and estimated theta values (Table 4.9). In contrast to the first approach, this additional analysis revealed that the keyed response option *did not have the strongest correlation* with total-correct and theta estimates for items 13, 15, 16, and 24. In other words, individuals who chose

item 13 “A” were more likely to have a higher ability level and score correct on the epidemiology dimension than those who chose the correct answer “C.” These findings added to the information that the quartile comparison gathered about the distractors because it provided more specific information about which particular options were possibly unfair or troublesome regardless of the number of participants who chose that option.

Table 4.9.
Item Response Option Correlation to Total Score (rS) and Theta (rθ)

Item (Key)	Item Response Option							
	A		B		C		D	
	*r*S	*r*θ	*r*S	*r*θ	*r*S	*r*θ	*r*S	*r*θ
1 (D)	-0.13	-0.14	-0.05	-0.11	-0.08	-0.14	0.15	0.24
4 (A)	0.04	0.08	-0.13	-0.12	-0.06	-0.09	0.07	0.06
5 (A)	0.19	0.25	-0.14	-0.17	-0.05	-0.05	-0.06	-0.10
7 (C)	-0.14	-0.17	-0.10	-0.16	0.17	0.24	-0.04	-0.03
8 (A)	0.10	0.15	-0.23	-0.25	0.01	-0.03	0.04	0.02
9 (D)	-0.17	-0.19	-0.18	-0.23	-0.15	-0.23	0.33	0.43
10 (B)	0.00	-0.09	0.06	0.16	-0.22	-0.23	0.15	0.12
11 (B)	-0.16	-0.18	0.12	0.20	0.09	0.05	-0.09	-0.13
12 (A)	0.21	0.28	-0.21	-0.25	-0.01	-0.07	-0.01	0.02
13 (C)*	0.03	0.02	-0.02	-0.09	-0.07	0.01	0.10	0.11
14 (B)	0.16	0.15	0.12	0.22	0.01	-0.05	-0.14	-0.21
15 (C)*	0.25	0.22	-0.19	-0.24	0.09	0.18	-0.20	-0.24
16 (D)*	-0.34	-0.37	-0.02	-0.05	0.23	0.25	0.13	0.19
17 (B)	-0.06	-0.10	0.32	0.45	-0.26	-0.36	-0.01	0.00
18 (A)	0.36	0.48	-0.47	-0.54	0.24	0.20	-0.14	-0.13
19 (B)	-0.26	-0.31	0.50	0.64	-0.06	-0.13	-0.33	-0.4
21 (A)	0.23	0.36	0.07	0.03	0.02	0.02	-0.28	-0.37
22 (D)	0.06	0.02	-0.25	-0.25	-0.31	-0.41	0.48	0.61
23 (C)	-0.17	-0.21	-0.16	-0.26	0.28	0.41	-0.01	-0.03
24 (C)*	0.19	0.2	-0.14	-0.18	0.03	0.10	-0.06	-0.09
25 (B)	-0.11	-0.17	0.45	0.61	-0.24	-0.30	-0.21	-0.30
26 (D)	0.21	0.21	-0.49	-0.58	0.02	0.01	0.40	0.50
27 (B)	-0.28	-0.36	0.62	0.77	-0.09	-0.15	-0.41	-0.48
28 (D)	-0.12	-0.16	0.03	-0.02	-0.10	-0.14	0.17	0.27
29 (A)	0.22	0.33	-0.10	-0.15	-0.23	-0.31	0.13	0.14
30 (A)	0.21	0.27	-0.27	-0.28	0.04	-0.03	-0.07	-0.07

*Keyed answer did not have largest correlation to theta or score

Analyze the total item and test information produced from the BACES instrument

The final step in the IRT analysis was to estimate the item and test information for each BACES item and dimension. Recall, item information is the inverse of the standard error of estimate (SEE) along different values of theta (de Ayala, 2009). It is the IRT equivalent of CTT reliability because higher information converts to lower SEE, which indicates a more accurate estimate of theta. Unlike CTT, estimates of IRT information are put in terms of ability level, so each item has a particular range of theta that it is particularly accurate in measuring. Figures 4.3a – 4.3c display several of these item information functions (IIF). First, 4.3a shows the total information provided by both dimensions and all 26 remaining BACES items. Figures 4.3b and 4.3c show the IIFs for each of the 13 remaining items in each the clinical epidemiology and statistics dimensions, respectively. Overall, the results indicated that the clinical epidemiology dimension reached its maximum information of 2.04 at theta = 0.15, or, a slightly above-average level of ability. Meanwhile, the statistics testlet reached its peak information of 1.43 at theta = 1.20. Similarly to the ICCs, the overall test met in the middle with its highest information of 3.22 at theta = 0.45.

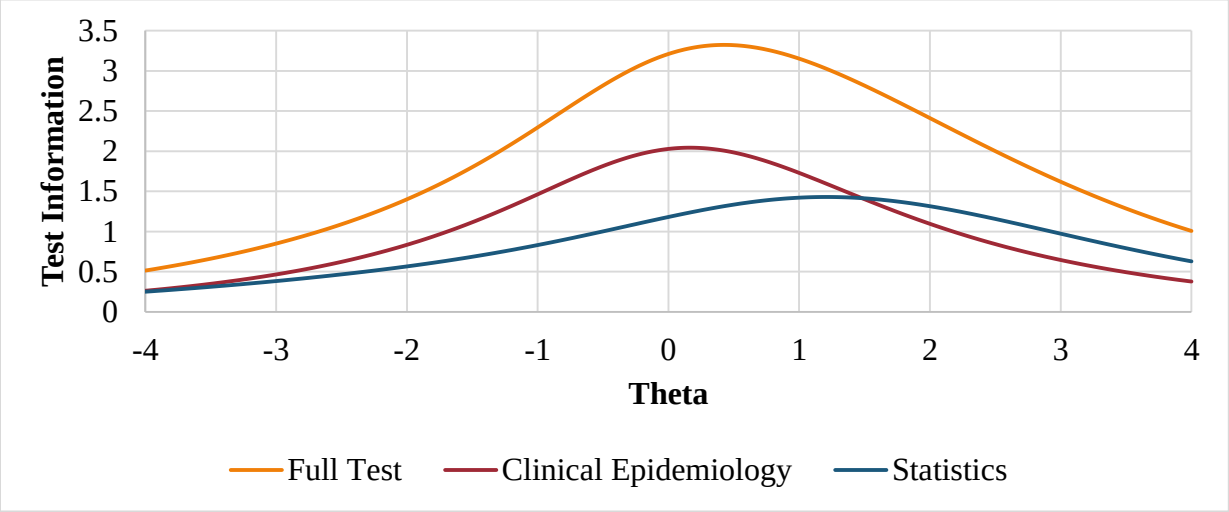


Figure 4.3a. Total Information Curves for Clinical Epidemiology, Statistics, and Full Test

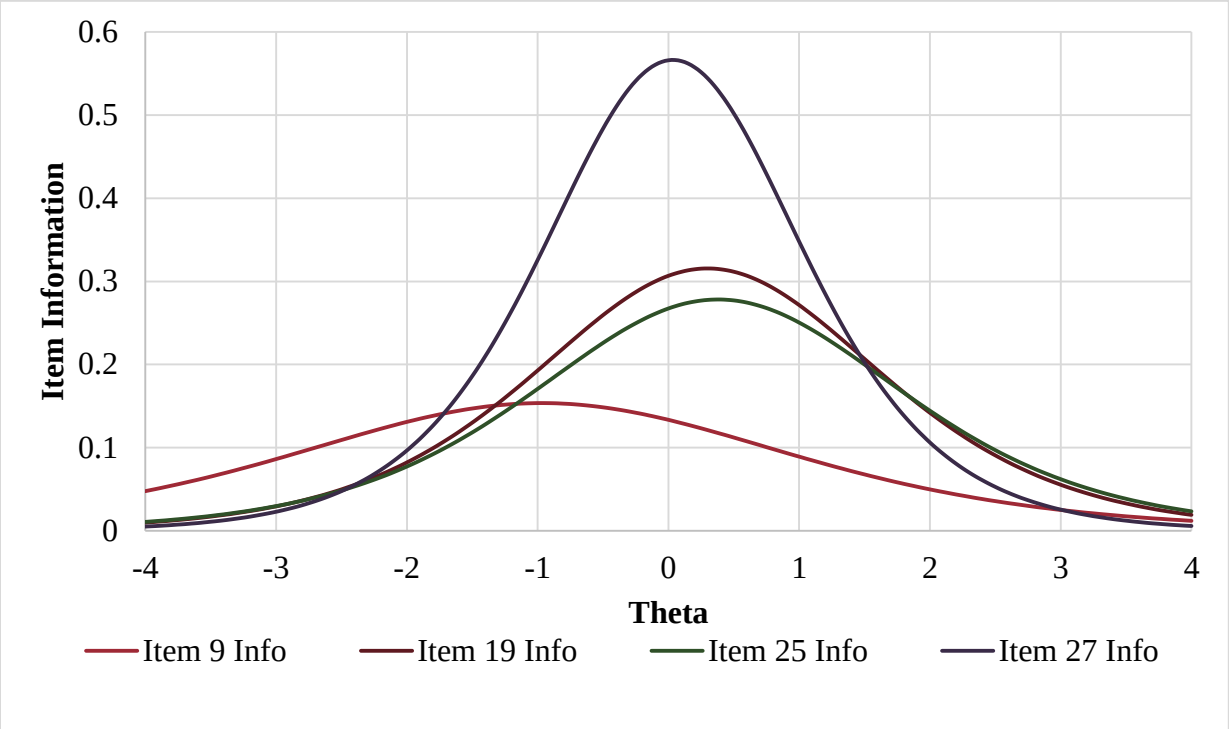


Figure 4.3b. Item Information Curves for Four Clinical Epidemiology Dimension Items

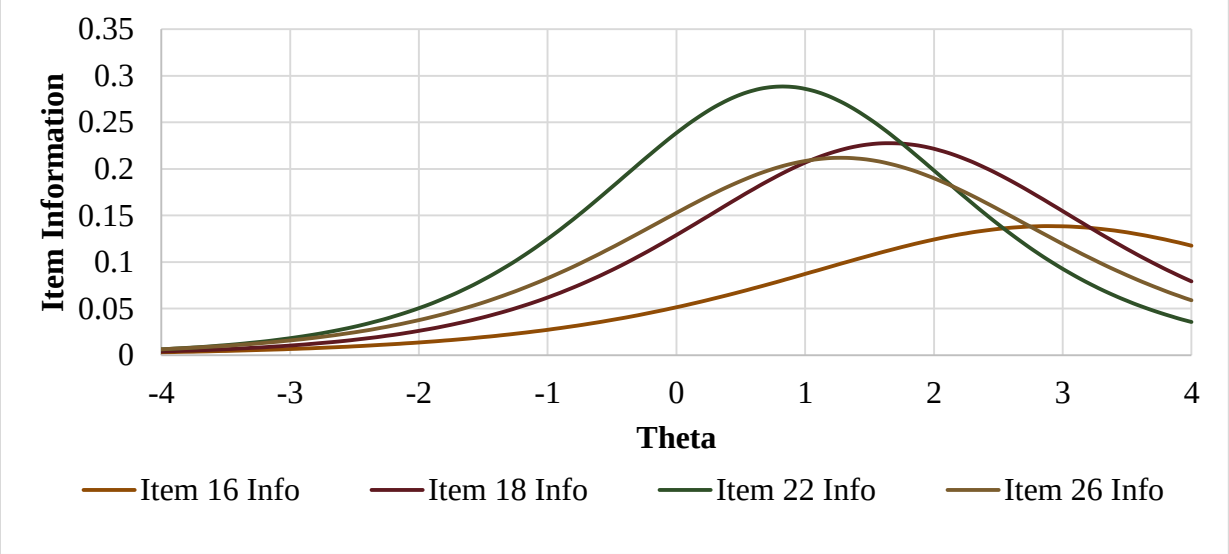


Figure 4.3c. Item Information Curves for Four Statistics Dimension Items

Compare person and item location estimates from IRT models to those of traditional CTT indices.

The final analysis was completed using the estimated IRT parameters and their CTT counterparts, so that the accuracy of the final model could be compared to what a researcher would have seen had CTT been the only method employed. For the analysis, the "a", "b", and theta estimates from the final 2PL model were entered into a separate datasheet along with each item's CTT difficulty and discrimination, and total-correct score values for comparative analysis. Pearson *r* correlations were used to quantify the extent to which CTT and IRT estimates were related to one another, and discovered that the estimates for "a", "b", and theta parameters were very strongly correlated with their CTT counterparts on each dimension as well as the full test. Specifically, CTT difficulty (P) was significantly, negatively related to IRT difficulty "b" ($r(24) = -0.980$, $p < 0.001$), and CTT discrimination (R) was significantly, positively associated with IRT discrimination "a" ($r(24) = 0.91$, $p < 0.001$). Similarly, theta estimates for person ability were significantly, positively related to CTT total correct scores for research methods, statistics, and the full test (Table 4.10).

Table 4.10.
Correlation Among Total Correct Scores and Theta Estimates for Best Fitting Model

Score	Research Methods Score	Statistics Score	Research Methods Theta	Statistics Theta	Full Test Score
Statistics Score	0.39**				
Research Methods Theta	0.98**	0.42**			
Statistics Theta	0.47**	0.98**	0.50**		
Full Test Score	0.87**	0.80**	0.87**	0.84**	
Full Test Theta	0.88**	0.75**	0.91**	0.81**	0.98**

Note. $^{**}p < 0.001$

Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.

Known-groups comparisons were completed on participants who provided their sex (n = 134), year of training (n = 129), and previous exposure to epidemiology (n = 135), biostatistics (n = 134), and EBM (n = 132). Due to the low group size, year of training and degree performance differences were examined using descriptive analysis only (Table 4.11). Also, correlational analyses were not possible for year of training because the distribution was highly skewed. The remaining comparisons showed only one significant difference in performance (Table 4.12), specifically, those who reported taking a course in biostatistics (M = 5.29, SD = 2.47) performed significantly better than those who did not take a course (M = 4.39, SD = 1.97) on the statistics testlet raw score $t(106.78) = -2.271$, $p = 0.025$. Although not statistically significant, participants who reported previous experience with EBM, epidemiology, or biostatistics scored slightly *lower* on the clinical epidemiology testlet than those who did not report such experiences.

Table 4.11.

Demographic Comparisons for Total-Correct Score and Theta Estimates of Final 2PL Model

Background Variable	Mean (SD)					
	Clinical Epi. Score	Clinical Epi. θ	Statistics Score	Statistics θ	Full Test Score	Full Test θ
Postgraduate Year						
PGY1	6.06 (2.6)	0.01 (0.85)	4.81 (2.32)	-0.01 (0.76)	10.87 (4.35)	0 (0.93)
PGY2	5.67 (2.7)	-0.03 (0.86)	4.33 (2.29)	-0.17 (0.75)	10 (3.85)	-0.11 (0.83)
PGY3	5.95 (2.84)	0.05 (0.94)	4.41 (2.11)	-0.06 (0.73)	10.36 (4.51)	-0.01 (1)
PGY4	5.6 (2.64)	-0.15 (0.81)	5.4 (1.96)	0.21 (0.67)	11 (2.88)	0.03 (0.64)
PGY5	6 (2.65)	-0.06 (0.99)	5 (1)	0.02 (0.41)	11 (3.46)	-0.04 (0.88)
PGY6	7 (0)	0.61 (0)	4 (0)	-0.21 (0)	11 (0)	0.33 (0)
PGY7	8 (0)	0.5 (0.05)	7 (1.41)	0.73 (0.5)	15 (1.41)	0.75 (0.24)
Degree(s)						
MD	5.77 (2.63)	-0.06 (0.85)	4.67 (2.16)	-0.05 (0.71)	10.44 (4.13)	-0.08 (0.88)
MD & PhD	8.5 (2.12)	0.99 (0.56)	6 (1.41)	0.49 (0.22)	14.5 (0.71)	0.94 (0.24)
MD & MPH	8 (0)	0.77 (0)	6 (0)	0.59 (0)	14 (0)	0.86 (0)
MD & MS / MA	6.67 (2.31)	0.33 (0.4)	6.67 (1.53)	0.66 (0.42)	13.33 (2.52)	0.6 (0.32)
DO	5.47 (3.16)	-0.06 (0.96)	5.24 (2.54)	0.1 (0.82)	10.71 (4.5)	0.03 (0.95)
MPH	9 (0)	0.85 (0)	3 (0)	-0.66 (0)	12 (0)	0.27 (0)
MS / MA	9.5 (0.71)	1.17 (0.3)	4 (1.41)	-0.23 (0.62)	13.5 (0.71)	0.68 (0.13)
Other	4.8 (1.64)	-0.4 (0.56)	5.2 (2.49)	0.06 (0.86)	10 (3.32)	-0.24 (0.74)

Table 4.12.

