The Pennsylvania State University

The Graduate School

College of Education

RESPONSE PROCESS VALIDATION OF EQUIVALENT TEST FORMS: HOW QUALITATIVE DATA CAN SUPPORT THE CONSTRUCT VALIDITY OF MULTIPLE TEST FORMS

A Thesis in

Educational Psychology

by

Sarah E. Zappe

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2007

The thesis of Sarah E. Zappe was reviewed and approved* by the following:

Hoi K. Suen
Professor of Educational Psychology
Thesis Adviser
Chair of Committee

Dorothy H. Evensen
Professor of Education

Pui-Wa Lei
Assistant Professor of Educational Psychology

James Rosenberger
Professor of Statistics

Kathy Ruhl
Department Head of Educational Psychology, School Psychology, and Special Education

*Signatures are on file in the Graduate School.

ABSTRACT

When developing multiple test forms, test developers need to be concerned with whether each form is measuring the intended construct in the same manner. This study examined the agreement between three methods of examining construct equivalence on a test of legal case reading and reasoning. The three methods utilized included response process information from think-aloud procedures, expert judgments of similarity, and statistical differential item functioning (DIF) methods. The study showed that the methods did not agree about which item pairs were considered construct equivalent. The think-aloud protocols identified instances of individual items functioning in a different manner than intended, due to construct irrelevant influences on student response such as item wording, ambiguities, item-writer oversight, and failure of the items to conform to item writing guidelines. The study concludes that all three methods provide some unique value in test development. In addition, the benefits of the think-aloud procedure outweigh the costs in terms of time and money.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Thank you sincerely to each of my committee members for their help and guidance with developing this project. To my advisor, Dr. Suen, your guidance and advice throughout the years is greatly appreciated. To Dr. Lei and Dr. Rosenberger, thank you for your support and guidance regarding the statistical analyses. To Dorie, Dr. Evensen, thank you so much for being a mentor during this time. I could not have pursued this project without you!

I would also like to thank the students and administrators at the University of Pittsburgh Law School. To the administrators and staff, I appreciate the help with coordinating the data collection activities. To the students who participated in the study, thank you so much for your great effort and your valuable thoughts and comments. It was a pleasure to work with all of you.

Thank you so much to my colleagues at the Schreyer Institute for Teaching Excellence. Thank you especially to Sue Cross. Sue, we've had quite the adventure exploring new realms in the statistical world! Thank you for putting up with my tears and frustrations and for always being there for me.

Thank you to my family and friends for your support. Hey, Adrienne, guess how many pages I have now?!? Of course, thank you to my husband Steve most of all. You've definitely had to bear the brunt of this experience through the tears, struggles, and the constant ups-and-downs. You were always there to listen to me, providing me with the support and help that I always needed. I could not have done this without you!

Chapter 1

INTRODUCTION

Multiple forms are used for a variety of test types, including achievement, standardized, and mastery tests (Gusky, 1997; Schmitt, Schmitt, & Clapham, 2001) in a variety of disciplines. On a practical level, multiple forms of a test, consisting of unique but theoretically equivalent items, are often desired to reduce problems with test security and possible cheating (Cizek, 1999). From a research perspective, having multiple forms of a test is often desirable to detect possible gains in learning after an intervention by using one version as a pre-test and an alternate version as a post-test. By using two unique instruments in this design, examinees are less able to learn from the test administration itself (Kolen & Brennan, 1995). Rather, potential learning gains can be attributed to an intervention which occurred between administrations. In using alternate forms, however, a significant concern is the determination that each form is measuring the same construct.

The most recent Standards for Educational and Psychological Testing (1999), as developed by a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, provide some limited guidance on how to determine the equivalence of multiple test forms. The standards state that, "A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably" (p. 57). The committee goes on to explain that the "easiest" way to establish equivalence is by following identical steps for test construction on each form and using statistical equating methods.

Most of the research on alternate forms focuses on the statistical equivalence in terms of difficulty levels. Holland and Rubin (1982), in their extensive treatment of equating, describe the importance of determining the equivalence of multiple test forms. "Only when tests are equated can it be fair to give them to different people and treat the scores as if based on the same test" (p. 1). Indeed, in order to maintain a level of fairness for all respondents, the score on one form should be analogous to the same score on the other form. Any differences in the difficulty levels of the instruments must be compensated for in the equating process. As Thissen and Orlando (2001) note, "Because it is obvious that one form may be 'easier' than another, some adjustment to the summed scores must be made so that the alternate forms yield comparable scores" (p. 133). Processes of equating should be used in order to place scores of the forms on the same scale. However, equating does not necessarily ensure that the tests are measuring the same construct. Rather, this is a precondition of the equating process. If the test forms in fact do not measure the same construct, equating the forms is not defensible.

In effect, the equating process is not sufficient in the creation of alternate test forms. Additional evidence is needed to ensure that the forms are measuring the same construct, if the scores are to be used interchangeably. As Savage (1982) responded in Holland and Rubin's anthology on equating, "What we want to do is find out how well the tests are measuring something that we are interested in. How well they measure each other is not really relevant." (p 343-344). The question of whether forms of a test *should* be used interchangeably is a concern related to validity. Evidence must be collected to support the interchangeability of tests such that that the interpretation of a test score on

one form of the test is equivalent to the interpretation of the score on the alternate form. This type of evidence supports the construct equivalence of the different forms.

Validity Evidence and Multiple Test Forms

The 1999 Standards list five potential sources of validity evidence that may be collected to support the interpretations of test scores. Evidence based on internal structure, content, and response processes should be considered during the initial development of the test to ensure construct equivalence of multiple test forms.[1] Table 1, available on the following page, provides a visual display of these three types of validity evidence with examples for supporting the use of one test versus multiple forms. While the Standards provide descriptions of each type of evidence in the context of the development and use of one test form, each type of evidence must also be collected to support construct equivalence of multiple forms. Each type of validity evidence, while necessary to support the interpretation of a test score, is not sufficient and has some inherent limitations if used alone.

Validity Evidence Based on Content

Validity evidence based on test content refers to the relationship between the content of the test and that of the intended construct. This type of evidence is primarily collected during the test construction stage. Carefully constructed test specifications, quality written items, and thorough expert reviews of items all support the validity of the

[1] While evidence based on external variables and consequences of the test are also important to establish an argument for validity, these will not be explored in this project. Evidence based on consequences cannot be explored this early in test development. Evidence based on external variables was not explored for two reasons. First, because of the length of time necessary to administer the test used in this study, a fatigue effect would likely have been pronounced if students were asked to complete an entire battery of tests. Second, law is a specific discourse and the reasoning in law is different than general reasoning, a finding supported by Graham and Anderson (2000-2001). Therefore, other instruments intended to measure reasoning skills are not likely to correlate highly with the test used in this study.