Known-Groups Demographic Comparisons for Total-Correct Score and Theta Estimates of Final 2PL Model

Background Variable	Mean (SD)					
	Clinical Epi. Score	Clinical Epi. θ	Statistics Score	Statistics θ	Full Test Score	Full Test θ
Sex						
Male	5.71 (2.72)	-0.04 (0.88)	4.71 (2.12)	-0.05 (0.7)	10.43 (4.07)	-0.06 (0.88)
Female	6.26 (2.51)	0.07 (0.83)	4.89 (2.42)	0.05 (0.81)	11.15 (4.11)	0.08 (0.89)
Training Location						
U.S.A	5.79 (2.69)	-0.04 (0.85)	4.92 (2.31)	0.01 (0.76)	10.71 (4.22)	-0.02 (0.9)
International	6.04 (2.95)	0.04 (0.96)	4.83 (1.87)	0.07 (0.62)	10.87 (3.97)	0.06 (0.87)
Epidemiology						
No	6.13 (2.65)	0.08 (0.84)	4.55 (2.04)	-0.08 (0.67)	10.68 (3.62)	0.01 (0.78)
Yes	5.86 (2.53)	-0.04 (0.85)	5.25 (2.54)	0.13 (0.85)	11.12 (4.72)	0.04 (1.01)
Biostatistics						
No	6.18 (2.63)	0.09 (0.85)	4.39 (1.97)	-0.11 (0.66)	10.58 (3.61)	0 (0.79)
Yes	5.6 (2.64)	-0.11 (0.86)	5.29 (2.47)	0.12 (0.82)	10.9 (4.67)	-0.02 (1)
EBM						
No	6.26 (2.58)	0.11 (0.84)	4.75 (2.23)	-0.01 (0.74)	11.01 (4.1)	0.07 (0.88)
Yes	5.52 (2.64)	-0.12 (0.85)	4.84 (2.31)	0 (0.77)	10.36 (4.08)	-0.08 (0.88)

Chapter Summary

The results described throughout this chapter have provided evidence for how the BACES assessment has performed in its first administration to 147 medical residents. The test was made up of two distinct dimensions as evidenced by both the DIMTEST (Nandakumar & Stout, 1993; Stout, et al., 2001) and DETECT (Kim, 1994; Zhang, 1996; Zhang & Stout, 1999) procedures. These two dimensions evenly split the original 30-items into 15 related to clinical epidemiology and 15 related to statistical interpretation. The IRT parameters for each of these "testlets" were calibrated separately in addition to the full 30-items. Initial IRT analysis showed the 2PL model to be the best-fitting model over the 1PL Rasch or 3PL options; however, four overly difficult items had to be deleted prior to achieving adequate model fit for both the clinical epidemiology and statistics testlets.

The next step was to appraise the quality of the remaining 26 items' response options and distractors. These investigations found that the response options, overall, were performing correctly for each of the test dimensions. In other words, the number of participants who chose each distractor was significantly lower in the top 25% of participants compared to the bottom 25%. On the other hand, additional biserial correlations with each response option to total-correct score and theta estimates revealed possible problems with at least one distractor on items 13, 15, 16, and 24.

After these distractors had been analyzed, the reliability of the best-fitting model was established through examining the SEE and information values at the item, dimension, and test levels. These findings concluded that the clinical epidemiology testlet reached its maximum information at slightly above-average level of ability while the statistics testlet was most accurate at roughly one standard deviation above the average. The accuracy of the final model was

additionally examined through correlations among the IRT and CTT parameters, which supported previous research in finding very strong relationships between IRT estimates and their CTT counterparts.

Construct validity evidence was statistically inconclusive with the exception of a significant increase in scores on the statistics dimension in individuals who reported previous biostatistics coursework. There were no differences between men and women on performance across any dimension of the BACES assessment nor was there a drastic difference in performance between residents trained in the U.S.A. versus internationally.

Chapter five will position these findings within the context of the study itself, the larger GME atmosphere, and future of the BACES assessment. Specific conclusions, recommendations for practice, and suggestions for future research will all be discussed for each of the three primary research objectives.

Chapter Five: Discussion

The primary purpose of this chapter is to position the results from developing the BACES assessment within the larger body of literature from which it arose. First, the primary purpose and objectives of the study will be reviewed. Second, the results from chapter four will be reviewed one research objective at a time to illustrate specific links with previous research. The third section will acknowledge and review a number of the limitations associated with study's methods, results, and conclusions while at the same time offering suggestions for future researchers to improve upon these limitations. Finally, a number of implications for GME policy and practice will be described along with directions for future investigations.

Summary of Study Purpose, Objectives, and Method

As reviews of top tier medical journals over the past 30 years have shown a steady *increase* in the frequency and complexity of statistical methods (Horton & Switzer, 2005; Reed, Salen, & Bagher, 2003; Weiss et al., 1980; Windish et al., 2007), it is essential that medical residents possess an adequate knowledge of clinical epidemiology and biostatistics if they are to effectively integrate EBM into their practice (Sahai, 1999; Hatala & Guyatt, 2002) In reality, studies from the past several decades have shown a consistently low, variable knowledgebase among graduate medical students (Berwick et al., 1981; Novack et al., 2006; Weiss & Samet, 1980; Windish et al., 2007). Moreover, the highly variable course designs used to teach these skillsets, the qualifications of the course instructor, and the GME learning environment has made assessment of these skills difficult (M. L. Green, 2001; M. Green, 1999; Hatala & Guyatt, 2002).

Although there have been numerous assessments of instruments for assessing these topics (e.g. Berwick et al., 1981; Enders, 2011; Fritsche et al., 2002; Windish, 2011), the lack of formal psychometric analysis and use of CTT item parameters make the instruments' difficulty,

discrimination, and reliability values irrelevant to residents outside of their original sample. Without a consistent, generalizable instrument to gauge resident competency, GME educators are left with no answer to the question of *how do educators effectively prepare and assess physicians in biostatistics and clinical epidemiology?*

The purpose of the present study was to address this question. Specifically, to establish preliminary item characteristics and validity evidence for the Biostatics and Clinical Epidemiology Skills (BACES) assessment. Rather than use CTT to develop the instrument, Item Response Theory (IRT) was used in order to offer educators item and person ability parameters that are independent of the sample from which they are estimated. This invariance trait could provide GME educators the freedom to choose specific, relevant biostatistics and clinical epidemiology assessment topics, and administer it to their residents while maintaining the item's difficulty, discrimination, and ability estimates. The study specifically aimed to address three primary objectives:

1. Establish content validity evidence of the BACES assessment
2. Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model
    a. Test for violations of essential unidimensionality and local independence
    b. Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment
    c. Analyze the quality of item distractors on the BACES assessment
    d. Analyze the total item and test information produced from the BACES instrument
    e. Compare person and item location estimates from IRT models to those of traditional CTT indices.

3. Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.

For brevity and clarity, Table 5.1 provides an overview of the methods used to address each of the primary study objectives. Recall, the purpose of this study was to develop the BACES instrument, and to obtain *preliminary* evidence for its quality and validity. The progress made towards meeting this purpose is described in the following section.

Table 5.1
Summary of Methods by Study Objectives

	Study Objective	Primary Methods
1	Establish content validity evidence of the BACES assessment	An expert in assessment, epidemiology, medicine, and clinical pharmacy reviewed the BACES items and provided feedback.
2	Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model	Each model was fit and compared for the best-fitting model, which was assessed via chi-square goodness of fit index, and by comparing change in -2 Log Likelihood statistics between models (de Ayala, 2009)
2a	Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment	Item parameters were estimated using an expectation-maximization method for each model. Item fit was assessed using standardized residuals and ICCs.
2b	Identify the Person-Location (Theta) Distribution for the BACES assessment	Person locations were estimated using an *expected a-posteriori* (EAP) approach with a standard normal prior distribution ($M_{(\theta)}$=0.00 $SD_{(\theta)}$=1.00)
2c	Analyze the total item and test information produced from the BACES instrument	Standard error of estimation (SEE) were calculated for each item, and item information functions were reviewed. Item information was summed to examine the total test information.
2d	Analyze the quality of item distractors on the BACES assessment	Distractor analysis was used to compare the frequency of distractor choices between the top and bottom 25% of examinees as well as the biserial correlations between each option, total-correct score, and theta estimates. Distractors that perform poorly were flagged for review.
2e	Compare person and item location estimates from IRT models to those of traditional CTT indices.	Pearson correlations were used to compare (1) IRT theta estimates with CTT total-correct scores, (2) IRT "b" parameters to CTT difficulty index, and (3) IRT "a" parameters to CTT discrimination index.
3	Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons.	A combination of descriptive analysis and independent *t*-tests were used to compare total-correct scores and theta estimates across different participant groups.

Implementation and Results of BACES Development Process

The BACES assessment was developed using a one-best-answer format (Case & Swanson, 2002) that presented the examinee with a clinical example or vignette based on broadly applicable medical conditions or procedures. Each question asked the resident to respond to a question regarding the example in a 4-option multiple choice question (MCQ) format. Response options were selected specifically to allow educators and residents distinguish precisely where their thinking went wrong on a given question (Suskie, 2009).

A cross-sectional, convenience sample of 147 residents was collected from 10 separate departments at three academic medical centers across the state of Tennessee. Each administration took place in a group, paper-and-pencil format during a scheduled didactic session or journal club meeting. The resident received one of two parallel BACES assessments along with a set of background demographic questions based on those used in previous research (CITES). The assessment took approximately 20-30 minutes to complete for each administration. Once completed, the resident was given a descriptive answer key (Appendix C) that provided the answers to their assessment, description of each response option, and a scannable link to online lecture resources. The primary findings from this study are summarized by research objective as follows:

1. *Establish content validity evidence of the BACES assessment*
    a. A four-person expert review group determined that the 30 BACES items met their standards for quality.
    b. A heterogeneous group of reviewers allowed for the items' content to be critiqued from multiple angles.
    c. Construct irrelevant difficulty, a major threat to content validity (Furr & Bacharach, 2009), was minimized by consulting with an advanced surgical

resident for broadly applicable medical situations and procedures, which were then used as the clinical vignettes for BACES items.

d. Two of the four expert reviewers were informally interviewed for additional follow-up discussion regarding changes to the instrument.

2. *Examine the model fit of the BACES items to a Rasch, 2PL, and 3PL IRT model*
    a. The DIMTEST (Nandakumar & Stout, 1993, and Stout, Froelich, & Gao, 2001) and DETECT (Kim, 1994; Zhang, 1996; Zhang & Stout, 1999) procedures concluded that the 30-item instrument was not strictly unidimensional, but splitting the test into two even dimensions satisfied both essential unidimensionality and local independence assumptions.
    b. The best-fitting model for the data was the 2PL model; however, items 2, 3, 6, and 20 were removed for being under-discriminating (2 and 20) or overly difficult (3 and 6).
    c. After these four items were removed, the remaining 26 items achieved an adequate level of model fit for the clinical epidemiology dimension (13 items, $p = 0.07$) and statistics dimension (13 items, $p = 0.06$), which indicated that using the IRT approach was appropriately used in this study.
3. *Identify the distribution of item discrimination values, difficulty, and pseudo-guessing parameters for the BACES assessment*
    a. The BACES assessment items covered a range of estimated item difficulty from
        i. -2.30 to 2.90. The addition of a pseudo-guessing parameter adversely affected the model fit, so this parameter was not included due to this poor fit and lack of an adequate sample size.

b. The estimated item discrimination (a-parameter) ranged between 0.42 and 1.51 with a mean of (0.75, SD = 0.31) for the clinical epidemiology dimension and between 0.35 and 1.07 with a mean of (0.68, SD = 0.20) for the statistics dimension.

    i. The most discriminating item was number 27, which required the residents to correctly apply number needed to treat (NNT) to a specific situation.

    ii. Conversely, the item that discriminated the poorest was number 4 which required residents to correctly interpret the term "covariates" in a research scenario.

c. The item difficulty values (b-parameter) ranged from -1.61 to 2.73 (M = 0.29, SD = 1.0) for the clinical epidemiology dimension and -2.30 to 2.90 (M = 0.91, SD = 1.45) for statistics.

    i. The most difficult question of those that remained was item 16, which required the residents to correctly apply an independent *t*-test to a research scenario.

    ii. On the other extreme, item 7 was the easiest of the 26 remaining items, and it asked participants to identify two different factors that influence statistical power.

4. Identify the Person-Location (Theta) Distribution for the BACES assessment

a. The raw score for the BACES data were nearly perfectly normally distributed (Skewness = 0.15, Kurtosis = -0.07); the homogeneous scores likely influenced the clustering of theta estimates near 0.00 (i.e. "average ability").

b. Average proficiency estimates varied slightly between the two dimensions, but their distributions both rounded to a mean of 0.00 and standard deviation of 1.00.

   i. Because the final distributions of theta estimates closely resembled the EAP prior distribution (M = 0.00, SD = 1.00), the choice of prior distribution was likely *not* biased (DeMars, 2010).

c. The highest frequency of proficiency estimates fell between -0.8 and -0.4 for both the clinical epidemiology and statistics dimensions while the full test was somewhat more spread out with 87 estimates falling between -0.8 and 0.4.

5. *Analyze the total item and test information produced from the BACES instrument*

   a. Information estimates were variable among the 26 items, but overall, the combined assessment items reached their peak information of 3.32 at $\theta = 0.45$. These results suggest that the preliminary BACES assessment has the lowest standard error (i.e. highest reliability) in measuring proficiency levels slightly above average.

   b. The most reliable item on the clinical epidemiology dimension was item 27 (max information of 0.56 at $\theta = 0.05$), which was a question about number needed to treat.

   c. Item number 22 (max information of 0.29 and $\theta = 0.08$) was the most reliable item on the statistics dimension. This item required residents to interpret a 95% confidence interval based on a research scenario.

6. *Analyze the quality of item distractors on the BACES assessment*

   a. Every one of the 120 possible response options was chosen at least once by the participants.

b. A total of two distractors were not chosen at all by individuals in the lowest 25% of raw scores, which compared to 19 that were not chosen by the highest 25%.

c. Several distractors such as items 1, 10, and 14 "A" or 7 "D" were very sparsely chosen, and were flagged for review.

d. Additional biserial correlations with each distractor to the raw score and theta estimates found that the keyed response option *did not have the strongest correlation* with total-correct and theta estimates for items 13, 15, 16, and 24. In other words, individuals who chose item 13 "A" were more likely to have a higher ability level and score correct on the epidemiology dimension than those who chose the correct answer "C."

7. *Compare person and item location estimates from IRT models to those of traditional CTT indices.*

a. The IRT and CTT person and item parameters were significantly related to one another.

b. CTT difficulty (P) was significantly, negatively related to IRT difficulty "b" ($r(24) = -0.980$, $p < 0.001$), and CTT discrimination (R) was significantly, positively associated with IRT discrimination "a" ($r(24) = 0.91$, $p < 0.001$).

c. IRT ability estimates for each dimension were also significantly associated with CTT total-correct scores for that dimension. Clinical epidemiology theta estimates were related to raw scores ($r(145) = 0.98$, $p < 0.001$), and statistics theta estimates were significantly associated with their raw scores ($r(145) = 0.98$, $p < 0.001$).

8. *Gather preliminary construct validity evidence for the BACES assessment by using known-groups validity comparisons*.
    a. Known-groups comparisons found no evidence of a significant difference between the sexes, among years in residency, or in those with previous exposure to epidemiology, or previous exposure to EBM.
        i. These findings contradict those found by Windish et al. (2007) who listed male sex as a significant contributor to performance.
    b. Those who reported previous biostatistics exposure (M = 5.29, SD = 2.47) performed significantly better than those who did not take a course (M = 4.39, SD = 1.97) on the statistics testlet raw score $t(106.78) = -2.271$, $p = 0.025$.
    c. Additional data will be needed to better assess the sensitivity of BACES items to demographic differences.

BACES Results – Alignment with Previous Research

This section describes the ways in which the results from the BACES assessment align themselves with findings from previous research. The numerous similarities can be summarized in three key areas which are (1) resident performance results, (2) item construction elements, and (3) sources for item content validity evidence.

The body of research on BEK has tested graduate medical students for measures of their competency for over 30 years, and each has consistently shown a low level of knowledge in these areas (e.g. Berwick et al., 1981; Windish et al., 2007; & Novak et al., 2006). Windish et al. (2007) found a mean score of only 41.4% and Novak et al. (2006) found an average of only 40%. When compared to the BACES assessment results, resident scores matched these previous studies with a mean raw score for the original 30 items was only 13.65 out of 30 (45.5%).

The second primary similarity between the BACES instrument and previous BEK instruments is the structure of the assessment itself. The BACES assessment used unique clinical vignettes derived from common medical procedures and conditions to construct the items. Each item provided one of these clinical examples in a one-best-answer format, which is the preferred approach for writing high quality MCQs (Case & Swanson, 2002).

An expert review approach was used in this study to gather evidence of content validity similar to previous instruments (Enders, 2011). This consistency was bolstered by using several existing instruments as a starting point for item construction. Moreover, this study, like others before it, used reviews of commonly used statistics in medical literature (Horwitz & Switzer, 2005; & Switzer & Horwitz, 2007) during the test blueprint process.

BACES Results – Expanding Upon Previous Instruments

While there have been studies on residents BEK for over three decades, the BACES assessment has addressed several of the shortcomings these previous items possessed. This next section explains three primary ways the BACES assessment has expanded upon existing instruments. The contributions have been organized by (1) strengthening the psychometric rigor of assessing BEK, (2) item-writing improvements, and (3) filling content gaps.

The most dramatic addition to previous instruments was using an IRT approach to instrument development, which sets the stage for more generalizable measurements in the future. The BACES assessment was the first study of BEK in GME to make the psychometric integrity of the instrument its top priority as opposed to residents' performance. Until this point, very brief discussions of psychometric properties were included with previous instruments (Enders, 2011). Additionally, the use of CTT item analyses for these instruments has muddied the ability to separate their psychometric properties from the samples on which they were originally tested

(Hays et al., 2000). It has been shown in previous research (Fan, 1998; Hays et al., 2000; Stage, 1998; Xu & Stone, 2011) as well as within this study, that the difficulty, discrimination, and estimates of ability in CTT are near identical to those from IRT; however, the parameters generated from CTT are so dependent upon the sample from which they are taken that the exact same instrument may look completely different in the second sample (DeMars, 2010). The BACES assessment, in contrast, fit a 2PL IRT model to the item response data, and the parameters that were generated from that model can be easily tested in additional samples. Rather than drastically change across administrations, the IRT parameters ought to remain invariant (Stage, 1998; DeMars, 2010; de Ayala, 2009, Furr & Bacharach, 2008). This property allows for items to be broken apart, rearranged, and reassembled into new test versions without losing their accuracy or consistency in estimating person trait levels. In other words, the BACES items could be broken up into smaller tests or topics depending on the needs of the instructor.

On a smaller scale, this study has addressed several common flaws existing BEK instruments contained in terms of best item-writing practices. Guidelines identified by assessment experts in health education (Case & Swanson, 2002) and higher education (Suskie, 2009) were applied to existing BEK instruments for this study. Unfocused stems and item dependencies were the two most commonly seen writing errors among the existing instruments. To review, an unfocused stem is one that fails to give the respondent enough information to answer correctly (De Champlain, 2010) while an item dependency occurs when the answer to one item directly influences the answer to another item through salient response options or a common example, vignette, etc. (DeMars, 2010). The BACES assessment was developed to minimize these two flaws in particular by using a unique case vignette or example for each

question. Item independence was confirmed through meeting the IRT assumption of local independence, which specifically tests for such interlocking items (DeMars, 2010).

One of the other important purposes for developing the BACES assessment was to fill content gaps that had been identified in previous BEK instruments. One such gap was noted by Enders (2011) who specifically concluded that more emphasis needed to be placed on within-subjects research designs and analysis. Item 26 was added to the BACES assessment to address this gap by asking residents to correctly identify a scenario where a within-subjects analysis of variance would be used. According to the IRT parameters, this item was considered to be one of the most difficult ones (b = 1.28), and only 27% of the residents correctly answered this question. 30 of those in the lowest 25% chose option "B" "Independent (unpaired) t-test."

Study Limitations

While these preliminary results from developing the BACES assessment have been positive, it is important to note three key limitations to the design, conduct, and interpretation of results. The most important of these limitations is that the interpretation of these results are intended to be preliminary only, and any causal conclusions based on these results would be inappropriate without additional studies. Specifically, the invariance of IRT parameters is only possible if the IRT model fits the data (DeMars, 2010). Although results showed a 2PL model fit this sample, the estimates may (and likely will) change as a larger sample of residents is tested. The corollary to the first limitation is that the sample size used for the BACES data was smaller than what would be preferred for IRT analysis. Simulation studies have shown between 100 and 500 participants is an adequate number for estimating a 2PL model (Lord, 1983; Drasgow, 1989); therefore, the 147 residents in this study was a relatively small sample size, which could

have produced a higher overall standard error (Orlando & Reeve, 2007). Third and finally, the non-randomized, cross-sectional design used in this study permits a great deal of possible sampling error, which could have impacted the results of the study. For example, departments self-selected to participate in the study, and the administrations were held in a rather uncontrolled environment (i.e. residents coming and going frequently). Since a completely controlled testing situation was not possible, there may be an element of cheating or lack of motivation impacting the BACES results.

Conclusions and Implications

The final section of this chapter shares the implications of this study to graduate medical students, GME faculty, and future research. The section ends with a brief, overall conclusion on the study as a whole.

Implications of the BACES Assessment Results for Graduate Medical Students

The BACES assessment results have broad implications to the graduate medical student population, which begin with the test construction itself. Each item for the assessment was specifically crafted to mimic a realistic clinical or literature example. The content for these examples was derived from broadly applicable medical and surgical conditions while at the same time incorporating many of the most commonly used statistics in major medical journals (e.g. Horwitz & Switzer, 2005; & Switzer & Horwitz, 2007). The BACES items were also given to four content experts in medicine, public health, surgery, higher education assessment, clinical pharmacy, and MCQ test development to ensure each item was a valid self-assessment for residents' ability.

This assessment also holds promise as a valuable possible source of information for graduate medical students. These two topics, statistics in particular, have a history of being self-

reported gaps in physician confidence with as little as 17.6% of respondents reporting their training as adequate (West & Ficalora, 2007; Reznik et al., 1987; & Swift et al., 2009). At the same time, nearly 80% of respondents to one of these surveys indicated that knowledge of statistics was important (Swift et al., 2009). To answer this need, the BACES assessment added a detailed answer key, which does not appear with any of the other existing assessment instruments. After completing the self-assessment, the examinee is able to receive immediate feedback on their success while at the same time getting a thorough explanation as to *why* their particular answer choice was correct or not. Although it was beyond the scope of this study, additional investigation must be done, possibly using qualitative methods, to look for evidence on how, if at all, the descriptive answer key was used by both examinees and instructors.

Implications of the BACES Assessment Results for GME Educators

The GME educator community also stands to gain from the BACES assessment. Researchers have shown that the methods by which BEK is taught at the GME level varies considerably (Green, 2001; Green, 1999); Rao, 2008). At the same time, the ACGME requires these topics to be addressed in their core competencies (Hatala & Guyatt, 2001; ACGME, 2013_a). The BACES assessment, its blueprint, and its descriptive answer key could all be used by GME faculty to plan their BEK curricula. Faculty with a high degree of knowledge in these areas may benefit from reviewing the content of the assessment because it was developed from what their residents will commonly encounter in the literature. On the other hand, less experienced instructors may find that relying on the descriptive answer key for its detailed explanations is helpful for their own education as well as their residents'.

The second important implication of these results for GME educators revolves around the possibility of a flexible, psychometrically rigorous assessment of BEK. Should the IRT parameters for BACES items be further tested, it would be possible for users to break the

assessment up by topic, test only those topics they are teaching, and *not* lose the reliability, difficulty, or discrimination of those items.

Implications of the BACES Assessment Results for Future Research

The entirety of this study could be considered a preamble to a long road of future research ahead. Now that *preliminary* evidence for content validity, construct validity, and item parameters have been estimated, it is up to future research to confirm them. Specific steps that must be taken by future researchers include (1) modifying the problematic items found during this study; (2) generating additional items to ensure the assessment includes *all* relevant topics; (3) administer the improved instrument to a far larger population; (4) test the stability of item parameters found in this study within the much larger sample; and (5) continue to investigate the BACES items for construct validity evidence and differential item functioning. These five steps will keep the development process moving forward, and ultimately create a much stronger and valid instrument.

With regards to additional items, it would be beneficial to work towards developing a much larger bank of items that could include several different items per concept. Also, this bank would include some of the concepts not covered in the BACES items such as dealing with clustered data, and interpreting values for absolute risk reduction or attributable risk. One possible approach to writing additional items would be to make a large-scale call to other GME educators to participate in writing items for the item bank. Each participant would receive the current BACES assessment items as well as instructions for how to write MCQs in the correct format. Such an approach would greatly increase the volume of items for future iterations of the assessment as well as generate more possible buy-in for larger samples of residents by engaging faculty across the country. Eventually, the goal would be to use what was started in this study as

the basis for a new computer-adaptive self-assessment that residents and GME educators access to study specific topics on their own time.

Final Summary

The BACES assessment was designed with a focus on *increasing* medical residents' competency in consuming the clinical epidemiologic and biostatistical methods in the medical literature as opposed to simply *diagnosing* it. In pursuit of this goal, the BACES assessment was developed and tested for its preliminary content and construct validity as well as its individual item parameters. In contrast to previous studies, this study was developed using an IRT approach, and its results have paved the road for a flexible yet psychometrically rigorous instrument for measuring the biostatistical and clinical epidemiologic knowledge of graduate medical students.

List of References

Accreditation Council for Graduate Medical Education (ACGME), (2013_a). *ACGME Common Program Standards*. Retrieved from http://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramRequirements/CPRs2013.pdf

Accreditation Council for Graduate Medical Education (ACGME), (2013_b). *List of ACGME Accredited Programs and Sponsoring Institutions*. Retrieved from http://www.acgme.org/ads/public/

Ahmadi-Abhari, S., Soltani, A., & Hosseinpanah, F. (2008). Knowledge and attitudes of trainee physicians regarding evidence-based medicine: a questionnaire survey in Tehran, Iran. *Journal of evaluation in clinical practice*, *14*(5), 775–9. doi:10.1111/j.1365-2753.2008.01073.x

Angelo, T., & Cross, K. C. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers.* San Francisco: Jossey-Bass .

Brunnquell, A., Degirmenci, U., Kreil, S., Kornhuber, J., & Weih, M. (2011). Web-based application to eliminate five contraindicated multiple-choice question practices. *Evaluation & the health professions*, *34*(2), 226–38. doi:10.1177/0163278710370459

Berwick, D. M., Fineberg, H. V, & Weinstein, M. C. (1981). When doctors meet numbers. *The American Journal of Medicine*, *71*(6), 991–998. Retrieved from http://www.sciencedirect.com/science/article/pii/0002934381903259

Case, S., & Swanson, D. (2002). *Constructing written test questions for the basic and clinical sciences* (3^{rd} ed.). Philadelphia: National Board of Medical Examiners (NBME).

Cheatham, M. L. (2000). A structured curriculum for improved resident education in statistics. *The American surgeon*, *66*(6), 585–8.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. doi:10.1037/1040-3590.7.3.309

De Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory.* New York: Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education*, *44*(1), 109–17. doi:10.1111/j.1365-2923.2009.03425.x

DeMars, C. (2010) *Item Response Theory*. New York: Oxford University Press.

DeVellis, R.F. (2012). *Scale Development: Theory and Applications* (3rd ed.). Thousand Oaks: Sage.

Downing, S., & Baranowski, R. (1995). Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. *Applied measurement …*. Retrieved from http://www.tandfonline.com/doi/full/10.1207/s15324818ame0802_5

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education : Theory and Practice*, *10*(2), 133–43. doi:10.1007/s10459-004-4019-5

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, 13(77). doi:10.1177/014662168901300108

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of life research : an*

international journal of quality of life aspects of treatment, care and rehabilitation, *16 Suppl 1*, 5–18. doi:10.1007/s11136-007-9198-0

Emerson, J. D., & Colditz, G. A. (1983). Use of statistical analysis in the New England Journal of Medicine. *The New England journal of medicine*, *309*(12), 709.

Enders, F. (2011). Evaluating Mastery of Biostatistics for Medical Researchers: Need for a new assessment tool. *Clinical and translational science*, *4*(6), 448–454. doi:10.1111/j.1752-8062.2011.00323.x.Evaluating

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. doi:10.1177/0013164498058003001

Fritsche, L., Greenhalgh, T., Falck-Ytter, Y., Neumayer, H.-H., & Kunz, R. (2002). Do short courses in evidence based medicine improve knowledge and skills? Validation of Berlin questionnaire and before and after study of courses in evidence based medicine. *BMJ*, *325*(7376), 1338–1341. Retrieved from http://www.bmj.com/content/325/7376/1338.abstract

Furr, M.R., & Bacharach, V.R., (2008) *Psychometrics: An Introduction*. Thousand Oaks: Sage.

Guyer, R., & Thompson, N.A. (2012). *User's Manual for Xcalibre item response theory calibration software, version 4.2*. St. Paul MN: Assessment Systems Corporation.

Green, M. (1999). Graduate medical education training in clinical epidemiology, critical appraisal, and evidence-based medicine: a critical review of curricula. *Academic Medicine*. Retrieved from http://journals.lww.com/academicmedicine/Abstract/1999/06000/Graduate_medical_education_training_in_clinical.17.aspx

Green, M. L. (2001). Evidence-based medicine training in graduate medical education: past, present and future. *Journal of Evaluation in Clinical Practice*, *6*(2), 121–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10970006

Hatala, R., & Guyatt, G. (2002). Evaluating the Teaching of Evidence-Based Medicine. *JAMA: The Journal of the American Medical Association*, *288*(9), 1110–1112.

Hays, R., Morales, L., & Reise, S. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1815384/

Hellems, M. a, Gurka, M. J., & Hayden, G. F. (2007). Statistical literacy for readers of Pediatrics: a moving target. *Pediatrics*, *119*(6), 1083–8. doi:10.1542/peds.2006-2330

Holt, K. D., Miller, R. S., & Nasca, T. J. (2010). Residency Programs' Evaluations of the Competencies: Data Provided to the ACGME About Types of Assessments Used by Programs. *Journal of Graduate Medical Education*, *2*(4), 649–55. doi:10.4300/JGME-02-04-30

Horton, N., & Switzer, S. (2005). Statistical methods in the journal. *New England Journal of Medicine*, 1977–1979. Retrieved from http://www.nejm.org/doi/full/10.1056/NEJM200511033531823

Jekel, J. F. (1991). Understanding Biostatistics. *The Yale Journal of Biology and Medicine*, *64*(5), 545–546.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine : Journal of the Association of American Medical Colleges*, *77*(2), 156–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11841981

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data* (Doctoral dissertation, University of Illinois).

Lord, F. (1952). A theory of test scores. *Psychometric Monographs, 7*, x, 84.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika,* 48(2).

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lorenzo-Seva, U., & Ferrando, P. J. (2012). TETRA-COM: a comprehensive SPSS program for estimating the tetrachoric correlation. *Behavior Research Methods*, *44*(4), 1191–6. doi:10.3758/s13428-012-0200-6

Looney, S. W., Grady, C. S., & Steiner, R. P. (1998). An update on biostatistics requirements in U.S. medical schools. *Academic Medicine*. doi:10.1097/00001888-199801000-00018

Morreale, M. K., Balon, R., & Arfken, C. L. (2012). Teaching statistical literacy to psychiatry residents: a pilot study of training directors. *Academic Psychiatry : The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, *36*(2), 152–3. doi:10.1176/appi.ap.11070133

Mulvihill, M. N. (n.d.). Faculty development and resident training in epidemiology and biostatistics. *The Mount Sinai Journal of Medicine, New York*, *48*(4), 350–2.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality. *Journal of Educational and Behavioral Statistics*, *18*(1), 41–68. doi:10.3102/10769986018001041

Novack, L., Jotkowitz, A., Knyazer, B., & Novack, V. (2006). Evidence-based medicine: assessment of knowledge of basic epidemiological and research methods among medical

doctors. *Postgraduate Medical Journal*, *82*(974), 817–822. Retrieved from http://pmj.bmj.com/content/82/974/817.abstract

Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200–219. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1994.tb00443.x/full

Rao, G. (2008). Physician numeracy: essential skills for practicing evidence-based medicine. *Family Medicine*, *40*(5), 354–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18465286

Reed, J. F., Salen, P., & Bagher, P. (2003). Methodological and statistical techniques: What do residents really need to know about statistics? *Journal of Medical Systems*, *27*(3), 233–238. doi:10.1023/A:1022519227039

Reznick, R., Dawson-Saunders, E., & Folse, J. (1987). A rationale for the teaching of statistics to surgical residents. *Surgery*. Retrieved from http://europepmc.org/abstract/MED/3576452

Rigby, A. S., Armstrong, G. K., Campbell, M. J., & Summerton, N. (2004). A survey of statistics in three UK general practice journal. *BMC Medical Research Methodology*, *4*(1), 28. doi:10.1186/1471-2288-4-28

Sackett, D., & Rosenberg, W. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal.* Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2349778/

Sahai, H. (1999). Teaching biostatistics to medical students and professionals: Problems and solutions. *International Journal of Mathematical Education in Science and Technology*, (April 2013), 37–41. Retrieved from http://www.tandfonline.com/doi/full/10.1080/002073999287978

Shadish, Cook, & Campbell. (2002). *Experimental and Quasi Experimental Designs for Generilized Causal Inference* (pp. 33–63). Boston: Houghton Mifflin.

Stage, C. (1998). A comparison between item analysis based on Item Response Theory and Classical Test theory. A study of the SweSAT Subtest WORD. *Educational Measurement*. Retrieved from http://www.sprak.umu.se/digitalAssets/59/59551_enr2998sec.pdf

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical teacher*, *28*(6), 566–8. doi:10.1080/01421590600711153

Stout, W., Froelich, A., and Gao, F. (2001). Using Resampling Methods to Produce an Improved DIMTEST procedure. In Boomsma, A., van Duijn, M.A.J., Snijders, T.A.B. (Eds.) Essays on item response theory. New York: Springer-Verlag.

Suski, L., & Banta, T. (2009). *Assessing Student Learning: A Common Sense Guide.* San Francisco: Jossey-Bass.

Swift, L., Miles, S., Price, G. M., Shepstone, L., & Leinster, S. J. (2009). Do doctors need statistics ? Doctors' use of and attitudes to probability and statistics, (November 2008), 1969–1981. doi:10.1002/sim

Switzer, S., & Horton, N. (2007). What your doctor should know about statistics (but perhaps doesn't...). *Chance*, 17–21. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:What+Your+Doctor+Should+Know+about+Statistics+(but+Perhaps+Doesn?t...)#0

Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics, (6th ed.).* Boston: Allyn and Bacon.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*(2), 191–208. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1989.tb00328.x/abstract

Waller, J., Ostini, R., Marlow, L. a V, McCaffery, K., & Zimet, G. (2013). Validation of a measure of knowledge about human papillomavirus (HPV) using item response theory and classical test theory. *Preventive Medicine*, *56*(1), 35–40. doi:10.1016/j.ypmed.2012.10.028

Wang, Q., & Zhang, B. (1998). Research design and statistical methods in Chinese medical journals. *JAMA : The Journal of the American Medical Association*, *280*(3), 283–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9676683

Weiss ST, Samet JM. An assessment of physician knowledge of epidemiology and biostatistics. *Journal of Medical Education* 1980;55(8): 692-7.

West, C. P., & Ficalora, R. D. (2007). Clinician Attitudes Toward Biostatistics. *Mayo Clinic Proceedings*, *82*(8), 939–943. Retrieved from http://www.mayoclinicproceedings.com/content/82/8/939.abstract

Windish, D. M., Huot, S. J., & Green, M. L. (2007). Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *JAMA: The Journal of the American Medical Association*, *298*(9), 1010–1022. Retrieved from http://jama.ama-assn.org/content/298/9/1010.abstract

Wise, J.M. (n.d.). *Item Analysis: Techniques to Improve Items and Instruction.* Retrieved from distance.fsu.edu/docs/assessment/ItemAnalysis.ppt

Xu, T., & Stone, C. a. (2011). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453–468. doi:10.1177/0013164411419846

Yen, W. M. (1981). Using Simulation Results to Choose a Latent Trait Model. *Applied Psychological Measurement*, *5*(2), 245–262. doi:10.1177/014662168100500212

Yu, C. H., Popp, S. O., Digangi, S., & Jannasch-pennell, A. (2007). Assessing unidimensionality : A comparison of Rasch Modeling , Parallel Analysis , and TETRAD. *October*, *12*(14).