Table 1: Types of validity evidence in supporting construct equivalence

Type of Validity Evidence	Evidence for one test	Evidence for multiple forms	Limitations
Evidence based on content validity	▪ Test specifications ▪ Item-objective congruence index ▪ Paired comparison procedure (Sireci, 1998)	▪ Test specifications guiding development of both forms ▪ Adapted version of paired comparison procedure	▪ Experts, item-writers, and test-takers have different perspective ▪ Experts focus on content rather than cognitive process ▪ Perceived subjectivity of expert judgment
Evidence based on internal structure	▪ Factor analysis ▪ Structural equation modeling (SEM) examining relationships between parts of a test	▪ Latent trait models showing similar structure between forms ▪ Differential item functioning (DIF) techniques	▪ Based on patterns of responses rather than specifics ▪ Assumptions of unidimensionality for DIF techniques
Evidence based on response processes	▪ Verbal protocols ▪ Interviews of participants	▪ Similar cognitive processes between forms as judged on verbal protocols or interviews across equivalent items	▪ Question of generalizability to larger sample ▪ Small sample size due to labor intensive nature of approach

instrument as it relates to the content of the test. While the test specifications help to define the content domain, the expert reviews of items help to support the relevancy and representativeness of each item to that content domain. A variety of methods have been developed to summarize expert judgments (i.e. Sireci,1998).

The use of experts to determine the construct equivalence of multiple test forms has some limitations. Experts are very proficient in the content and may possibly overlook wording that may not be appropriate for examinees' level of understanding or which may be possibly misinterpreted. Additionally, while experts are well-versed in the domain, they are not typically trained to identify the cognitive response processes in which respondents engage. Even if experts are given this training, this type of judgment does not identify the actual response processes of the test-taker, but only experts' hypotheses. As Leighton (2004) states,

> …[C]ognitive models of domain mastery and test specifications serve only as initial guidelines or hypotheses for defining expertise within a domain and creating test items – even if students *hypothesized* cognitive processes are taken into account in the development of these models. How students *actually* react and respond to those finally formed items… will ultimately have a defining, empirically based influence on both the revision of the construct measured and the items developed (p. 8).

Similarly, as Haney and Scott (1987) state, the content validity of a test "depends not on what…experts or critics *think* it measures nor on what item statistics say about the item but rather on how the individual test-takers perceive and react to the test or item" (p. 301).

Validity Evidence Based on Internal Structure

A second type of validity evidence is based on the internal structure of the test, or "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (Standards, p. 13). The items on a test should a have structure which corresponds to the hypothesized structure of the construct. In test development, this type of evidence is most frequently supported through factor analysis. Evidence to support the construct equivalence of multiple forms also uses latent trait modeling. For example, Loehlin (1998) discusses the application of path modeling to test the hypothesis that two tests are parallel, or share "equal amounts of a common factor" and "same amount of specific variance" (p. 98). Ding and Hershberger (2002) used structural equation modeling (SEM) to determine if equivalent forms are measuring the same construct.

Scale level modeling, such as factor analysis and structural equation modeling, is not sufficient to demonstrate construct equivalence. As Zumbo (2003) points out, tests that may be statistically equivalent at the scale level may still have nonequivalence at the item level. Therefore, additional analyses, such as differential item functioning (DIF) should be conducted at the item level to provide evidence for the equivalence of the multiple forms in terms of internal structure. For a single test, DIF techniques are most often used to determine test item bias between majority and minority groups or between genders. However, DIF techniques are not limited solely to finding biased test items. As Camilli and Shepard (1994) note, "Item bias statistics are useful for studying the internal structure of a test . . .[S]o-called item bias statistics work like item-level factor analyses to examine item functioning" (p. 154). Therefore, in addition to studies of bias, DIF

methodology can be extended to the examination of the equivalence of test forms translated into different languages (i.e. Allalouf, 2003; Allalouf, Hambleton, & Sireci, 1999; Sireci & Allalouf, 2003) and the equivalence of tests with varying modes of administration (i.e. Schwarz, Rich, and Podrabsky, 2003; Schwarz, Rich, Arenson, Podrabsky & Cook, 2002; Bielinski, Thurlow, Ysseldkye, Freidebach, & Freidebach, 2001). The potential exists for using DIF techniques to examine the equivalence of theoretically equivalent items across different test forms, in a form of item-level equating, although recent literature in testing and measurement does not indicate such a use.

DIF techniques do have some limitations that may lead one to question the conclusions relating to construct equivalence of multiple test forms. As mentioned above, scale level analyses are not sufficient to demonstrate construct equivalence across forms, as construct nonequivalence in theoretically parallel items is still possible (Zumbo, 2003). DIF analyses are then needed to be performed to ensure equivalence at the item level. However, the identification of DIF is dependent upon the use of an internal criterion of ability, which is usually the total test score on the form. If the total test score has a different meaning on one version of the test as compared to the other version, then the DIF analyses may be suspect. Test developers first need to demonstrate the internal criterion, which is usually the ability estimate as measured by the total test score, is equivalent between the forms. As Sireci and Allalouf (2003) state,

> [B]efore performing DIF analyses in cross-lingual assessment, the viability of the matching criterion (in this case, total test score) must be defended by ruling out construct bias [in order to conclude that] the total score derived for these common items represented the same construct in each language group (p. 157).

A second limitation of DIF analyses is the assumption of unidimensionality. Some DIF analyses assume a single dimension, as evidenced by the single estimate of ability. If the test is complex, incorporating several cognitive dimensions, this one unit of ability may not sufficiently capture the relationship between the items and the construct. The DIF analyses then may overlook items that are potentially not measuring the same construct, due to this limitation of the ability estimate.

Validity Evidence Based on Response Processes

Validity evidence based on response processes constitutes a new aspect to the most recent Standards for testing, although this has been previously discussed by some measurement specialists including Messick (1989). The standards describe this type of evidence as "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (p. 12). Although few details are provided on how to determine whether a test is valid based on the response processes of test respondents, the committee does provide some suggestions including the analysis of students' responses through interview techniques. "Questioning test takers about their performance strategies or responses to particular items can yield evidence that enriches the definition of a construct" (p. 12). Other types of verbal protocol methods include the use of think-aloud techniques (Ericsson & Simon, 1999; Rzasa, 2003) in which respondents verbalize their thought-processes as they complete a task.

The test takers' perspective can add additional insight into the construct of interest. This type of data collection, while infrequently used in test construction, has been more frequently utilized in survey research in order to ensure that "all respondents are answering the same questions" (Jabine, Straf, Tanur, & Tourangeau, 1984, p. 13).