Zhang, J. (1996). *Some fundamental issues in item response theory with applications*. (Doctoral dissertation, University of Illinois).

Zhang, J. and Stout, W. (1999).The Theoretical DETECT Index of Dimesionality and Its Application to Approximate Simple Structure. *Psychometrika, 64*, 213-249

Appendix

Appendix A
Expert Review Rubric for Content Validity Evidence
BACES SELF-ASSESSMENT ITEM REVIEW RUBRIC

Thank you again for being willing to participate in the expert review process of my dissertation research developing the Biostatistics and Clinical Epidemiology Skills (BACES) self-assessment. This document contains a rubric to guide you through the review process for each item as well as a score sheet to provide your feedback.

INSTRUCTIONS: This first page asks broad questions about the overall format of the assessment including the ***instructions, length***, and ***item order*** (i.e. "flow). After these first questions, please use the rubric on Page 2 to review the components of each item (the vignette, stem, response options, and content), and rate them on the worksheet I have created on Page 3. Finally, Page 4 contains an area for you to provide any additional comments or suggestions you may have for improving the instrument before it is tested.

Tip for being efficient! The rubric may seem daunting at first, but it can be greatly simplified by reading from the top row where the numbers show the general rating for each component with 1 = "Heavy revisions necessary" and 4 = "Keep as is, no revisions necessary." The descriptions in each cell of the rubric are simply to assist you in the review process should you be confused or need more clarification. It may also be helpful to print the rubric out and have it at your side while reading through the items rather than flipping back-and-forth.

Thank you again for your willingness to lend your expertise, and happy reviewing!

What comments or suggestions do you have for the ***instructions***? Will the participant know what is expected of them when they are given the instrument? >
Is the ***length*** of the instrument appropriate? Are there enough items to sufficiently address biostatistics and clinical epidemiologic research methods? >

What comments or suggestions do you have for the way in which the items are ***ordered***?

>

Component	1 (Heavy Revisions Necessary)	2 (Some Revisions Necessary)	3 (Minimal Revisions Necessary)	4 (Keep as is, no revisions necessary)
Vignette (the example or narrative part of the question)	Vignette is missing important information that is necessary for answering the question. There is irrelevant or "trick" information that would prevent a student who knows the concept from answering the question. There are clues or hints that would help a student with no knowledge of the concept to answer the question correctly.	Vignette does not clearly provide the information necessary to answer the question, and contains at least one of the following: • some irrelevant or "trick" information; • clue or hint to the correct answer • grammatical or content-related errors	Vignette clearly provides the information necessary to answer the question, but contains one of the following: • some irrelevant or "trick" information; • clue or hint to the correct answer • grammatical or content-related errors	Vignette clearly provides the information necessary to answer the question. The length for the vignette is appropriate. There is no irrelevant or "trick" information, clues or hints to the correct answer, or grammatical errors.
Item Stem (the question itself)	Item stem is unclear, does not ask a question, or asks multiple questions, or the stem does not follow naturally from the vignette (i.e. it fits logically with the vignette presented).	Item stem needs to be clarified to ask a clearer question, there are grammatical clues in the stem that connect to the correct answer, or the stem does not follow naturally	The item stem needs minor clarification, but it asks a single, relevant question. There are no grammatical clues to the correct answer, and the stem follows	The item stem clearly asks a single, relevant question that does not provide grammatical clues to the correct answer. The stem follows naturally from the vignette.

		from the vignette.	naturally from the vignette.	
Response Options (A – D)	The response options are *not* clearly written, the keyed answer is factually inaccurate, *or* there are significant errors in any of the following: • No clear "Order" of correctness • Grammatical connections to stem and/or vignette • Uneven length • Response option links to other items in the instrument.	Some response options are clearly written, and the keyed answer is the correct option, but there are errors in least two of the following: • No clear "Order" of correctness • Grammatical connections to stem and/or vignette • Uneven length • Response option links to other items in the instrument.	All response options are clearly written, but contain minor errors in any of the following: • No clear "Order" of correctness • Grammatical connections to stem and/or vignette • Uneven length • Response option links to other items in the instrument. • Grammatical errors	All response options are clearly written, and can be arranged in order of "correctness" with the keyed answer as the single best option. There are no grammatical links from the response set to either the vignette or stem. All options are of similar length, and do not link themselves to other items in the instrument.
Content (the medical and statistical concepts / terminology used in the item)	The content chosen for the question is not relevant to research methods and/or statistics. The question is not appropriate for the resident population. (OR) There are significant errors in the interpretation of medical terminology, or plausibility that could affect responses.	The content chosen for the question is relevant to research methods and/or statistics, but may not be appropriate for the resident population. (OR) There are errors in the interpretation of medical terminology, or plausibility that could affect responses.	The content chosen for the question is relevant to research methods and/or statistics, and is appropriate for the resident population. (OR) There are minor errors in the interpretation of medical terminology, or plausibility that need revision.	The content chosen for the question is relevant to research methods and/or statistics, and is appropriate for the resident population. (AND) There are no errors in the interpretation of medical terminology, or plausibility that need revision.

ITEM REVIEW WORKSHEET

1. Please use the rubric on the previous page to rate each item in the table below (each is rated between 1 = "Heavy revisions necessary" and 4 = "Keep as is, no revisions necessary").
2. After you have rated each component, please rate the overall quality of the item from 1 = "Very poor" to 5 = "Excellent."

Item #	Item Component				
	Vignette	Item Stem	Response Options	Content	Overall
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					

28					
29					
30					

ADDITIONAL COMMENTS

Please use this page to write any additional comments have about specific items or components of the BACES instrument (and example is given). Thank you!

>Example Q1: "Rearrange the order of response options to make it more logical to the reader".

Appendix B

Final BACES Assessment Form "A"

BIOSTATISTICS AND CLINICAL EPIDEMIOLOGY SKILLS (BACES) ASSESSMENT

Research Methods and Statistics Knowledge Self-Assessment

Today you are being asked to participate in an ongoing research project conducted by Patrick Barlow, a PhD candidate in Evaluation, Statistics, & Measurement at the University of Tennessee, Knoxville.

For this study, you will be asked to answer several multiple-choice questions about your statistics and research methods skills. Your participation in this project is completely voluntary, and you may choose to decline participation at any point. Please take the next few minutes to answer the following questions; if you are unsure of a particular question, you should give your best guess. Although more than one response option may be plausible, each question has a *single best answer*.

Your responses and total scores on this assessment will remain confidential, and your participation in today's study will not affect your standing with your institution in any way. After everyone has completed this self-assessment, you will receive a copy and description of each question, so that you may use it to study in the future.

Thank you!

1) An internal medicine resident is interested in looking at the effect of adherence to ribavirin and interferon therapy on virologic response in patients with hepatitis C (HVC). His team uses a national case registry to identify and follow all HVC patients with either 50% or 100% medication adherence between January 2003 and June 2008. He then analyzes both early and sustained virologic response across two different adherence groups.

This is an example of:

a. A randomized control trial
b. A case-control design
c. A nested case-control design
d. A retrospective cohort design

2) A researcher is investigating the association between a patient's history of colonoscopy and subsequent risk for colorectal cancer (CRC). She obtains the medical history for 1688 CRC patients and 1932 healthy patients, and finds that colonoscopy is associated with 77% lower risk of having CRC with OR=.23 (95% CI, .019 to .27).

This is an example of:

a. A longitudinal design
b. A case-control design
c. A case series design
d. A cross-sectional design

3) A research team conducts a randomized, prospective study to compare a 1-day, 4-drug regimen with a 7-day, 3-drug regimen in their efficacy at eradicating a particular infection. They are hypothesizing that the two groups would not differ in the proportion of patients whose infection was eradicated by a clinically meaningful amount (+/- 15% eradication percentage).

Which of the following statistical analysis would be the *most* appropriate to way to test this hypothesis?

a. Chi-Square Test of independence
b. Independent samples *t*-Test
c. Non-inferiority test for proportions
d. Equivalence test for proportions

4) A 2005 study used The Nurses Health Study (NHS) II population to assess the association between self-reported diagnosis of psoriasis and risk for diabetes and hypertension. In the article, the authors list age, height, Body Mass Index (BMI), smoking status, alcohol intake, and physical activity as covariates in their analyses.

The word "covariates" most likely indicates that the researchers…

a. controlled these variables as possible confounders in their analysis
b. excluded some patients that based on these variables
c. matched patients in each group according to their values on these variables
d. identified an interaction between these variables, diabetes, and hypertension

5) A pharmacist studies the effect of treatment intensification in type II diabetes patients. He investigates the likelihood of patients being readmitted within 90 days for whose treatment was intensified versus those whose treatment was maintained. The results showed an odds ratio (OR) of 0.26 (95% CI = .08 to .82).

This OR would best be interpreted as:

a. Patients whose treatment was intensified were 74% less likely to be readmitted versus those whose treatment was maintained.
b. Patients whose treatment was intensified were 26% more likely to be readmitted versus those whose treatment was maintained
c. Patients whose treatment was maintained were 74% less likely to be readmitted versus those whose treatment was intensified
d. Patients whose treatment was maintained were 26% more likely to be readmitted versus those whose treatment was intensified

6) A third-year family medicine resident is developing a pre-intervention survey to give to his clinic patients before beginning a smoking cessation intervention. He wants to ask the participants about their smoking history, so he can decide which patients are most in need of the additional service.

What scale of measurement would give the resident the most precise data about his patients' smoking history?

a. Discrete
b. Ordinal
c. Interval
d. **Ratio**

7) Which of the following is an effective way to *increase* the statistical power of a study?

a. Increase β (beta) from .20 to .40
b. Decrease α (alpha) from .05 to .01
c. Increase the sample size from 100 to 150
d. Use an ordinal scale rather than an interval scale of measurement

8) You submit an article that looks at the difference between two different statin regimens in their ability to lower LDL cholesterol. One of the journal reviewer claims your study results are likely a Type I error.

The researcher is most likely claiming that your team…

a. concluded there is a statistically significant difference between the statin regimens when in fact there is not a difference.
b. did not have enough patients in the study to show the difference between the two statin regimens.
c. concluded there is not a statistically significant difference between the two statin regimens when in fact there is a difference.
d. did not control for possible confounding variables when testing the difference between the two statin regimens.

9) A medical school faculty member wants to know residents' attitudes towards research design and statistics. He administers a nationwide survey to look at these concepts. This faculty member's study is an example of a(n):

a. Longitudinal design
b. Ecological design
c. Prospective cohort design
d. Cross-sectional design

10) Consider the following table (below) from an article on risk factors for acute kidney injury (AKI).

Table 1. Descriptive Statistics of Study Sample

Variable	Frequency (%) or Mean (SD)	
	Standard	Extended
Sex		
Male	38 (32.2%)	31 (26.3%)
Female	21 (17.8%)	28 (23.7%)
Race		
White	55 (46.6%)	58 (49.2%)
Black	4 (3.4%)	1 (1.7%)
Taking Medications		
NSAID	24 (64.9%)	13 (22.4%)
Diuretics	35 (47.3%)	29 (39.2%)
Age	60.88 (18.58)	57.32 (17.90)
Height (in).	67.80 (3.87)	67.12 (4.19)
Weight (kg).	84.27 (27.67)	80.39 (27.21)

In this case, "NSAID" is considered to be a(n):

a. Ordinal variable
b. Nominal variable
c. Interval variable
d. Ratio variable

11) A surgical resident conducts a study looking at a mouse model for surgical site infections (SSI) and local anesthetic use. She finds that mice injected with a lidocaine/marcaine mixture had a significantly lower risk of SSI compared with those injected with saline with a relative risk (RR) of RR = .45 (95% CI = .25 - .89).

What is the effect size of her analysis?

a. 45%
b. 55%
c. 64%
d. 95%

12) An OBGYN resident is conducting a cross-sectional study to look at the relationship between contraceptive medication prices and typical income for a given area. She suspects that family income will not be normally distributed.

Which of the following measures of central tendency should she use in order to accurately represent typical family income?

a. Median
b. Mean
c. Mode
d. Range

13) The clinical trial results of an investigational diagnostic test report the sensitivity and specificity values for the test as 88% and 71%, respectively.

From this statement, you can conclude that the new test was…

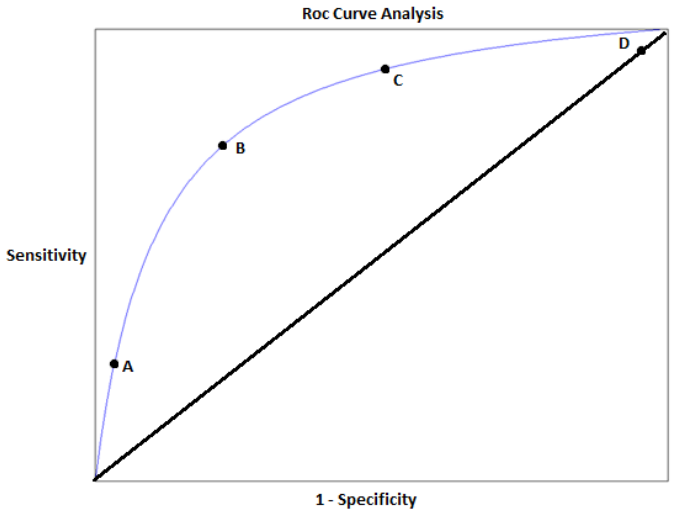
a. is 88% effective at detecting negative disease states, and 71% effective at detecting positive disease states.

b. is 88% effective at detecting true positive disease states, and 71% effective at detecting true negative disease states.

c. is 88% effective at detecting positive disease states, and 71% effective at detecting negative disease states.

d. is 88% effective at detecting true negative test results, and 71% effective at detecting true positive test results.

14) A study in which the researcher aimed to develop a new screening test procedure for pancreatic cancer gave the following Receiver Operating Characteristic (ROC) curve to show the sensitivity and specificity of the new test.



Of the coordinates labeled in the ROC curve, which would be most likely to give the researcher the maximum sensitivity and specificity values for her new screening test?

a. Point "A"
b. Point "B"
c. Point "C"
d. Point "D"

15) A group of first-year residents are given a review of proper technique for chest tube placement, and then are observed as they perform the task in the simulation center. Three faculty researchers rate the residents' performances using a skills checklist before and after they work with a skills coach.

What would be the most important type of *reliability* evidence for this research study?

a. Test-retest reliability
b. Internal consistency reliability
c. Inter-rater reliability
d. Split-half reliability

16) A nuclear medicine resident is comparing the average heart and lung uptake values (Standardized Uptake Value) between two different tracers, Rb82 and N13. He gathers data from 50 patients who had Rb82 and 51 patients who had N13, and compares the mean uptake in both groups.

Which statistical test would the resident likely use?

a. Multiple regression analysis
b. Paired-samples *t*-test
c. Chi-square test
d. Independent *t*-test

17) A research team conducted a prospective cohort study of 88,757 women to investigate the association between high dietary fiber intake and colorectal cancer (CRC). They compared the likelihood of developing CRC over the 16-year follow-up period among five quintiles of dietary fiber intake.

What would be the most *accurate* measure of association to use in this situation?

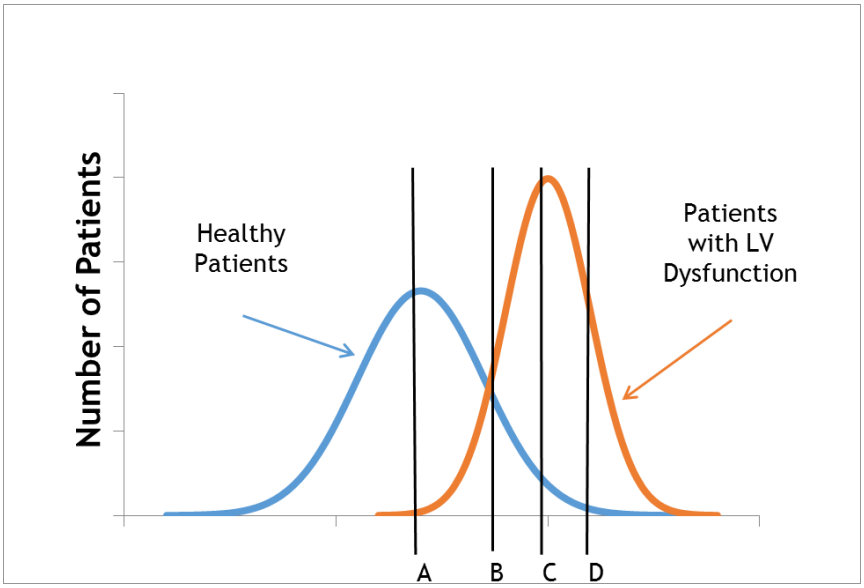
a. Odds ratio
b. Relative risk
c. Incidence ratio
d. Absolute risk

18) A clinical trial randomly assigns 3202 patients to receive one of two possible treatments for acute symptomatic pulmonary embolism. The authors concluded that their experimental treatment was statistically significantly non-inferior ($p = 0.03$) at preventing clinically relevant bleeding within 10% of the standard of care treatment.

The authors are concluding that their experimental treatment…

a. no more than 10% less effective than the standard of care
b. 10% more effective than the standard of care
c. no more than 10% more or less effective than the standard of care
d. 10% less effective than the standard of care

19) A study of using Brain Natriuretic Peptide (BNP) to screen for Left-Ventricular Dysfunction considers four possible cut-off values for classifying a patient as testing "positive" for dysfunction.



Which of the four values would provide the researchers with a test that would minimize *both* false positive and false negative test results?

a. Cut-Off A
b. Cut-Off B
c. Cut-Off C
d. Cut-Off D

20) Two-hundred-and-thirty steelworkers with hypertension participated in a randomized trial to see if adherence to antihypertensive drug regimens could be improved. The men were randomly allocated to see either their own family doctors outside of work-hours or company physicians during work; they were also randomly allocated to receive or not receive a hypertension educational program.

Which research design best describes this scenario?

a. 2x2 factorial trial
b. Double cohort design
c. 2x2 cross-over trial
d. Prospective cohort design

21) A study of middle-aged, obese individuals examined the association between obstructive sleep apnea syndrome and hypertension. They evaluated the patients' Apnea-hypopnea index score, age (in years), sex, and neck circumference (in centimeters) as possible predictors for increased systolic and diastolic blood pressure.

Which statistical approach would be the most appropriate way to address this objective?

a. Multiple linear regression
b. Multiple logistic regression
c. Multiple Cox regression
d. Multiple Ordinal regression

22) A nested case-control study was conducted to evaluate whether increases in the inflammatory markers interleukin 6 (IL-6) and C-reactive protein (CRP) were associated with an increased risk of type II Diabetes in otherwise healthy middle-aged women. They found that women in the third (RR = 8.7, 95% CI = 3.6 – 21.0) and fourth (RR = 15.7, 95% CI – 6.5 – 37.9) quartile of CRP were at a significantly higher risk for Type 2 Diabetes than those in the first quartile.

Given these results, what can be said about the researchers' estimates for the association between CRP and Type 2 Diabetes?

a. The estimate for women in the third quartile of CRP is less accurate than the estimate for the fourth quartile.
b. 95% of the estimates for risk of Type 2 Diabetes in the fourth quartile of CRP would be RR = 15.7.
c. 95% of the estimates for risk of Type 2 Diabetes in the third quartile of CRP would be RR = 8.7.
d. The estimate for women in the third quartile of CRP is more accurate than the estimate for the fourth quartile.

23) A group of 362 elderly patients were followed over 50 weeks to collect data on risk factors for accidental falls and injurious falls. The researchers reported that the incidence rate for falls was 45.5 per 1,000 person-months.

What would be the most appropriate way to interpret these results?

a. 45.5% of the 362 elderly patients fell during 1,000 months of follow-up
b. The 362 elderly patients fell 45.5 times during 1,000 months of follow-up
c. We expect 45.5 falls for every 1,000 months of follow-up
d. We expect 45.5% of elderly patients to fall for every 1,000 months of follow-up

24) You are interested in looking at how two different surgical procedures influence the length of patient survival (in days) when controlling for age, sex, and if the patient has a history of heart problems.

What statistical test you would choose to analyze this question?

a. Kaplan Meier analysis
b. Chi-Square test
c. Multiple Cox regression analysis
d. Multiple Logistic regression analysis

25) A retrospective study aimed to find risk factors for chronic exertion compartment syndrome (CECS). Two-hundred athletes who were evaluated at a local orthopedic clinic for lower leg pain were selected for the sample. The researchers then interviewed 100 athletes with CECS and 100 athletes without CECS about their sports training and overall lifestyle habits.

To which type of bias is this study most susceptible?

a. Misclassification bias
b. Recall bias
c. Experimenter bias
d. Medical surveillance bias

26) A cardiology fellow implements a 12-week protocol to investigate the impact of statin treatment regimens on the patients' low-density lipoprotein (LDL) cholesterol levels (mg/dL). At baseline, patients are randomly assigned to either continue their daily statin regimen or change to a three times per week regimen. After six weeks, the patients changed to the opposite regimen (e.g., daily changes to three times per week) for the remaining six weeks.

What would be the most effective way to statistically compare LDL cholesterol levels at baseline, six weeks, and twelve weeks?

a. Dependent (paired) t-test
b. Independent (unpaired) t-test
c. Between subjects analysis of variance (ANOVA)
d. Within subjects analysis of variance (ANOVA)

27) An interventional trial investigated the efficacy of antibiotic prophylaxis on reducing sepsis and mortality in patients with acute necrotizing pancreatitis (ANP). The researchers found the relative risk for sepsis to be RR = 0.69 (95% CI = 0.86 – 0.40). They would like to know what it would take to prevent one death from sepsis in ANP patients.

What measure of risk could the researchers use to answer this question?

a. Number needed to harm
b. Number needed to treat
c. Absolute risk
d. Attributable risk

28) Susan and John are neighbors who are both enrolled in opposite arms of a clinical trial investigating the efficacy of a new medication for patients with GERD (versus a placebo). John experiences a great relief in his heartburn symptoms, but Susan tells him she is feeling no better. John decides to offer Susan some of his medication because he figures she must be on the placebo.

What threat to internal validity of the study has John most likely increased as a result of his actions?

a. Attrition
b. Compensatory Rivalry
c. Demoralization
d. Diffusion

29) A pilot study sought to develop a suitable training model for laparoscopic appendectomy by using the uterine horns of three female pigs. After surgical preparation, ethanolamine oleate (EO) was injected into the uterine horn of each pig, which then simulated the inflamed human appendix. A critic of the study wrote a letter to the authors in which he cautioned them against generalizing their pig model to human subjects.

To what type of validity is the author of the letter most likely referring?

a. External
b. Internal
c. Ecological
d. Construct

30) An interdisciplinary patient intervention is initiated for all Type 2 diabetes patients at a local hospital. For 12 months, a pharmacist, general practitioner, and dietician meet with the patients for their regularly scheduled appointments, so that they may provide a more holistic approach to care. The researchers believe that there will be a significant decrease in the rate of hospitalizations when comparing 12 months before the intervention to 12 months after their intervention began.

How would the research team most accurately write their *null* and *alternative (research)* hypotheses for this study?

a. H_0: Rate prior to intervention = Rate after the intervention
H_1: Rate prior to intervention > Rate after the intervention

b. H_0: Rate prior to intervention = Rate after the intervention
H_1: Rate prior to intervention ≠ Rate after the intervention

c. H_0: Rate prior to intervention ≠ Rate after the intervention
H_1: Rate prior to intervention < Rate after the intervention

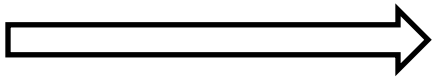
d. H_0: Rate prior to intervention = Rate after the intervention
H_1: Rate prior to intervention < Rate after the intervention

BIOSTATISTICS AND CLINICAL EPIDEMIOLOGY SKILLS (BACES) ASSESSMENT
ANSWER SHEET

Directions: Please ***use the bubble sheet*** to mark your answers to the each of the items on your test form. You may also mark your answers directly on the test form if you would like to keep your test for later. Finally, do not forget to ***write which form*** you are using on the line below.

Form ID **A** ○ **B** ○

Question	Answer Choices			
	A	B	C	D
1	○	○	○	○
2	○	○	○	○
3	○	○	○	○
4	○	○	○	○
5	○	○	○	○
6	○	○	○	○
7	○	○	○	○
8	○	○	○	○
9	○	○	○	○
10	○	○	○	○
11	○	○	○	○
12	○	○	○	○
13	○	○	○	○
14	○	○	○	○
15	○	○	○	○
16	○	○	○	○
17	○	○	○	○
18	○	○	○	○
19	○	○	○	○
20	○	○	○	○
21	○	○	○	○
22	○	○	○	○
23	○	○	○	○
24	○	○	○	○
25	○	○	○	○
26	○	○	○	○
27	○	○	○	○
28	○	○	○	○
29	○	○	○	○
30	○	○	○	○



Please Turn Over to Complete the Background Questions

BACKGROUND QUESTIONS

Directions: Please use this page to answer a couple background questions by selecting the single answer that best describes you.

1. What residency year are you currently completing?

- ☐ First
- ☐ Second
- ☐ Third
- ☐ Fourth
- ☐ Fifth
- ☐ Sixth
- ☐ Seventh

2. What degree(s) have you attained? (check all that apply)

- ☐ MD
- ☐ DO
- ☐ PhD
- ☐ MS / MA
- ☐ MPH
- ☐ DrPH (DPH)
- ☐ Other (please specify) ____

3. In what country did you graduate medical school? _____

4. What is your sex?

- ☐ Male
- ☐ Female
- ☐ Prefer not to answer

Have you ever taken a course in:	**No**	**Yes**
5. Epidemiology	☐	☐
6. Biostatistics	☐	☐
7. Evidence-based medicine	☐	☐

Appendix C

Descriptive Answer Key for BACES Form "A"

BIOSTATISTICS AND CLINICAL EPIDEMIOLOGY SKILLS (BACES) ASSESSMENT

ANSWER DESCRIPTIONS (FORM A)

This document provides the set of answers for the BACES items as well as descriptions of *why* each answer was the most correct option for the question. In addition, you see a relative "Correctness" scale for each item, so that you can see where each response option was intended to be located in terms of correctness. My intention is that this document provides you, the examinee, with not only the correct answers to the test, but also with a useful insight into where and how you may have gone wrong in your thinking. Use this document in conjunction with the BACES test as a study tool for your future work. Also, please scan or type in the link at the end of this document for access to all of my online teaching materials.

1) Correct Answer: ***D. Retrospective Cohort Design***

Description: The most appropriate study design for this research scenario would be (d) a retrospective cohort design. The retrospective nature and non-randomized sample rule out (a) a randomized control design. Both (C) a nested case-control design and (B) a case-control design are non-randomized, retrospective designs; however, both designs require groups to be chosen based on their outcome rather than exposure. Since the resident has chosen his retrospective sample based on their exposure to HVC, and he is following them over a 5-year period for sustained virologic response, the strongest research design will be (D) a retrospective cohort.

2) Correct Answer: ***B. Case-Control Design***

Description: Similar to question one, the correct answer comes down to the selection of groups for the study, which in this example is (B), a case-control design. Both a case series design (C), and (D) a cross-sectional design could be used; however, they both lack a comparator group. Response (A) a double-cohort design would not be appropriate because the groups are selected based on the *outcome* (i.e., CRC) as opposed to some exposure of interest (i.e., colonoscopy). Since the researchers chose a group of healthy patients (no CRC) and a group of sick patients (CRC), the correct answer would be (B) a case-control design.

3) Correct Answer: ***D. Equivalency test for proportions***

Descriptions: A (D) equivalency test for proportions analyzes the similarity between a pair of proportions under the null hypothesis that the pair *does differ*. The researchers' hypothesis states, "That the two groups would not differ in the proportion of patients whose infection was eradicated by a clinically meaningful amount ***(+/- 15% eradication percentage)***." In other words, the researchers are testing to see if the regimens are no more than 15% better or worse than one another at eradicating infection. The closest other response is (C) non-inferiority test for proportions because it also tests for the similarity between two proportions; however, non-inferiority analyses are only concerned with the regimens being no more than 15% *worse* that one another rather than 15% better or worse. A (A) chi-square test of independence would be the appropriate analysis to use if the researchers were trying to find a significant *difference* between the regimens rather than a similarity. Finally, (B) Independent samples *t*-test, could not effectively test their hypothesis because the analysis tests for a *difference* between two *means* rather than the similarity of two proportions.

4) Correct Answer: ***A. The researchers controlled these variables as possible confounders in their analysis***

Description: The researcher (A) controlled for these variables as possible confounders in their analysis. In statistics, a covariate is a variable that is statistically controlled for or "held constant" across all patient groups during a particular analysis. Oftentimes, researchers will include a number of covariates such as age, height, BMI, and smoking status because these variables may distort (confound) the association they wish to assess. Responses (B) and (C) are both methods for reducing the impact of confounding variables either by excluding patients with those characteristics from the study (B), or by matching patients with similar characteristics across both the experimental and control group. Response (D) would occur as a possible result for not accounting for these confounding variables in that the association between psoriasis and diabetes may *depend* on the individuals smoking status or BMI, for example.

5) Correct Answer: ***A. Patients whose treatment was intensified were 74% less likely to be readmitted versus those whose treatment was maintained.***

Description: The vignette provided the results testing regimen intensification as a predictor for 90-day readmission, and an OR of 0.26; therefore, the correct way to interpret their results would be (A) *Patients whose treatment was intensified were 74%* (1.00 – 0.26) *less likely to be readmitted versus those whose treatment was maintained.* There are two primary ways to interpret and odds ratio (OR). First, if the odds ratio is *above 1.0*, then the exposure (intensified regimen in this case) *increases* the odds of the outcome by OR – 1.00 percent. Second, if the odds ratio is *below* 1.0, then the exposure *decreases* the odds of the outcome by 1.00 – OR percent. Response (B) describes the correct hypothesis being tested in the vignette, but the OR has been interpreted incorrectly as 1.26 (i.e. 26% increase in odds) rather than 0.26. Option (C) correctly interprets the decrease in odds, but interprets the incorrect hypothesis, and (D) fails to provide either the correct decrease in odds or hypothesis.

6) Correct Answer: ***D. Ratio***

Description: Of the four different measurement scales provided, the (D) ratio scale of measurement will provide the researcher with the most accurate estimate of a phenomenon because the scale consists of an infinitely divisible number of ordered values as well as a true zero point. The interval (C) scale of measurement is similar to the ratio scale in that it is ordered and numeric, but it lacks a true zero point (e.g. 0 degrees Fahrenheit is not an "absence of temperature"). The ordinal scale of measurement (B), as the name suggests, provides an order to a series of values; however, these values can be any distance apart from one another such as position in a marathon. Finally, (A) discrete is an umbrella term for both the ordinal scale *and* the nominal scale, and it refers to a type of measurement where individuals are classified into discrete categories (e.g. male or female, first or second place).

7) Correct Answer: ***C. Increase the sample size from 100 patients in each group to 150 patients in each group***

Description: The most effective way to increase statistical power in this example would be to increase the sample size (C). Increasing β (beta) from 0.20 to 0.40 would actually *decrease* the statistical power of the study because power is equal to 1.00 – β (beta). Similarly, decreasing α (alpha) from 0.05 to 0.01 would also decrease the statistical power because the threshold for determining statistical significance has been increased from 95% (0.05) to 99% (0.01). Finally,

using an ordinal scale of measurement would also likely decrease the statistical power of the study because power is reduced whenever a less-precise scale of measurement is used (D).

8) Correct Answer: ***A. Your team concludes there is a statistically significant difference between the statin regimens when in fact there is not a difference.***

The journal reviewer is suggesting the team has made a Type I error in determining the results of their study, which means that the team claimed that there was a statistically significant difference between the statin regimens when in fact there was not a difference (A). Response (C) defines a Type II error, which occurs when a researcher fails to find a statistically significant difference when one truly exists. The low sample size described in option (B) is more likely to result in a Type II error rather than a Type I error because small sample size usually equates to low statistical power, and therefore a high chance of a Type II error. Finally, (D) defines an issue with confounding rather than Type I error although if the significant association was due to some other variable (confounder) the researchers failed to address, then it could be a plausible *cause* of the Type I error in question.

9) Correct Answer: ***D. Cross-sectional design***

The best approach for the research question would be to use a (D) cross-sectional design to assess residents' attitudes of research design and statistics. A cross-sectional design assesses the phenomenon of interest at in an entire population at a single point in time, and does not have a specific comparison group. Option (B) would not be appropriate because the faculty member is prospectively gathering information from participants at a single point in time, and is surveying an entire population rather than selecting a group of cases and controls to compare to one another. A (A) longitudinal design would be appropriate if the researcher planned to survey the same group of residents at various time points of a follow-up period rather than at a single point. Similarly, (C) includes both a follow-up duration *and* a comparison group to follow over time.

10) Correct Answer: ***B. Nominal variable***

Description: NSAID use in this descriptive table is presented as a (B) nominal variable. A nominal variable is one that is measured as discrete categories such as Male/Female, White/Black, and Yes/No. In tables, nominal variables are presented by frequencies and percentages rather than means and standard deviations. Since the table provides the number and percent of patients taking NSAIDs (Yes taking it/No not taking it), it is considered a nominal variable. A full description of scales of measurement can be found in the description to answer 6 (form A) or 22 (form B).

11) Correct Answer: ***B. 55%***

Description: Relative risk ratios (RR) are interpreted as those over 1.0 being an *increase* of risk by RR – 1.00 percent, and those below 1.00 are a *reduction* in risk by 1.00 – RR percent. The effect size, or magnitude of the difference, is the percent difference in risk between the exposed and unexposed groups. In this example, the RR for mice exposed to the lidocaine/Marcaine mixture compared to those injected with saline was RR = 0.45, and since it is below 1.00 the effect size is 1.00 – 0.45, or (B) 55%. Option (A) would be correct if the RR was 1.45 rather than 0.45. Option (C) equates to 0.89 – 0.25, or, the width of the 95% confidence interval for the RR. While the width of the confidence interval can be an estimate for the *accuracy* of the effect size, it is not a measure of effect itself. Finally, 95% (D) is the parameter for the type of confidence interval, and not a measure of effect.

12) Correct Answer: ***A. Median***

Description: In situations where the variable of interest is not normally distributed, (A) median will usually be the preferable measure of central tendency as opposed to the other options given. In a normal distribution, both the arithmetic average (mean) and the middle of the distribution (median) are located very close to one another; however, the mean can be skewed by extreme values on either the low or the high end of the distribution. The median will not be affected by these extreme values, and therefore it will be a more accurate measure. The mode (C) only provides the most frequent response, which may or may not be near the middle of the distribution, and the range (D) is the width of the distribution from the lowest value to the highest value; it measures the *spread* of the data as opposed to the center of it.