Similarly, Czaja and Blair (1996) state that validation of a survey or instrument can be supported by the use of "cognitive interviews." As the researchers note, "Validity requires, first, that the questions measure the dimension or construct of interest; and second that the respondents interpret the question as intended" (p. 94). The application of verbal protocols has seen only limited use in the development of scalable instruments, such as rating scales or achievement tests, although recent investigators are encouraging this exploration (i.e. Gorin, 2006; Leighton, 2004).

As with other types of validity evidence, the methods for gathering information on the response processes of test-takers also have some limitations. One limitation is the generalizability of the sample of test-takers who complete the verbal protocols to the overall intended population of test-takers. Most test developers are unable to gather this data from large samples due to the time-consuming nature of collecting and analyzing verbal protocols. One way to reduce the problem of generalizability is to gather data until a level of saturation is collected, in which the researcher is no longer hearing or seeing unique information (Strauss & Corbin, 1998).

While not yet used extensively in test development, think-aloud procedures have been used in some testing and measurement aspects such as in the comparison of items given in different formats (Kobrin & Young, 2003) and in the identification of potential sources of DIF (Ercikan, Law, Arim, Domene, Lacroix, & Gagnon, 2004).

Validity Argument Applied to Alternate Test Forms

Validation is an argument, a collection of evidence to support a given interpretation of a test score (Messick, 1989; Kane, 1992; Standards, 1999). Messick (1989), who argues for unified theory of validity, stated that, "construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 17). When using multiple test forms, evidence must be sufficient to justify the interchangeability of test scores. Too often in the development of alternative forms of tests, the cognitive processes that examinees employ are neglected in exchange for an overabundance of statistical support for equivalence. Information on cognitive processes is potentially very rich and can add valuable and unique information to support or counter this evidence. One way to examine the construct equivalence of forms is to compare theoretically equivalent item pairs, or item pairs that should be similar in content, elicit similar response strategies, and function statistically in the same manner. As mentioned above, all the methods of demonstrating construct equivalence have limitations, including collection of verbal protocols. However, test creators must consider whether this added information has enough value to warrant integration in test development.

Research Hypotheses

The purpose of this study is to examine if there is sufficient evidence to support the development of a more complete method of determining construct equivalence in the creation of multiple test forms. This study aims to gather evidence to reveal whether each of three methods: 1) gathering expert opinion, 2) the collection of verbal protocols of respondents, and 3) DIF techniques, adds a unique value in determining the equivalence of test forms. Two forms of a test of legal case reading and reasoning will be

analyzed (Stratman, Evensen, & Oates, 2005). Methods of collecting evidence based on content (expert judgment), response process (verbal protocols), and internal structure (DIF techniques), will be used to demonstrate the construct equivalence of the test forms. Theoretically equivalent items across the two forms will be analyzed to determine if they could indeed be used interchangeably. Item pairs intended to be equivalent according to the item writers will be used as the object of measurement in the study. These item pairs should ideally elicit similar response strategies, be similar in content, and function statistically in a similar manner. The following hypotheses will be tested:

1. *The set of items identified through verbal protocols as being nonequivalent across the test forms will be different from those identified as nonequivalent by expert judgment.* Using the procedures commonly used to support validity based on content, experts will be more likely to focus their attention on the explicit content of the item rather than on hypothesized cognitive structure. Even if they focus on the cognitive processes, because experts are so fluent in the area, they will be unable to judge the processes that inexperienced test-takers will use. Additionally, experts are often not even aware of their strategies in completing a task (Lundeberg, 1985), which may hinder their ability to hypothesize the equivalence in terms of response process.

2. *The set of items identified through the verbal protocols as being nonequivalent will be different from those identified through DIF methods.* DIF analyses rely on using the total test score as the criteria by which to measure equivalence. On a complex cognitive test, such as a test of legal case reading and reasoning, ability as measured by the total test score may not be the most appropriate measure of equivalence across two

forms of the test. The think-aloud method may be better able to account for complexities in cognitive task.

3. *The set of items identified through expert judgment as being nonequivalent will be different than those identified through DIF methods.* Camilli and Shepard (1994) highlighted several studies that compared expert judgment in the identification of potential item bias to statistical DIF results. In general, experts have not been able to a priori predict which items will function differently for minority groups. A similar comparison of a priori expert judgments to DIF results for items on alternative forms has not been performed. However, it is hypothesized that the results of such a comparison will follow the line of research for DIF studies of minority versus majority participants. That is, while experts may be able to identify some sources of DIF, there will be items that are flagged by the statistical analyses that the experts do not identify.

On the test, students will be asked to read legal cases and then answer a series of multiple-choice items. Test items are categorized along two dimensions: single versus cross-case and determinate versus indeterminate. Some items focus on a single case while others require information from across several cases. Along the second dimension, test items that focus on the explicit legal information in the text are called *determinate* items. Test items that ask students to consider implicit information (such as ambiguities, silences, contradictions) in the text are called *indeterminate* items. Test items become increasingly complex as they move from single to cross-case and from determinate to indeterminate.

An additional area that will be investigated in this project is the relationship between methods of detecting potentially non-equivalent item pairs and the item type.

Specifically, do certain methods have a greater tendency to flag item pairs of a particular type (cross-case versus single case; determinate versus indeterminate)? One potential theory is that as the complexity of the items increases, from single-case to cross-case and from determinate to indeterminate, the think-aloud procedure may become more likely to identify differences among item pairs. As the cognitive load required from the test-taker increases, perhaps the types of response processes that emerge in the verbal protocols will become more varied between theoretically equivalent items. If this is the case, the expert judgment may fail to detect this type of difference, as the experts may be more focused on the content of the items themselves. The use of experts to identify similarities among these types of items may be confounded by their focus on the cases. Experts may pay more attention to nuances and differences in the content of the cases rather than identifying the similar, cognitive skills necessary to answer each of the items. In addition, the DIF procedures may fail to capture these more sophisticated differences in response process, as the statistics are dependent upon the unidimensional ability estimate. In terms of this general theory, two additional hypotheses will be tested:

4. *As compared to the other methods, the think-aloud procedure will identify more of the cross-case items as being nonequivalent.* As the items move in complexity from single-case to cross-case, the think-aloud procedure will be more sensitive to differences in theoretically equivalent item pairs.

5. *As compared to the other methods, the think-aloud procedure will identify more of the indeterminate items as being nonequivalent.* As the items move in complexity from determinate to indeterminate, the think-aloud procedure will be more sensitive to differences in theoretically equivalent item pairs.

For over a decade, researchers in education and psychology have stressed the importance of combining concepts related to both cognitive psychology and measurement in the development and creation of tests (i.e. Snow & Lohman, 1993; Pellegrino, Baxter, & Glaser, 1999). This promise of a "New Generation of Tests," as termed by Frederiksen, Mislevy, and Bejar in 1993, has yet to emerge. This study attempts to build a bridge between these two fields by examining whether response process information should be incorporated into the validity argument.

In addition, the creation of alternative test forms, while quite frequently used, has not been studied extensively, particularly regarding the issue of validity. While psychometricians have developed very precise methods for statistically equating alternate test forms, the issue of construct equivalence has not been given proper emphasis in the literature. This project will help to provide input in how to utilize response process information when developing multiple test forms.

Chapter 2:

LITERATURE REVIEW

Throughout the modern history of testing, alternate test forms have been common. If alternate forms are utilized, test developers need to ensure that each form is measuring the intended construct in the same manner. In order to determine that two forms are construct equivalent, a validity argument needs to be crafted from various types of evidence. This chapter will describe several types of validity evidence that need to be collected and how this evidence can be used in the creation of alternate test forms. The first section of the chapter will provide a discussion of validity evidence based on response processes. Specifically, this section will contain a review of the literature on how the think-aloud protocols are used in test construction and validation as well as a discussion of the potential use of think-aloud protocols in the construction of alternate forms. The second section of this chapter will discuss validity evidence based on test content focusing on judgmental approaches involving the use of experts. In the third section of this chapter, validity evidence based on internal structure of the test will be discussed focusing on the statistical procedures that have been used in the creation of alternate test forms.

Use of Think-Alouds for Validity Evidence Based on Response Process

As mentioned in Chapter 1, the concept of validity related to response processes was included in the 1999 *Standards for Educational and Psychological Testing.* The standards, however, provide only some very basic guidelines on how this information may be helpful and on how to collect this type of data. The standards define evidence based on response process as the "fit between the construct and the detailed nature of

performance or response actually engaged in by examinees" (p. 12). The Standards state that for example,

> …if a test is intended to assess mathematical reasoning, it becomes important to determine whether examinees are, in fact, reasoning about the material given instead of following a standard algorithm. For another instance, scores on a scale intended to assess the degree of an individual's extroversion or introversion should not be strongly influenced by social conformity (p. 12).

In other words, evidence based on response process needs to support that the test measures the intended construct in that students' responses are eliciting the intended reasoning. In addition, evidence based on response process needs to show that the responses are not being influenced by construct-irrelevant factors. Standard 1.8, the specific standard that relates to the collection of response process follows:

> If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided (p. 19).

The standards do not discuss in great detail how response process information should be collected. They do provide a few examples of gathering evidence such as questioning test-takers, documenting performance through tracking eye movements, analyzing relationships among parts of a test, and examining the ratings of judges.

Other than the brief statements quoted above, the Standards do not provide much information on why response process information is important in test development nor do

they provide practical guidance to the test developer as to methods collecting this type of evidence. The remainder of this section will review the literature regarding 1) arguments on why response process validity is a necessity in test development and 2) how response process data can be collected. The primary focus of the latter section will be on the use of verbal protocols, including suggested methodology and associated problems or concerns.

Benefits of response process information in test development.

While the concept of response process validation is relatively new, research throughout the past several decades has begun to shift in the direction of linking psychometric theory and practices to research in cognitive development (i.e. Frederiksen, Mislevy, & Bejar, 1993; Pellegrino, Baxter, & Glaser, 1999; Snow & Lohman, 1993). Several researchers have emphasized the importance of collecting this type of information in test development in order to support the validity argument and to learn more about the construct of interest.

One of the primary arguments for the collection of response process data concerns the issue of alignment of the test developers' intentions and the nature of the cognitive activity actually elicited during test administration. In their examination of alignment between test developers' intent and elicited information from students taking a middle school science state assessment, Ferrara, Duncan, Freed, Velez-Paschke, McGivern, Mushlin, Mattessich, Rogers, & Westphalen (2004) provide an informative definition of this concept of alignment:

> We define alignment as the degree of correspondence between content area knowledge, content area skills, broader cognitive processes, and response

> strategies (i.e., Knowledge, Skills, Processes, and Strategies: KSPS) that (a) test developers intended to assess, and (b) examinees bring to bear when they respond to test items (p. 1).

The authors further discuss that when the intended and observed KSPS align, construct validity for this item is supported. When the intended and observed are misaligned, this may indicate that an item is not construct-valid and thus inferences made from test scores could potentially be suspect.

Misalignment is not necessarily uncovered through the use of test specifications, expert judgment, or statistical analyses. Gorin (2006) recognized the limitations of the traditional methods of test development and encouraged more extensive examination of items in order to explore the cognitive processes employed. As she states,

> Test developers must consider even more rigorous methods of item examination before operational use that provides explicit evidence regarding the skills, knowledge, and processes measured by the items. Item design should proceed from sources of cognitive complexity related to the construct of interest, rather than unrelated surface features (p. 33).

A 1997 study by Gierl supported that test developers are often not able to predict the type of cognitive processes that are elicited by items. In his study, Gierl used think-aloud protocols to examine the putative cognitive processes elicited on a test of mathematics and compared these to the anticipated cognitive processes as described in the test specifications. The intended and observed cognitive processes were only found to coincide 54% of the time. As he noted, "Cognitive features such as strategy selection and higher order thinking are often poorly evaluated because item writers are not trained to

identify the cognitive processes required to solve items "(p. 26). Additionally, he noted, "If test developers hope to assess students' cognitive processes successfully, researchers should use think-aloud protocols to evaluate directly the problem-solving strategies that *students* use to solve achievement test items" (p. 31). Thus, when response process information is included in test specifications, test developers should still examine the actual process elicited as misalignment can often occur.

Especially when a test is designed to be cognitively complex and to require specific cognitive processes in order to correctly respond, items should be closely examined to see if they perform as expected. Linn, Baker, and Dunbar (1991) in an examination of the cognitive processes utilized in performance assessment agree with this argument. As they stated,

> It should not simply be assumed, for example, that a hands-on scientific task encourages the development of problem-solving skills, reasoning ability, or more sophisticated mental models of the scientific phenomenon. Nor should it be assumed that apparently more complex, open-ended mathematics problems will require the use of more complex cognitive processes by students (p. 19).

Although they do not go into detail on how test developers should examine their tests with regard to the cognitive processes, the authors note that "analysis of open-ended responses" (p. 19) can be quite valuable and informative.