13) Correct Answer: ***C. The new test is 88% effective at detecting positive disease states, and 71% effective at detecting negative disease states.***

a. **Description:** In diagnostic testing, Sensitivity (SE) refers to the ability for the test to identify *positive* disease states, and Specificity (SP) refers to the detection of *negative* disease states, regardless of accuracy. For example, a highly sensitive test will be excellent at detecting individuals that have the disease of interest; however, there will also be many healthy individuals who are false positives. Response option (C) is the best way to interpret the values in the clinical trial because it correctly defines SE and SP in the context of the example. Option (A) incorrectly interprets the values of SE and SP in the example by reversing their definitions. Both options (B) and (D) incorrectly assume SE and SP detect *true positive* and *true negative* disease states when it is the predictive value of a test that provides these measures of accuracy.

14) Correct Answer: ***B. Point "B"***

a. **Description:** A ROC curve displays the values of sensitivity and 1-specificity along various possible cut-points for a diagnostic test. The objective of the question is to maximize both SE and SP, which is always the value closest to the upper left-hand corner of the curved line (Option B: Point "B"). Option (A) would provide a highly specific test because 1 – SP would be very low, but the test would have very low SE. Point (C) provides a reasonable balance of SE and SP although the SP would be smaller than point "B." Option (D) is located along the straight line of the figure, which represents a test with 1:1 odds of identifying disease states, and therefore not part of the test in question.

15) Correct Answer: ***C. Inter-rater reliability***

Description: Inter-rate reliability (C) is the extent to which multiple observers of the same phenomenon are similar in their ratings. When conducting a behavioral observation or other similarly subjective data collection, it is important to use multiple observers (i.e. raters) to be sure individual biases do not affect the results. Responses (A), (B), and (D) are additional forms of reliability that are usually used in self-report or survey instruments. Test-retest reliability (A) is the extent to which students perform similarly when taking the same instrument multiple times while internal consistency reliability (B) describes the similarity among answers to similar

questions on a single instrument. Finally, split-half reliability looks at the consistency in participants' responses on the first half of an instrument compared to the second half.

16) Correct Answer: ***D. Independent t-test***

Description: Option (D), independent *t*-test, is the most appropriate method for comparing the average heart to lung uptake between two separate patient groups. Option (A) would not be appropriate because the researcher is only concerned with a single variable (tracer) rather than *multiple* variables as the name implies. A paired *t*-test (B) is not correct because there is no evidence that the groups have been individually matched, and there are two separate treatment groups as opposed to one. Finally, a (C) chi-square test of independence is not appropriate because it would test the difference in proportions between two groups rather than a difference in means.

17) Correct Answer: ***B. Relative Risk***

This case involves a prospective cohort study in which the true population is known over the 16-year follow-up period; therefore, the most accurate measure of association to use in this situation would be (B) relative risk. An odds ratio (OR) is only an estimate of RR, and it is used in retrospective or descriptive studies when the researcher can only assess the prevalence of the outcome rather than its incidence. An incidence rate ratio (C) could be used in this study if the research question was interested in the rate at which CRC developed between the five groups rather than simply the "likelihood" of CRC developing. Finally, (D), absolute risk, is simply the proportion of patients in each quintile who developed CRC; therefore, it is descriptive rather than a measure of association.

18) Correct Answer: ***A. no more than 10% less effective than the standard of care***
Description: The authors are concluding that their experimental treatment was (A), no more than 10% less effective than the standard of care. When investigating a non-inferiority study, the researchers aim to prove that the experimental treatment is no worse than the standard of care by a predetermined margin (10% in this case). In non-inferiority studies, a statistically significant finding indicates that the experimental treatment *did not* perform any worse than the margin (10%) compared to the standard of car. Option (C) would be correct if the researchers were looking for the two treatments to be statistically *equivalent* rather than *non-inferior* because equivalence studies are concerned with the experimental treatment performing no better *or* worse than the standard of care as opposed to simply no worse. Finally, options (B) and (D) would both correctly test the difference between the proportions of pulmonary embolisms in each treatment group, but they would not be able to prove the experimental treatment is "No worse than the standard of care."

19) Correct Answer: ***B. Cut-off "B"***
Description: In diagnostic testing, balancing the values of SE and SP provide the researcher with the fewest false positive and false negative results. Cut-off "A" would result in a test that is very sensitive, excellent at detecting *positive* disease states; however, the false-positive rate would be very high. Conversely, Cut-off "D" would result in a highly specific test that was effective at detecting negative disease states, but would also identify a large number of false negatives. Both cut-off "B" and "C" provide some balance of SE and SP; however, "B" would be the ideal balance to minimize false negative results.

20) Correct Answer: ***A. 2x2 factorial design***
Description: The best research design to describe this situation is option (A), 2x2 factorial trial. There are exactly two variables (physician type and educational program) that each has two different possibilities (family or industrial physician, and received or did not receive the educational intervention). Each participant received only one combination of physician type and educational program, so there was no crossover involved (C). The use of randomization makes (D) and (B) both less accurate descriptions of this scenario than option (A).

21) Correct Answer: ***A. Multiple Linear Regression***

Description: The most appropriate statistical approach in this study would be to use (A), multiple linear regression. All of the responses would allow the researchers to include Apnea-hypopnea index, age (in years), sex, and neck circumference (in centimeters) as possible predictors; however, multiple linear regression is the only approach that could evaluate the increase in a continuous outcome such as diastolic blood pressure. Multiple ordinal regression (D) would be appropriate if the outcome was measured on an *ordinal* scale such as 1 = "No change in BP" to 5 = "Substantial change in BP" while multiple logistic regression (B) would be useful if the outcome was a dichotomy (e.g. increased BP or not). Finally, (C), multiple Cox regression would be the least appropriate option as it is designed not only for dichotomous outcomes but also for time-to-event analysis.

22) Correct Answer: ***D. The estimate for women in the third quartile of CRP is more accurate than the estimate for the fourth quartile.***

Description: Given these results, (D) The estimate for women in the third quartile of CRP is more accurate than the estimate for the fourth quartile. This question asks the reader to evaluate the results, and make a judgment based on the 95% confidence intervals presented.

The 95% CI indicates the range in which the researcher is 95% certain that the true association (RR in this study) exists. The width of the 95% CI is therefore directly related to the accuracy of the estimated association where a wider interval indicates a less-accurate estimate, and a smaller interval indicates a stronger estimate. In the study, the 95% CI for the third quartile of CRP is 3.6 – 21.0 while the fourth quartile is 6.5 – 37.9. Since the latter interval is wider than the former, it is the less-accurate estimate. Both options (B) and (C) provide an inaccurate interpretation of the 95% CI, so they are not an appropriate way to describe the researchers' estimates.

23) Correct Answer: **C. We expect 45.5 falls for every 1,000 months of follow-up**

Description: The correct way to interpret these results is as an *incidence rate*, which would be that (C) we expect 45.5 falls for every 1,000 months of follow-up. A *rate* is the incidence of an event of interest (falls) *per a unit of person-time* (1,000 person-months) while a *proportion* is simply the number of times an event occurred out of the total number of trials. A rate may be anywhere between 0 and infinity, and is set to whatever unit makes sense to the researcher. For example, the same rate could be 45.5 falls per 1,000 person-months, 0.0455 falls per 1 person-month, or 455 falls per 10,000 person-months. On the other hand, a proportion will *always* fall between 0 and 1.0 (i.e. 0 to 100%). Options (A) and (D) are not correct because they interpret the

incidence of falls as proportions rather than rates. Option (B) is incorrect because it uses the arbitrary unit (1,000 person-months) as the total number of months the 362 patients were followed rather than the actual total follow-up the researchers observed in these patients over 50 weeks.

24) Correct Answer: ***C. Multiple Cox Regression***
Description: The best choice in this study would be to use multiple Cox regression (C) to look at patient survival between two surgical procedures while accounting for age, sex, and cardiovascular history. Kaplan Meier analysis (A) also allows for the analysis of survival data; however, it can only look at a single predictor (i.e., surgical procedure in this case) at a time rather than the multiple predicts the researcher wants to test. Multiple logistic regression (D), on the other hand, could assess the multiple predictor variables at once time, but would not be able to address the researcher's survival analysis question. Finally, chi-square tests (B) are only used to assess a single predictor variable and a single outcome variable, and cannot address time-to-event analysis, which makes it the least desirable option.

25) Correct Answer: ***B. Recall Bias***
Description: Case-control studies are susceptible to each of these biases; however, this particular case-control study design is most susceptible to (B) recall bias. The researchers have identified a group of athletes with CECS and a group *without* CECS and then interviewed them about their previous exposures to possible risk factors. Recall bias stems from the inaccurate recollection of a patient's experiences the exposure of interest. In general, those individuals who have the disease (the "case" group) are more likely to recall their exposures to possible risk factors (whether truthfully or not) compared to their healthy controls. If recall bias occurs, then case-control designs are also susceptible to misclassification bias (A) because poor or inaccurate recall of exposures (i.e. recall bias) may lead to one group being classified as "exposed" more or less than the other group. Experimenter bias (C) is a bias associated with the way the researcher treats each study group. For example, an interviewer may ask patients with CECS different questions compared to those without CECS. Standardizing interview protocols/procedures, and blinding the interviewer are two ways to minimize this bias. (D) Medical surveillance bias can occur if the case-control study uses hospital *cases* and population *controls*, and when the exposure of interest is associated with visiting the hospital. This study is least susceptible to this because *both* the cases and controls came from a group of patients at a medical clinic with the same chief complaint (lower leg pain).

26) Correct Answer ***D. Within-subjects analysis of variance (ANOVA)***
Description: A within-subjects analysis of variance (ANOVA) is the most appropriate statistical test for this situation because it effectively measures the change in LDL cholesterol in the *same* group of patients over *two or more* difference measurements (baseline, six weeks, and twelve weeks). A dependent or "paired" *t*-test also tests repeated measures of a single group, but can only be used for *exactly two* measurements (e.g. baseline and twelve weeks). Between-subjects ANOVA (C) can address more than two measurements; however, it is a test reserved for comparing two or more *different* patient groups rather than a single group. Finally, an independent *t*-test (B) neither tests within-subjects nor allows for more than two measurements, which make it the least desirable option.

27) Correct Answer ***B. Number Needed to Treat***
In an interventional trial where the exposure decreases risk of a negative outcome, (B) number needed to treat (NNT) addresses the researchers' question about the number of antibiotic prophylaxis treatments would need to be administered to prevent a single death from sepsis. Had the study investigated a possible risk factor for an *increased* likelihood of sepsis, than (A) number needed to harm (NNH) would be the correct method for finding how many individuals would need to be exposed to the risk factor before one person became septic. Likewise, attributable risk (AR) (D) would be an appropriate measure of how much excess risk for sepsis was attributed to being exposed to the risk factor. In an interventional trial both NNH and AR would be negative numbers, which would not make sense to interpret.

28) Correct Answer ***D. Diffusion***
Description: The threat to internal validity that John is most likely increasing through his actions is (D), diffusion, or the spread of treatment effects across multiple treatment groups. By offering Susan his medication to help her symptoms, John is modifying Susan's treatment and her results as a member of the placebo group will no longer be valid. Both compensatory rivalry (B) and demoralization (C) could be playing a role in John and Susan's situation; however, they are not being directly influenced by the sharing of medications in the same way diffusion has been

affected. Finally, there is no evidence that (A), attrition has taken place in this situation, as both individuals are staying enrolled in the study.

29) Correct Answer ***A. External***

Description: When the critic claims the results would not generalize to human subjects, he is referring to the study's (A) external validity. External validity is most basically the ability for the results of one study to *generalize* to the larger population. Since this study was conducted in a pig model, the critic is claiming that it may not generalize to the human population. While the critic is not describing it in his letter, this study also has questionable (C) ecological validity, which is the extent to which a variable is measured in the way it would naturally exist. By manipulating the pig model to mimic a human appendix, the researchers were measuring their skill in the most natural setting, that is, an actual human appendix. Internal validity (B) is the strength of causal inferences that can be made *within a single study*. In this example, it would be the extent to which the researchers can claim that their training *caused* an increase in resident skill level. Finally, (D) construct validity broadly refers to the extent to which the researcher is measuring what they claim to be measuring (resident skill level), which is not the critic's concern.

30) Correct Answer ***A. H_0: Rate prior to intervention = Rate after the intervention;***
H_1: Rate prior to intervention > Rate after the intervention

Description: The most accurate way to articulate the researchers' hypotheses would be using response option (A) because it specifically states a *directional* alternative hypothesis (i.e., "There will be a significant decrease in the rate of hospitalizations"). Response (B) would be correct had the researchers simply wanted to investigate *any* difference in hospitalization rate, regardless of increase or decrease. Response (D) incorrectly specifies the direction of the alternative hypothesis, and response (C) incorrectly specifies both the null and alternative hypotheses.

WANT TO KNOW MORE?

Use your smartphone to scan the QR code below (or type in the web address) for full access to my online lectures and course materials.

Address: www.slideshare.net/pbbarlow1

Scan me!

Appendix D

Brief Project Description and Memorandum of Understanding for Participating Institutions

Brief Study Proposal Outline for Patrick Barlow's Dissertation

Title: *Development of the Biostatistics and Clinical Epidemiology Skills Assessment for Medical Residents*

Purpose: The purpose of the proposed study is to establish preliminary item characteristics and validity evidence for the Biostatics and Clinical Epidemiology Skills (BACES) assessment.

Background & Rationale:

- This is not a new problem in GME:
    - Studies back to the 1980's (newest study *just* published in JGME two months ago) show many physicians lack the fundamental understanding necessary to adequately read statistics they encounter in the medical literature.
    - While physician knowledge has remained steadily low yet variable over the past three decades, the frequency and complexity of statistics in the literature has risen dramatically.
    - Although the Accreditation Council for Graduate Medical Education (ACGME) includes these topics within their core program standards (medical knowledge and practice-based learning and improvement), assessment of these topics is sparse and generally done on a per-campus basis (i.e. no validated instrument).

- Need for an instrument that *addresses* the problem rather than diagnoses it.
    - Programs will need to develop new, better methods for assessing both clinical and non-clinical skills as the Next Accreditation System continues to be implemented.
    - A 2011 review of existing instruments (Enders, 2011) explicitly called for new, better assessment instruments in this area. This author has also offered to review the BACES items as an expert reviewer.
    - While study-after-study has confirmed how little physicians know about these areas, few have attempted to *do* anything about it.

Plan for Data Collection:

- I intend to give the BACES assessment to residency groups during either their regularly scheduled journal club / didactic time, *or* a separate time at the department chair's convenience.
- The residents will first complete the BACES assessment (approximately 30 minutes), and then my colleague(s) and I will spend the remaining 30 minutes going over the answers in a large group.
- Anyone who participates in the study will receive a copy of the answer key that describes the rationale for each answer in detail as well as several additional online educational resources from the Office of Medical Education, Research, and Development (OMERAD) at the UT Graduate School of Medicine.

- The total expected time investment is approximately *one hour*, and all data will be collected confidentially.

EXAMPLE ITEMS

What they read…

1. An internal medicine resident is interested in looking at the effect of adherence to ribavirin and interferon therapy on virologic response in patients with hepatitis C (HVC). His team uses a national case registry to identify and follow all HVC patients with either 50% or 100% medication adherence between January 2003 and June 2008. He then analyzes both early and sustained virologic response across two different adherence groups.

   Which research design is most appropriate for this scenario?
   A. A randomized control trial
   B. A case-control design
   C. A nested case-control design
   D. A retrospective cohort design

2. A research team conducted a prospective study of 88,757 women to investigate the association between high dietary fiber intake and colorectal cancer (CRC). They compared the likelihood of developing CRC over the 16-year follow-up period among five quintiles of dietary fiber intake.

   What would be the most *accurate* measure of association to use in this situation?
   a. Odds ratio
   b. Relative risk
   c. Incidence ratio
   d. Absolute risk

After the self-assessment, participants are given a description of the question and rationale for each response.

1. The most appropriate study design for this research scenario would be (d) a retrospective cohort design. The retrospective nature and non-randomized sample rule out (a) a randomized control design. Both (C) a nested case-control design and (B) a case-control design are non-randomized, retrospective designs; however, both designs require groups to be chosen based on their outcome rather than exposure. Since the resident has chosen his retrospective sample based on their exposure to HVC, and he is following them over a 5-year period for sustained virologic response, the strongest research design will be (D) a retrospective cohort.
2. The most accurate measure of association of those listed would be (b) relative risk. (a) Odds ratios are accurate *estimations* of relative risk; however, they are less appropriate in a prospective study as they tend to over-estimate the association compared to relative risk. (c) Incident ratios could be an accurate approach to measuring the rate at which new cases of CRC develop within the cohort, but are not used to look at the *likelihood*. Finally,

(d) absolute risk is a descriptive measure of risk within a single group but would not be used to statistically compare among the five quintiles.

Memorandum of Understanding for Sample Institutions

To the University of Tennessee Institutional Review Board,

I, (Name of DIO), give my permission for Patrick Barlow, a PhD candidate at the University of Tennessee, Knoxville, to conduct data collection for his dissertation research at (Name of Institution). As the Designated Institutional Official for graduate medical education, I understand that Mr. Barlow will be gathering data from the medical residents at my institution, and that this data collection process will be comprised of (1) a biostatistics and clinical epidemiology knowledge self-assessment, and (2) an educational follow-up discussion of the assessment answers.

I also give my permission for Mr. Barlow to have ownership of the data collected from my institution, and that he may use it for future academic work such as professional conference presentations and academic journal publications. I understand that no identifying information will be collected from my residents, and that no reference will ever be made that could personally identify any of the participants from my institution.