A second related reason for the importance of collecting response process information concerns the possibility of the introduction of construct-irrelevant influences on students' responses. Surface features of items, unrelated to the intended skills to be measured, can alter the cognitive processes that are elicited during the testing process. In

the item generation literature, the content of an item can be categorized along two dimensions (Gorin, 2005). *Incidentals* are those aspects of an item that are considered to be surface characteristics and can be manipulated without affecting the cognitive processes elicited. For example, if a student is given a hypothetical mathematics word problem, the names, locations, and exact numeric values might be considered incidentals. *Radicals* are those aspects of an item that correspond to the cognitive processes that the item is intended to elicit. For example, the algebraic formula necessary to complete the mathematics word problem might be considered a radical. Gorin explains these features below:

> Incidentals are useful when many items with identical processing components that may affect difficulty are generated, but perceptually different such that an examinee cannot perceive the similarity through surface processing alone. Radicals are critical components of the cognitive model…that are manipulated specifically in order to produce a predictable change in cognitive processing (p. 352).

If one considers the concept of construct relevance and irrelevance using the terminology associated with item generation, validity evidence should support or confirm that the incidental information included in an item did not unduly influence the cognitive response processes elicited. In other words, specific surface features of an item should not introduce construct irrelevant influences on a students' response. As Leighton and Gokiert (2005a) note, "…seemingly simple words and phrase choices in test items could a) disrupt how students understand the nature of the task, b) derail students' cognitive processing, and c) undermine the construct validity of individual items and the test as a

whole" (p. 3). In a related article (2005b), Leighton and Gokiert note that while these features may affect student performance, these may not be evident to the test developer. As they stated,

> Test items with problematic features such as words with multiple meanings or missing story information can *potentially* slow students' cognitive processing and lead them to generate faulty assumptions and incorrect responses. These item features are perhaps barely discernible to test developers who bring a very different knowledge base to bear on item development (Leighton and Gokiert, 2005a, p. 21).

Examples of surface features of an item that might influence a students' response may include ambiguous words, specific unclear phrasing, or clues that lead students to a specific response even though they may be lacking in the intended skill set required to answer the item. Evidence based on response processes can support that construct-irrelevant features of an item are not influencing students' responses.

Think-aloud protocols have been shown to be successful in supplementing other sources of information about a test. Two studies were found that used think-aloud procedures to help explain statistical sources of DIF between different groups of examinees. First, Uiterwijk and Vallen (2005) used both experts and think-aloud procedures to identify possible sources of DIF between native Dutch students and second generation immigrants (SGI) on a Dutch achievement test. The expert analyses focused on the linguistic features of the items, and were often unable to state with much certainty the sources of DIF between the two groups. "The great majority of the 16 experts concluded that it was extremely hard to determine which element in an item might cause

difficulty for SGI students" (p. 223). The use of retrospective think-aloud procedures, while not described in great detail, helped to contribute to an understanding of the linguistic features of the items that made them more difficult for the SGI group. Unfortunately, while Uiterwijk and Vallen provide descriptions on the potential sources of DIF for specific item pairs, they do not detail how the conclusions were arrived. In other words, the researchers do not differentiate whether the potential sources of DIF were determined from expert judgment or from the students' verbal protocols. In addition, no information is provided on whether the information from experts agreed or disagreed with the information gathered from the protocols on the potential sources of DIF.

Ercikan and her colleagues (2004) used think-aloud procedures to supplement statistical analyses in order to help identify potential sources of translation DIF in tests in English versus French. Ercikan et al. found that the think-aloud procedures helped to identify DIF sources approximately 35% of the time. As they note:

> Think-aloud protocol approach has proven to be useful in identifying sources of DIF not as a preferred method but more as a complementary method to other methods such as judgmental reviews and statistical methods. Yet it is clear that some of the supporting evidence obtained from using this approach, in support of the hypothesized source of DIF, could not have been obtained using neither judgmental reviews nor statistical analyses (p. 14).

The two studies described above suggest that think-aloud protocols can supplement DIF analyses and to potentially identify sources of DIF. However, in spite of the use of think-aloud protocols to identify sources of DIF, it is not likely that statistical

and think-aloud methods will agree completely if used exclusively in the identification of potentially problematic items. The question of agreement was not examined in either of the two studies above. Both employed think-aloud methods as a post-hoc analysis of the potential DIF sources. As such, this question of agreement, when each method is used independently, rather than as a post hoc explanation of the other method, still needs to be explored.

A final benefit of this type of method to gather validity evidence concerns the ability to compare processes that are elicited across different types of test formats and potentially across test forms. For example, Kobrin and Young (2003) used think-aloud protocols to identify that cognitive processes were equivalent on the same test administered through the computer or via traditional pencil and paper. They had hypothesized that there could potentially be differences in cognitive processes due to greater demands on working memory in the computerized context. In fact, Kobrin and Young found no differences between the two modalities. Martinez and Katz (1995) also provide an example of a study in which the construct equivalence of item types could be examined. Specifically, they explored whether multiple-choice versus constructed response items on an architecture assessment elicited similar or different cognitive processes collected through think-aloud procedures. Martinez and Katz found that while the items tended to elicit similar processes, differential difficulty between the items in different modalities was often a cause of the recall demands in constructed response items. These studies suggest that response process information can be used to support that alternate forms of a test are equivalent.

In summary, validity evidence based on response processes is important to 1) demonstrate that the item is measuring the intended construct, and 2) that the item is not affected by construct-irrelevant influences on response. Correct responses to the item should indeed mean that students possess the skills, knowledge, or processes that were intended by the test developer. Incorrect responses should mean that students do not possess the skills, knowledge, or processes intended to be measured by the item as opposed to some alternative strategy that may lead to the correct answer without possessing these characteristics. Think-aloud data can be used as supplemental data to explain statistical analysis or can be used on its own in the validity argument. In general, response process has not been examined much in the argument for construct equivalence although the potential exists as based on some studies that examined equivalence across item formats.

Collection of think-aloud data for evidence based on response process.

The methods of collecting evidence based on response processes are not yet well defined within the testing and measurement literature. Gorin (2006) describes a general method to consider when developing a new instrument:

> Evidence of this fact can be gathered using a variety of courses. First, develop preliminary items that you believe achieve this goal. Then, test your hypotheses about the properties of the items in several ways – verbal protocols, experimental manipulations, and visual eye-tracking data. The procedures may be more time consuming, but they should yield maximally accurate, valid, and useful test scores (p. 33).

A hypothesis about what items are measuring needs to be developed, then the methods for collecting response process information should follow. The most common sources of data on response process are verbal protocols consisting of either think-aloud data or structured interviews.

A comprehensive source of information on how to collect think-aloud data is provided by Ericcson and Simon (1999). Their directions on how to collect think-aloud protocols seem to be generally agreed upon in the literature. In providing instructions to the students, the proctor often provides a short practice task to warm students up. The proctor then tries to remain as unnoticeable as possible during the task to avoid distraction. If a student is silent for a period of time, prompts are kept as neutral as possible in order to avoid leading the students' verbalization in a certain direction. Some sort of audio- or video-recording is utilized to capture responses. In general, these directions seem to be standard across most think-aloud studies (Afflerblach & Pressley, 1995).

Limitations of think-aloud methodology.

The difficulty in utilizing think-aloud protocols in constructing the validity argument seems to often stem from how to analyze the data. Throughout the literature, there are different methods employed both in terms of the guiding research questions and in terms of the actual methods of data analysis. For example, Ferrara, et. al (2004) used think-aloud protocols to examine students' cognitive processes on a state science assessment examining alignment in terms of the intended and observed knowledge, skills and processes necessary to answer the items. They used three questions of alignment to guide their analyses. Ferrara, et al. noted that often the intended processes, skills, and

knowledge did not align and that when students were lacking in one of these areas they often resorted to use of prior knowledge to select their responses. Enright, Tucker, and Katz (1995) employed a different approach in examining the think-aloud protocols, in their study examining the differences in cognitive processes elicited by various item types on the Graduate Records Examination (GRE) general test. In this study, the researchers used the question of how students represented each problem and how they solved each problem to guide their analyses. Enright, et al. found that these processes of problem representation and problem solution varied based on item type. Norris (1992) used yet a different approach to analyze the think-aloud data, which he gathered from students completing a test of critical thinking. The protocols for each item were analyzed using a rubric to assign "thinking scores" which were correlated with the students' performance on each item. In addition, the protocols were analyzed to identify whether the examinees understood the task, whether critical thinking was used to select an option other than the key, and whether there were clues in the items that could lead students to the correct response without using critical thinking. As yet another example, Kobrin and Young (2003) in their investigation in the construct equivalence of computerized versus pencil and paper instruments, used content analysis, following a coding strategy detailing the frequency and types of 1) cognitive processes used during the initial reading of the passages, 2) cognitive processes while answering the test questions, 3) search strategies finding information in text passages, and 4) overall test-taking strategies.

Thus, in conclusion, while the methods for collecting think-aloud protocols may be relatively well established, the methods of analyzing the transcribed protocols vary

widely in the literature. Perhaps the best method of analyzing the data depends upon the hypotheses of what the test is intended to measure.

One of the major limitations of the think-aloud protocol as a method of data collection concerns the issue of generalizability of results. Because of the time-consuming nature of the task and the fact that individualized administration of the test is required, most think-aloud studies only use small sample sizes. These small sample sizes force researchers to consider whether the findings from these studies can be generalized to the larger population of test-takers. The answer to this question is "no" according to the technical definition of generalizability.

Another argument often made against the use of think-aloud procedures concerns whether or not the cognitive processes that examinees employ are altered by the introduction of thinking out loud. If indeed the cognitive processes as reflected by the verbal protocols are changed because of the process of thinking out loud, then the inferences made are not valid. On a test of critical thinking, Norris (1990) randomly assigned students to complete a test of critical thinking under one of several conditions including thinking aloud, retrospective interviews, and regular test-taking conditions. No differences were found in performance among the groups supporting his inference that perhaps no differences in cognitive processes were present among the groups. This is supported by other findings in cognitive psychology, as reviewed by Ericsson and Simon (1999) who emphasized that the think-aloud method does not alter cognitive processes individuals utilize when completing a task. The verbalization is an indication of what the subject is able to do. As Ericsson and Simon (1984) argue, "the report 'X' need not be used to infer that X is true but only that the subject was able to say X – i.e., had the

information that enabled him to say 'X'" (p. 7). As Deegan (1995) further explains, "In other words, subjects do not usually report what they are not doing" (p. 168).

In general, think-aloud protocols are not used in most test development procedure possibly due to the time-consuming nature of data collection and analysis. After recommending the use of think-aloud procedures to support that items are measuring intended cognitive processes, Haladyna (2004) noted that, "Unfortunately, we see too few reports of this kind of validity evidence in all achievement testing programs" (p. 201). However, this information may be critical to identify whether test items are functioning as intended and to reduce possible construct-irrelevant influences on student responses.

Use of Expert Judgment for Validity Evidence Based on Content

According to the Standards, validity evidence based on test content concerns the following:

> …[T]he relationship between a test's content and the construct it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring (AERA, APA, NCME, 1999, p. 11).

This type of validity evidence is generally supported through test specifications of the content domain, which would include the knowledge, skills, and possibly cognitive processes that the test is intended to measure. The test specifications act as a blueprint for how the construct is represented in the test format. While the specifications are an important part of supporting this type of validity evidence, the primary source of validity evidence in this area typically stem from the use of experts.

Methods of gathering expert judgment.

The use of experts has had a long-standing tradition in test development. The most frequently utilization of expert judgment is in gathering evidence to support validity based on test content (i.e. Sireci, 1998; Popham, 1992), determining cut-scores and standards for passing a test, (i.e. Plake & Impara, 2001), and reviewing instruments for possible test bias (i.e. Allalouf, Hambleton, & Sireci, 1999). Generally, the use of experts in the development of multiple test forms has not been covered extensively in the literature. Therefore, this section details the methodology and issues related to expert judgment for the development of one test form. Logically, these methods and issues can extend beyond to the development of multiple test forms, although additional research is necessary to refine methodology and increase understanding.

The 1999 Standards provide some limited guidelines on when and how to make use of expert judgment in test development. In establishing the validity argument, gathering evidence related to test content relies most heavily on the judgment of subject-matter experts. According to the Standards, evidence based on test content can come from "expert judgment of the relationship between parts of the test and the construct" (p. 11). Experts may often be asked to review test specifications and items, to determine if each is representative and relevant to the domain being measured. In this review process, experts are often able to "point to potential sources of irrelevant difficulty (or easiness) that require further investigation" (p. 12). If experts are used for sensitivity reviews, they are asked to identify characteristics of the items which could potentially be "inappropriate, confusing, or offensive" for a particular group of people. According to

the standards, when working with experts, demographic information including qualifications and relevant experiences should be well documented.

In developing the argument for validity based on test content, the item-objective congruence method has been used extensively, originally developed by Rovinelli and Hambleton (1976). In this method, experts are asked to use a simple rating scale in determining how well an item matches its intended objective from the test specifications or blueprint. Experts rate each item using a 3-point scale as being correctly classified, incorrectly classified, or that a determination could not be made due to uncertainty. The level of agreement among the raters is then used to calculate an index of congruence between the experts' perceptions of the item classification and the classification originally designated at the initial item development.