Sincerely,

Name (Printed)

Signature

Appendix E
Participant Informed Consent Document
INFORMED CONSENT STATEMENT
BACES Self-Assessment Study

INTRODUCTION

Today you are being asked to participate in a dissertation research study looking at biostatistics and research methods knowledge in medical residents. The purpose of the study is to develop a useful and flexible self-assessment tool that medical educators and residents will be able to use in their own work. Your participation today is completely ***voluntary***, and you may decline to participate at any time.

YOUR INVOLVEMENT IN THE STUDY

Should you choose to participate the study will take place in two parts. First, you will be asked to complete a multiple-choice self-assessment of your biostatistics and research methods knowledge as well as several background demographic questions. Second, we will go over the answers to each item as a large group to review the concepts that were covered. Each part of the study should take between 20 and 30-minutes to complete.

You may agree to participate in one, both, or neither pieces of today's study, and your refusal to participate in any portion of this study ***will not*** affect your standing with your organization in ***any way***.

RISKS

There are no foreseeable risks to your participation in today's study. Some individuals may feel uncomfortable answering questions about their biostatistics and/or research methods knowledge; however, be assured that all answers will be kept ***confidential***, and no identifying information will be collected that could link your responses to your assessment. Finally, your participation is completely ***voluntary***, and you may choose to withdraw or refuse to participate at any time.

BENEFITS/COMPENSATION

There is no incentive for participating in today's study; however, there is an educational benefit for your participation. Specifically, the answers to each self-assessment item will be discussed after everyone has completed their assessment, and everyone will receive a copy of the answer descriptions for their future reference. Finally, each participant will receive a number of online resources to use in their future work.

CONFIDENTIALITY

As we have said, all data collected today will remain confidential. Data will be stored securely and will be made available only to the researcher. No reference will be made in oral or written reports which could link you as a participant to the information you provide here today.

CONTACT INFORMATION

If you have questions at any time about the study or the procedures, (or you experience adverse effects as a result of participating in this study,) you may contact the principal investigator, Patrick Barlow, at:

A 503 Bailey Education Complex
The University of Tennessee
Knoxville, TN 37996.

If you have questions about your rights as a participant, contact the Office of Research Compliance Officer at (865) 974-3466.

PARTICIPATION

Your participation in this study is voluntary; you may decline to participate without penalty. If you decide to participate, you may withdraw from the study at anytime without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

CONSENT

I have read the above information. I have received a copy of this form. I agree to participate in this study.

Participant's signature _____ Date _____

Investigator's signature _____ Date _____

Appendix F

Item Characteristic Curves (ICCs) for Final 2PL Model

The figures within this appendix display the ICCs for each of the 26 items in the final 2PL model. The appendix is divided into two separate sections corresponding to the clinical epidemiology and statistics dimensions. The items within each dimension are displayed in individual figures along with their "a" and "b" parameter values. Figures F1a-F1b contain the ICCs for the clinical epidemiology dimension while Figures F2a-F2b display the ICCs from the statistics dimension.

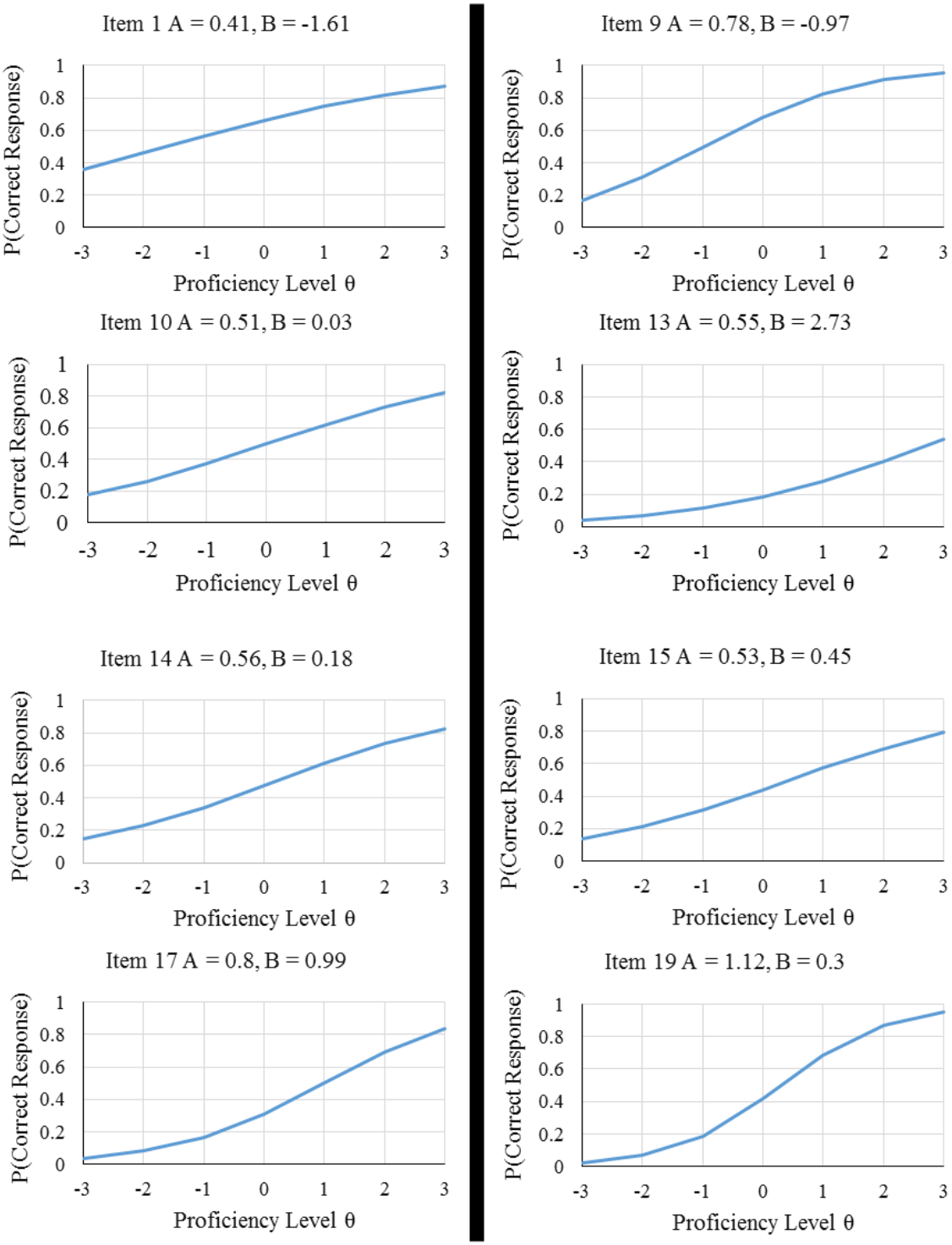


Figure F1a.. ICCs for Clinical Epidemiology Dimension Items 1, 9, 10, 13, 14, 15, 17, and 19

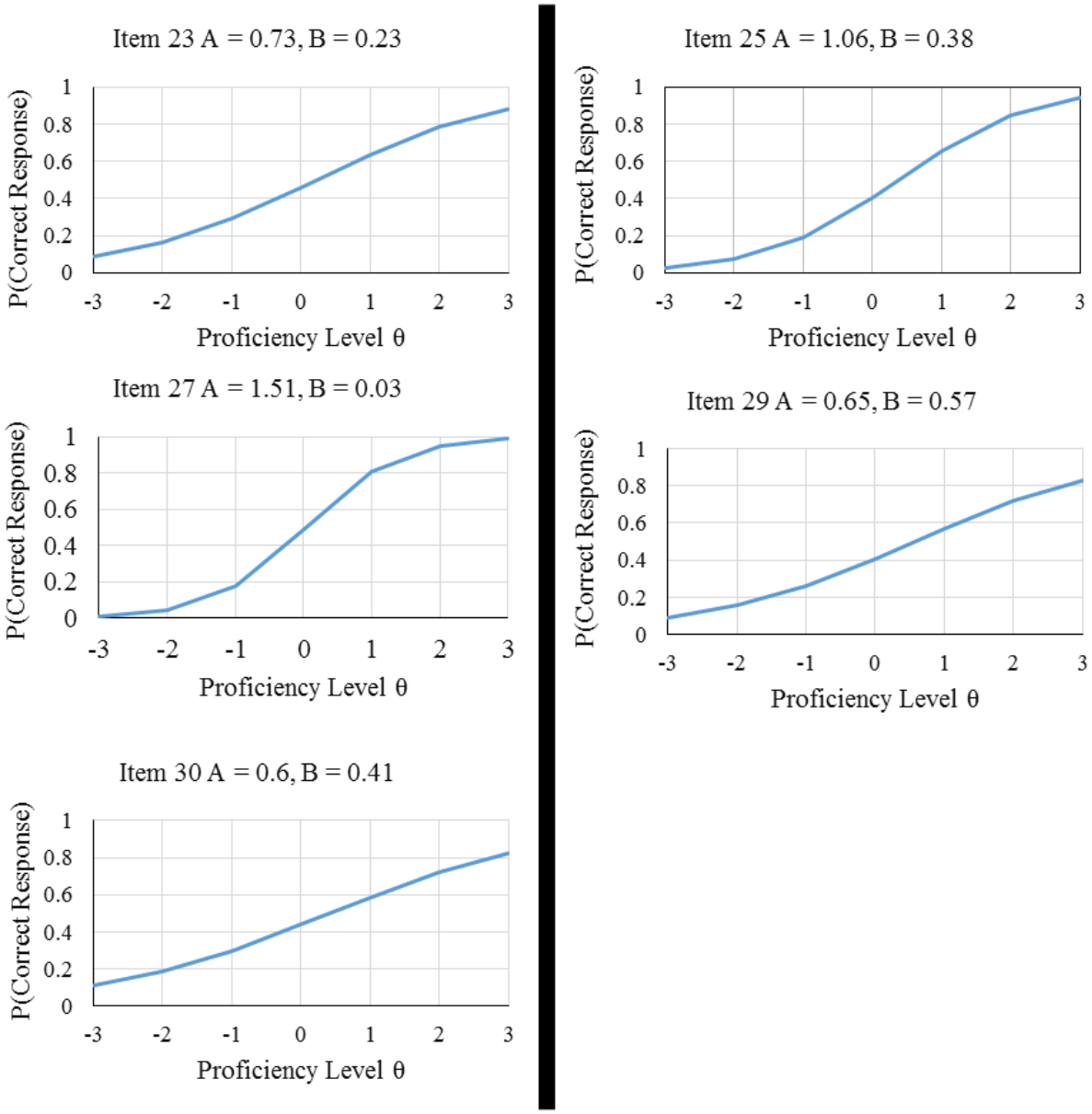


Figure F1b. ICCs for Clinical Epidemiology Dimension Items 23, 25, 27, 29, and 30

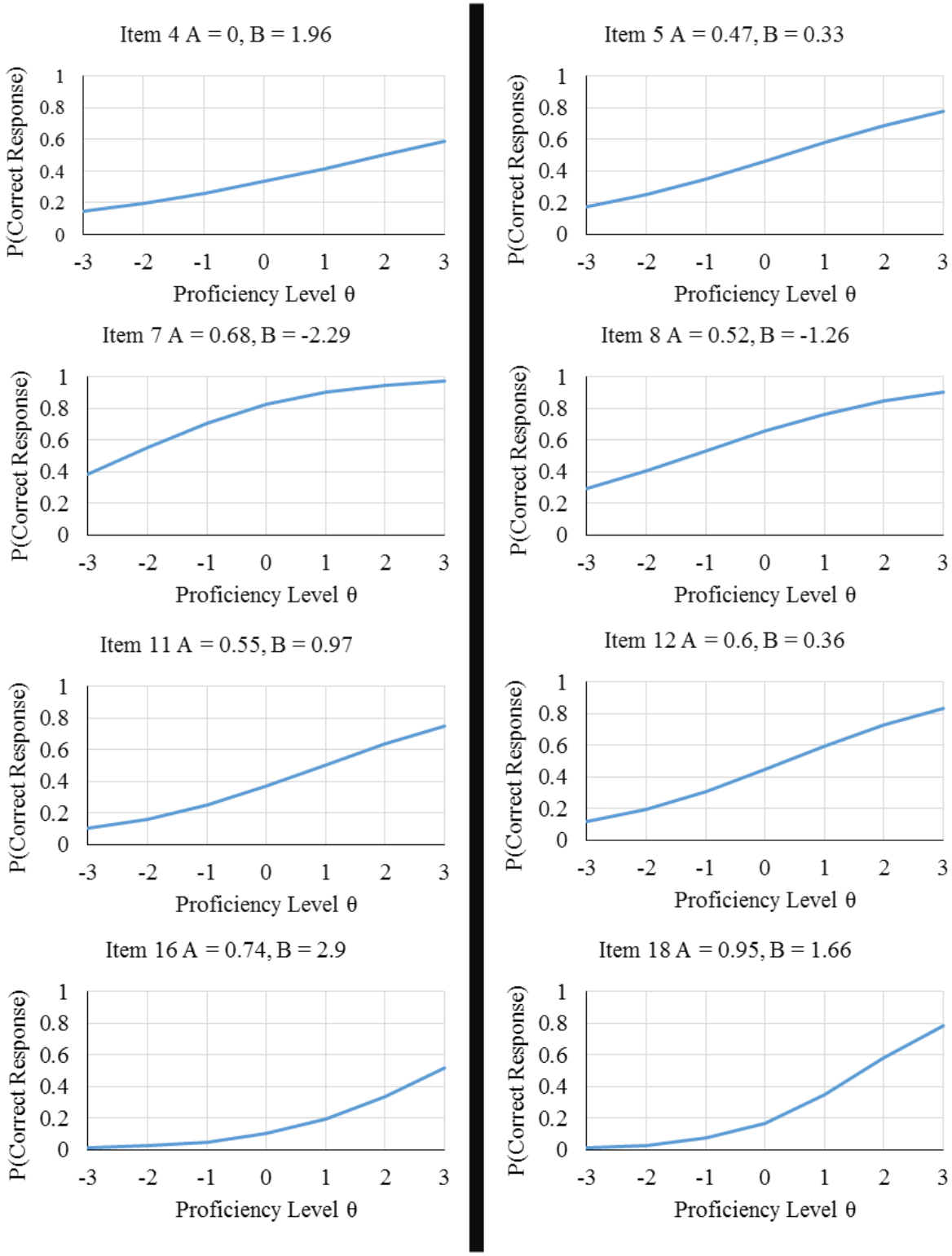


Figure F2a. ICCs for Statistics Dimension Items 4, 5, 7, 8, 11, 12, 16, and 18

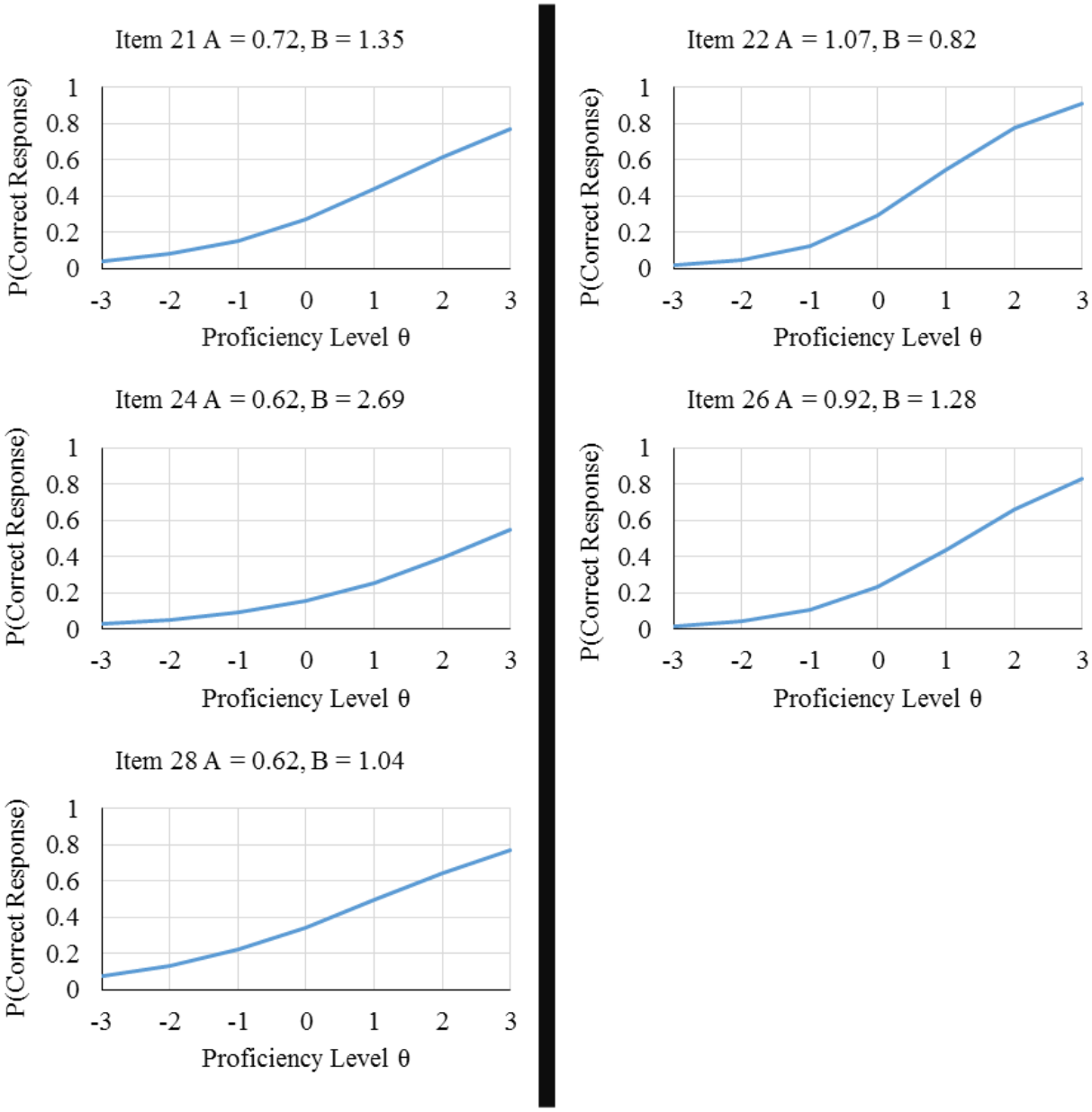


Figure F2b. ICCs for Statistics Dimension Items 21, 22, 24, 26, and 28

Appendix G

Item Information Functions (IIFs) for Final 2PL Model

Similar to Appendix F, this appendix contains the individual IIFs for each of the 26 items in the final 2PL model. These figures are also grouped together according to their dimension with Figures G1a-G1b representing the IIFs for the clinical epidemiology items and G2a-G2b representing the statistics items.