Since that time, other researchers have refined this item-congruence method to some degree. For example, Turner and Carlson (2003) recognized that a limitation of the strict item-congruence method, as introduced by Rovinelli and Hambleton, is its failure to adjust for items that measure more than one dimension or classification on the test blueprint. Turner and Carlson suggested that experts rate each item in relationship to its classification for multiple objectives and provide an adjustment to the calculated item-congruence index. This adjustment allows for the collection of validity evidence for items known to be multidimensional in nature.

Sireci (1998) argues that the item-objective congruence method is limited by the introduction of an a priori structure regarding the classification of the items. He argues that the limitation of the item-objective congruence method is that a predefined test structure is imposed upon the experts, thus opening the possibility that ratings may be

influenced by certain biases, such as social desirability. Because of this threat to the interpretation of the ratings, Sireci proposed a less constrained approach in which experts are simply asked to rate the similarity of items on a test. As he states,

> The paired comparisons procedure is a valuable method for discovering individuals' perceptions of the objects under investigation without informing them of what is being studied…[The task] is intentionally ambiguous. The directions do not impose or suggest strict criteria for conducting the ratings (p. 306).

In this method, experts are asked to rate the similarity of all possible pairs of items within a test. The latent data structure is then portrayed visually using multidimensional scaling (MDS). Sireci was able to use the results of the MDS to support that the structure of the test matches the theoretical structure of the intended domain.

While the paired comparison method is currently the most complete in terms of collecting validity evidence relating to content, the procedure can be quite cumbersome for the experts to complete due to the requirement of rating all possible pairs of items. For a test as short as 30-items, experts would have to rate 435 total pairs of items. Given such a lengthy task, test developers need to be concerned with fatigue effects and possible rater drift. While Sireci does provide some suggestions on how to reduce the number of items, such as selecting sets of representative items and utilizing incomplete designs, the task is still likely to be lengthy and cumbersome for the experts to complete.

Surprisingly, while the collection of validity evidence related to test content is well documented for the creation of one test form, no scholarly publications were found that discuss the extension of these techniques to the construction of alternate forms of the

same test. One could expect that these methodologies could be used to support that each test form measures the same construct. For the item-objective congruence index, experts could be asked to rate items from both forms of the test to determine the level of correspondence between the theoretically equivalent items on the test forms and their intended classification from the test specifications.

Benefits and limitations of using expert judgment in test development.

Likely, the use of expert judgment will continue indefinitely in the use of test development. As Berk (1995) noted, "Human judgment is the foundation of every method developed since prehistoric times." The use of expert judgment in test development has some strengths as well as obvious limitations. Two areas of questions emerge when considering the use of expert judgments in test development. The first area concerns the accuracy and reliability of expert judgments of items. In other words, are experts able to accurately predict or detect certain characteristics of items? Are these ratings generalizable across different panels of experts or across time? The second area of questioning concerns the characteristics of expert judgments and of items. What processes do experts employ when making decisions about items? What characteristics of items do experts attend to while performing their ratings? What external variables may influence their decisions in ratings? These two general areas of questioning are discussed below in light of the available literature on test-development as well as standard-setting. Although standard-setting is somewhat external to the research questions being posed in this study, the literature on this is more complete and may shed some insight into influences on expert decision making.

Regarding the question of the accuracy and reliability of expert judgments, one needs to look at the recent history of how expert judgments have been utilized. In 1982, Thorndike discussed how judgment can be used to establish prior knowledge about item parameters to be used in the equating process. Confronted with the problem of establishing equivalent test forms for civil service examinations, Thorndike began to "explore the possibility of using groups of judges to estimate the difficulty of test items" (p. 310). Thorndike and his colleague found that experts were able to accurately rank items in terms of their difficulty levels relative to one another. Establishing the estimation of a more absolute level of difficulty for items on a test was harder to achieve and required various procedures, such as the utilization of anchor items with already established difficulty indices, to enhance experts' ability to accurately predict item difficulties. Thorndike suggested that in order to obtain the best difficulty estimates for test items, experts should have sufficient knowledge in test development and be well trained in the estimation procedure.

Since Thorndike's study, the use of experts in the estimation of item difficulty levels has been relatively common in both equating (i.e. Mislevy, Sheehan, & Wingersky, 1993) and standard-setting (i.e. Plake & Impara, 2001). In standard-setting exercises, experts are asked to review items and determine the probability that a minimally competent individual would be able to correctly respond. In essence, most standard-setting exercises are asking the experts to estimate the level of difficulty for each item in the context of imagining the ability level of a minimally competent individual. This information is then compiled in setting passing or cut-scores for the test. Plake and Impara (2001) investigated both the reliability and validity of expert judgments of item

performance estimates on a certification test in financial management. Regarding reliability, the researchers found that estimates of item difficulty were similar across different panels and were likely to be similar even when ratings occurred several years apart. Regarding the accuracy of the experts' judgments, Plake and Impara found that the estimates of item proficiency of the minimally competent candidates were quite similar to empirical data collected from a large sample of test-takers.

While both the Thorndike and the Plake and Impala studies provide support that experts can accurately predict the difficulty level of items, these only provide limited evidence that the use of expert judgments can enhance test validity. Another area in which accuracy of expert ratings has been explored is in the study of item bias. Studies comparing the results of expert sensitivity reviews have not always matched up to statistical methods of identifying item bias (see review in Jensen, 1980). According to Camilli and Shepard (1994), "The failure of judgmental methods to provide a satisfactory means of screening test items for differential difficulty gave impetus to the development of statistical item bias procedures" (p. 136).

Englehard, Davis, and Hansche (1999) investigated whether or not experts were able to detect items that were known to contain cultural and technical flaws. After an extensive training period, experts were asked to review a set of items which had been known to contain these types of flaws. The researchers found that the experts were generally quite adept at being able to identify problematic issues within the items, particularly those containing cultural flaws. While this does provide some hope about the ability of experts in the review process to detect flawed items, Englehard, et al. acknowledge that the study may not be generalizable to other item review panels or

across other item sets. Given that the experimenters knew the flaws that would appear within the items to be reviewed, the specific training provided to the experts could very well have primed them to identify certain aspects within the items. In addition, the types of features that were problematic within the items were known to the researchers and were perhaps more obvious to detect.

In conclusion to the first area of questioning, the results are mixed as to whether experts are able to make accurate and reliable ratings concerning test items. There is evidence from Plano and Impala that judgments of items in standard setting exercises are generalizable across panels and across time. Evidence also supports that experts are able to predict the difficulty level of items to some degree. There is mixed support as to the accuracy of experts in their ability to detect items that function differently between groups. However, with training, experts may be able to detect known cultural and technical defects in items.