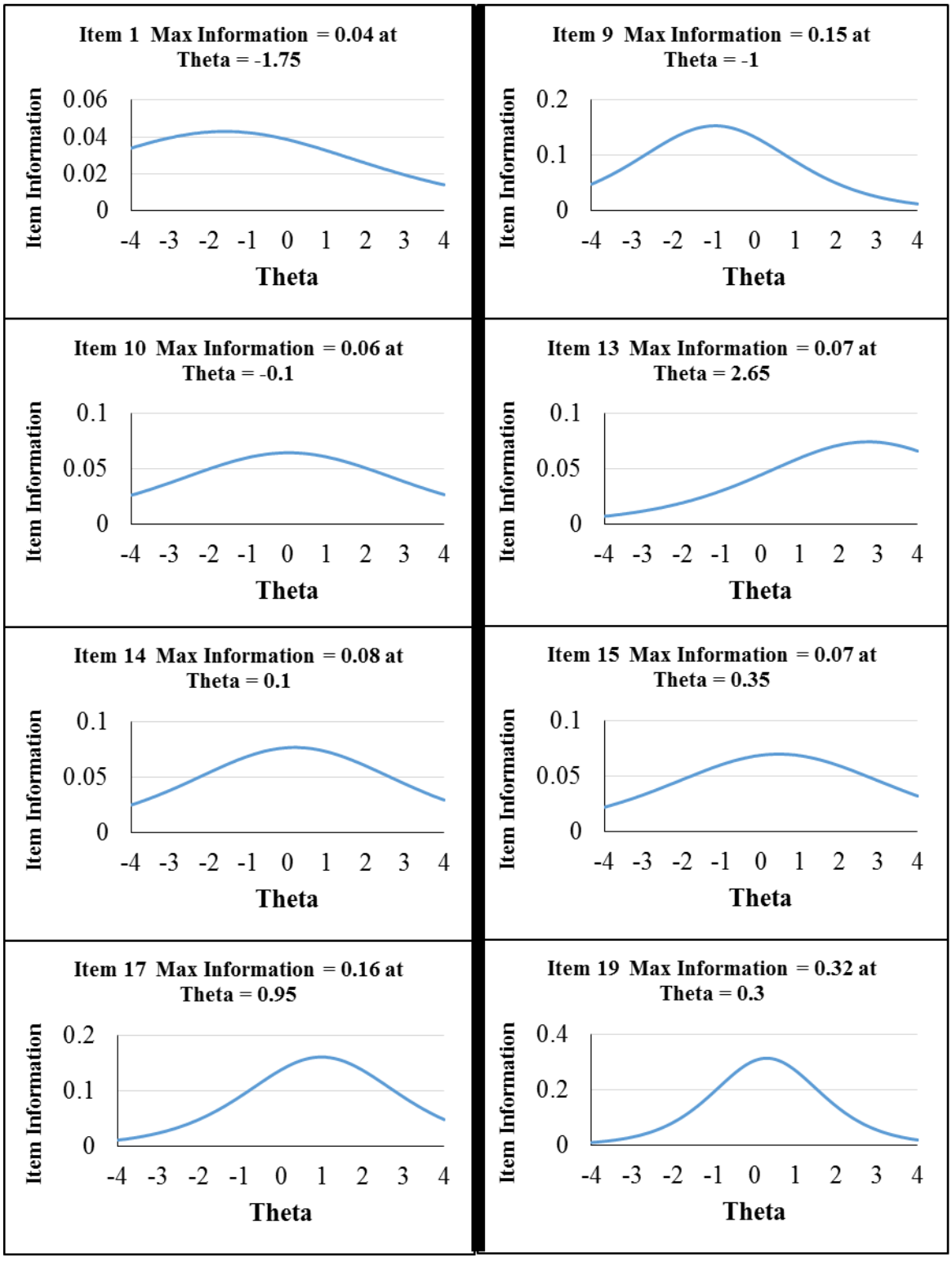


Figure G1a.. IIFs for Clinical Epidemiology Dimension Items 1, 9, 10, 13, 14, 15, 17, and 19

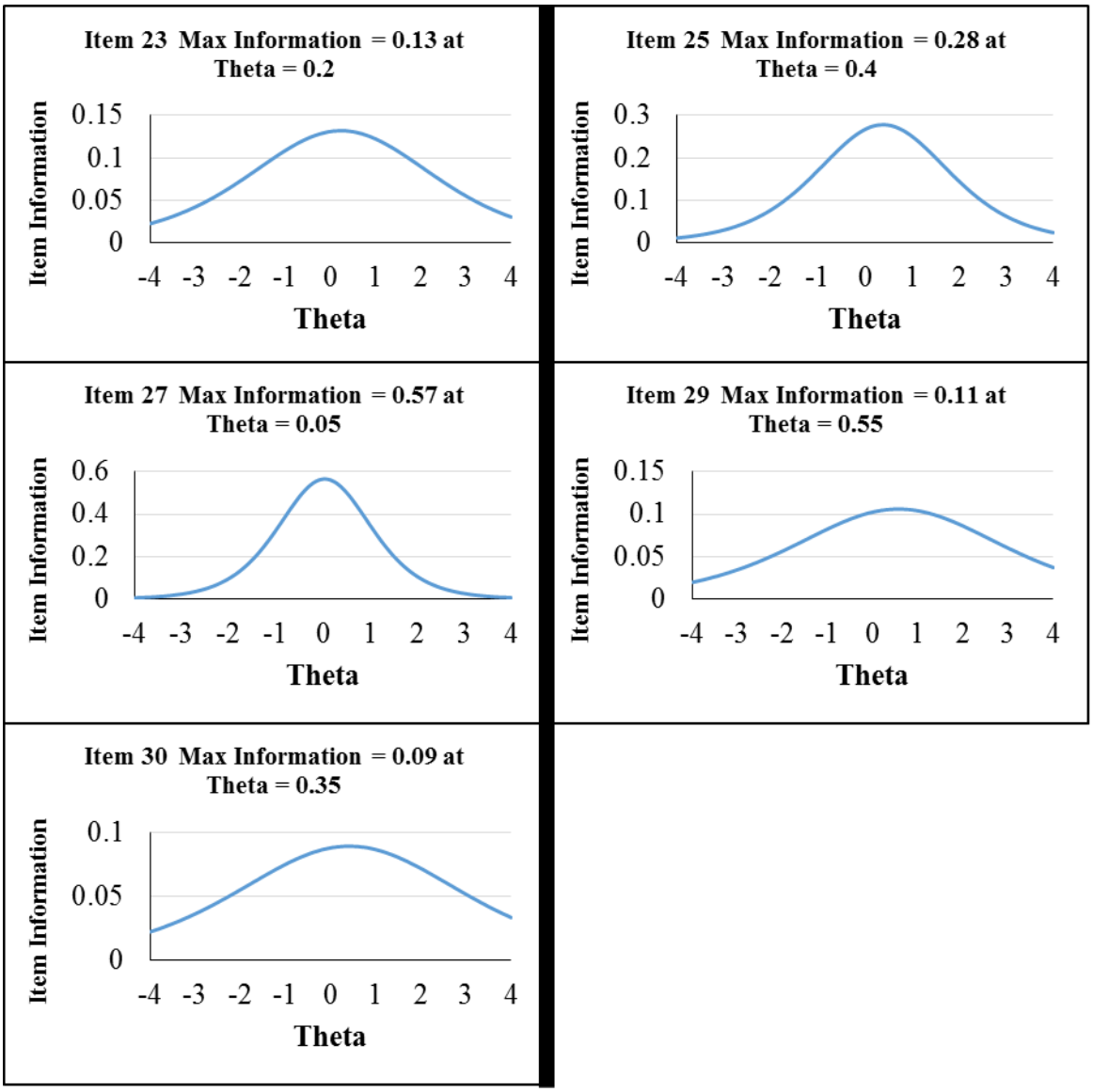


Figure G1b. IIFs for Clinical Epidemiology Dimension Items 23, 25, 27, 29, and 30

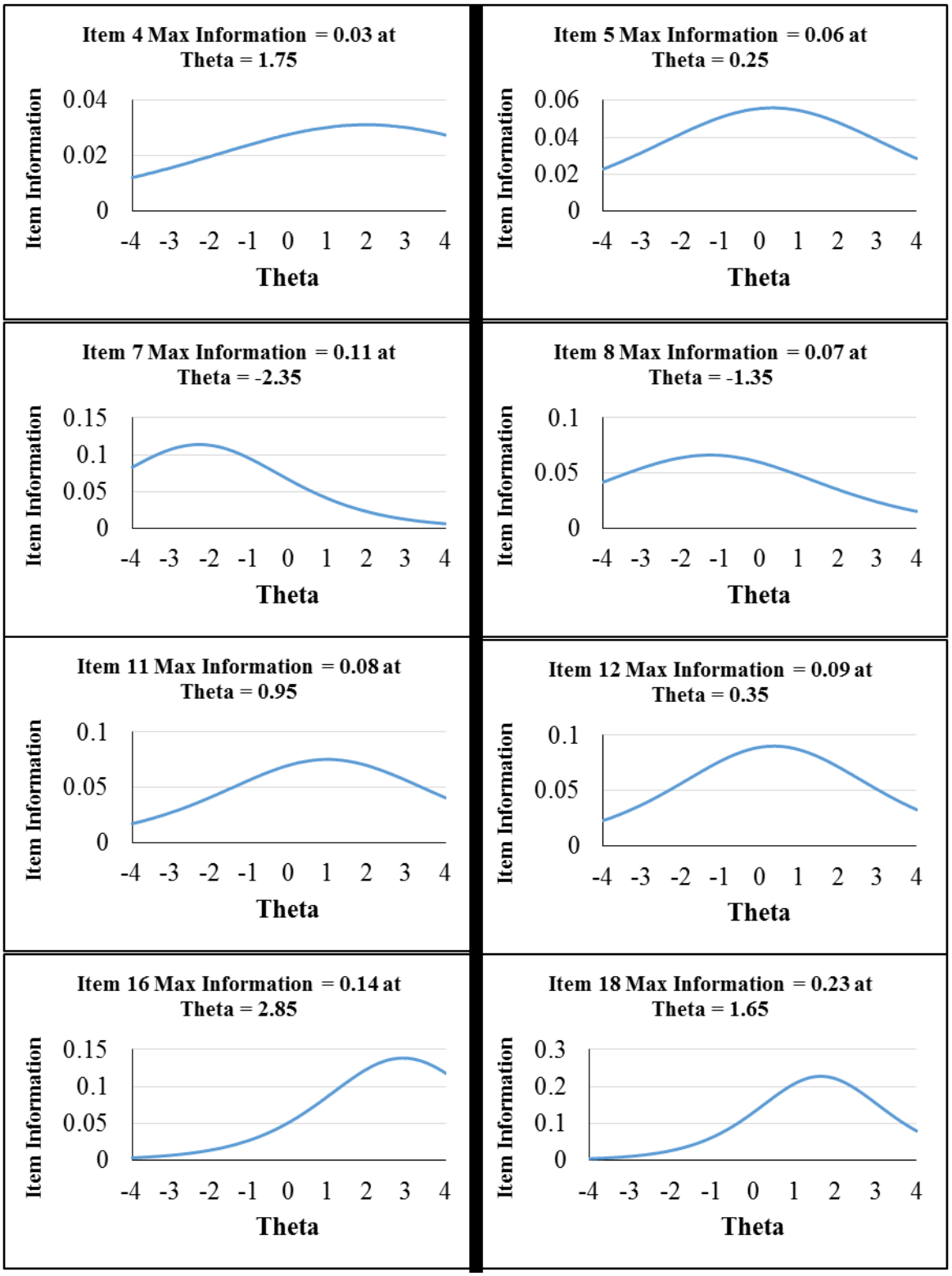


Figure G2a. IIFs for Statistics Dimension Items 4, 5, 7, 8, 11, 12, 16, and 18

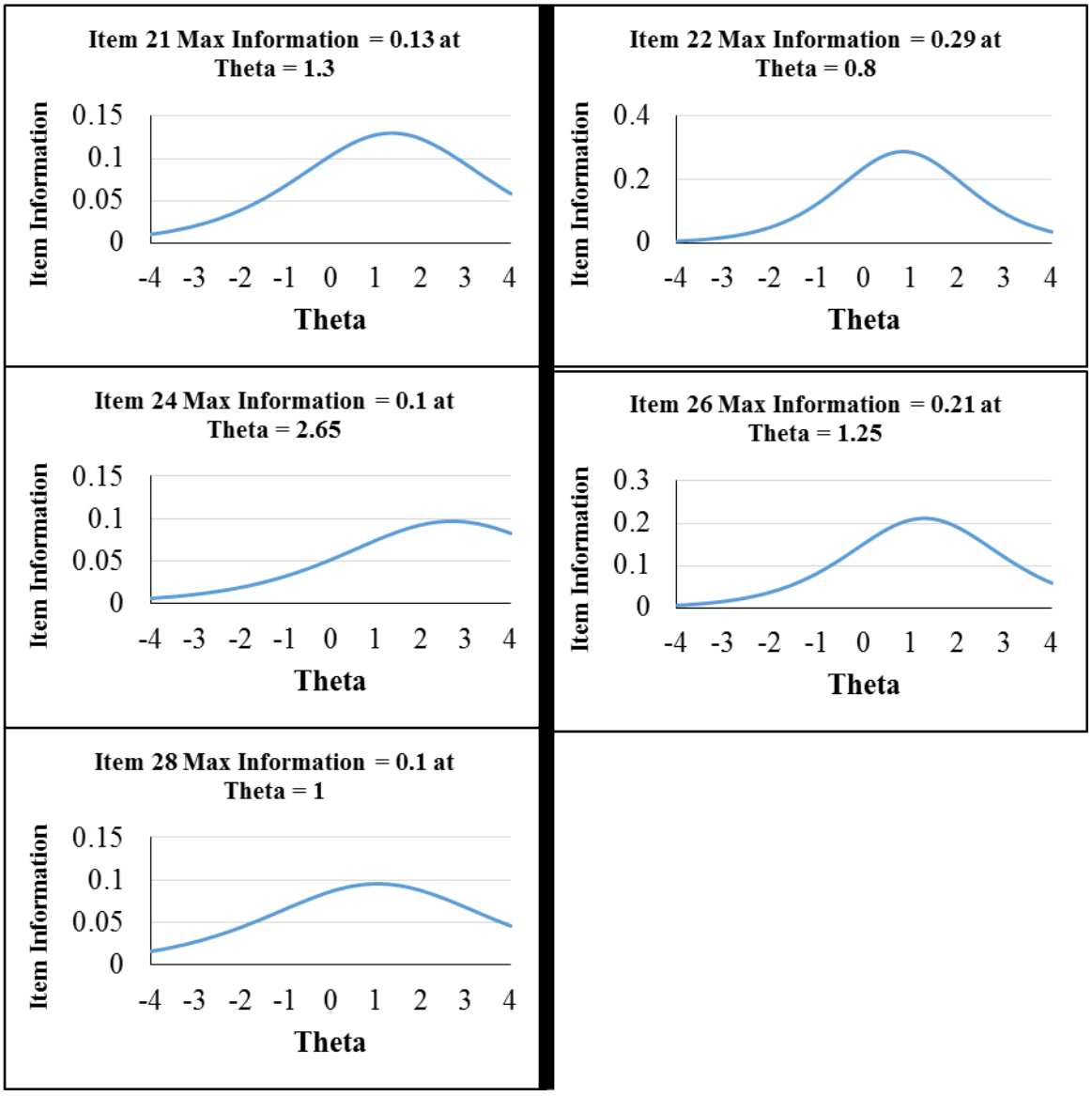


Figure G2b. IIFs for Statistics Dimension Items 21, 22, 24, 26, and 28

Vita

Patrick Brian Barlow was born in Bangor, Maine, to parents Ken and Theresa Barlow. He is the oldest of three children, and has two sisters, Meredith and Caroline. He grew up in Maple Grove, Minnesota, graduating from Maple Grove Senior High School in 2006. His family moved back to the East Coast summer of 2006, and Patrick stayed in Minnesota to attend St. John's University.

While at Saint John's Patrick began working as an instructor for a faculty class on classroom assessment under a grant from the Teagle Foundation. This work, directed by Dr. Philip Kramer and Dr. Ken Jones was the inspiration for pursuing a PhD in Evaluation, Assessment, and Measurement after graduating from Saint John's in 2010 with a double major in English and Psychology.

During his four years in the Evaluation, Statistics, and Measurement program at the University of Tennessee, Patrick completed a number of evaluation research projects as either principal or co-principal investigator. He also worked with Drs. William Metheny and Eric Heidel at the University of Tennessee Graduate School of Medicine where they provided a number of research and statistical consulting services for medical and pharmacy residents. This experience in addition to his time creating and teaching the graduate medical education curriculum in statistics and research methods has led Patrick to specialize in assessment in graduate medical education environments. As of March 2014, he has taken a position as a Post-Doctoral Research Associate in the Department of Surgery at the University of Wisconsin-Madison.