The second area of questioning explores the characteristics of both experts and items which may influence judgments. Two studies were found in standard-setting which explores some of the characteristics of experts that may influence decision making. First, Skorupski and Hambleton (2005) present an interesting study examining the processes that experts engage in when participating in a standard setting exercise. In the study, experts were asked to use the item mapping method in order to identify the probability that students at varying levels of ability would be able to correctly respond to an item. At various points during the exercise, participants were asked to complete questionnaires requesting information about their thought processes in making rating decisions. Generally, the focus on the study was collecting data on how confusing or understandable

the experts perceived the task at hand. They found that panelists who were not clear on the procedure or how they should rate the items were more variable in their ratings between rounds. While Skorupski and Hambleton draw some interesting conclusions on how standard setting exercises can be conducted in the future, they do not focus on the types of thought processes that influenced experts' decisions. They also acknowledge that some additional research into the cognitive processes of experts on standard setting panels is necessary in order to better understand the decision making process.

Ferdous and Plake (2005) performed a similar study examining the thought processes of experts completing a standard-setting exercise. Following their estimation of item performance, the raters were divided into groups based on their total score and asked to participate in a focus group on the influences of their ratings. The group of experts who rated the items as most difficult and the group who rated the items as least difficult were more likely to consider prototypical students in their own classes when estimating the item performance. Ferdous and Plake found that those individuals who rated items as the least difficult were more likely to be aware of the consequences of the test in terms of No Child Left Behind legislation.

Both the Skorupski and Hambleton study and the Ferdous and Plake study provide interesting methods into the factors which could potentially influence experts' ratings on the difficulty of items. However, neither study focuses on the characteristics of the items but rather focuses on the characteristics of the experts in terms of their background knowledge and level of comfort with the task. Other studies were found that explored the characteristics of items that experts may consider when making judgments. For example, a study by Enright and Bejar (1989) found that experts in analogy writing

were quite adept at predicting the difficulty levels of the items. The experts were asked to provide information as to what attributes of the item influenced their ratings. "While vocabulary difficulty plays a role in determining item difficulty, other item attributes such as rationale difficulty, stem-option similarity, and syntactic-order of word pairs are also important" (p. 16). A limitation of this study is that these item types, while often requiring complex reasoning skills to correctly respond, are probably easier for experts to judge in that, by nature, their format is very consistent without much superfluous information which could influence ratings. In addition, the experts participating in this study were not subject-matter experts in the usual sense; rather, they were test-development experts who were highly trained in writing analogy items. These types of experts are likely to employ different reasoning strategies than subject-matter experts within a domain. While this study does have some limitations, it does provide some insight into characteristics of items that experts may attend to. Likely, having extensive familiarity with the item types would be helpful for experts of any type.

Hambleton, Sireci, Swaminathan, Xing, and Rizavi (2003) asked five experts to describe the characteristics of items which influenced their estimates of the difficulty level of items from two subscales of the Law School Admission Test (LSAT), namely reading comprehension and logical reasoning. The characteristics that experts listed that influenced their estimation included: complexity of the stem, complexity of the distractor, the necessity of reading the entire passage, the question reaching beyond the explicit information in the reading passage, difficulty responding to the question as the expert, and the reading load of the stem and options. Other characteristics included word count, placement of key, category of the item, and attractiveness of the distractors. One

panelist stated that he relied heavily "on intuition" (p. 13). Hambleton et al. note that, "[p]anelists brought to the task their own experiences and ideas about how to judge item difficulties" (p. 23). In other words, no consistent method was used.

In a study examining sources of translation DIF, Sireci and Allalouf (2003) asked experts to hypothesize why certain pairs of items functioned differently on the Hebrew versus the Russian form of the Psychometric Entrance Test (PET), a university admissions test used in Israel. The following reasons were listed as possible sources of DIF: changes in difficulty of words, changes in content, changes in format of the item, and differences in cultural relevance. Experts were unable to hypothesize sources of DIF 16.7% of the time.

One question of particular interest in this study is whether experts consider the cognitive complexity of items when performing item reviews. O'Neil, Sireci, and Huff (2003-2004) asked expert science teachers to rate the consistency in content between two forms of a 10^{th} grade science assessment. The purpose of this study was to ensure that differences in scores over time were due to "student or instructional characteristics, rather than to changes in the test. Although test forms from different years are statistically equated to maintain a common scale across years, parallel content is a prerequisite to score equating" (p. 131). The teachers were first asked to rate the similarity of items within each test. This similarity rating task did not provide any predefined criteria for the experts to base their ratings on. O'Neil et al. argue that this helps to reduce the possibility of response biases influencing their ratings. Following this task, the experts were asked to classify items using Bloom's taxonomy and the five content areas the test was intended to measure. Using multidimensional scaling (MDS) and following up with

short questionnaires, the researchers found that the primary characteristic influencing the experts' similarity ratings concerned the cognitive complexity of the item, based on their classification of the item into the levels of Bloom's taxonomy. The other characteristics which influenced ratings included item content, item format, and graphical components. O'Neil et al. note that while the content areas were found to be relatively consistent between the test forms, the experts felt that there were some differences in the cognitive complexity of items from one form to the other. This study shows support that experts can potentially consider cognitive features of items when making ratings or judgments. However, the primary limitation of this study concerns the characteristics of the experts as being science teachers who have likely been exposed previously to various taxonomies of cognitive skills. Whether the findings of this study are replicable needs further investigation with different types of tests and experts. In addition, several of the experts participating in the study had been previously involved with the development of the test items, which may have primed them to look for features related to the cognitive complexity of the items.

Some of the limitations of expert judgment include the difficulty of making generalizations from such a small sample. In addition, experts approach the task from a certain perspective. They have the necessary domain knowledge to respond to the question and may miss important item features that may influence test-takers' responses. However, the use of expert judgment is critical in supporting test validity related to content. According to Berk, (1995), "Given enough time, judges can be trained to do just about anything. [They] can rate individual items, exercises, clusters of items, or profiles of behavior" (p. 100). From the literature reviewed, experts seem to be quite capable,

with training, of estimating the difficulty levels of items. They have potential to identify both cultural and technical flaws, although some studies have not yielded a great match between statistical and judgmental methods of detecting DIF. Characteristics of the experts, such as comfort with the task, background knowledge, and understanding of possible test consequences, have been shown to influence perceptions of items. Although this area has not been investigated extensively, the features of items also likely have an impact on the judgments of experts. The literature on the use of experts in various aspects of test development is still quite limited. Additional research is necessary, particularly in the extension of the methods used in traditional test development to the development of alternate test forms.

Use of Statistical DIF Techniques for Validity Evidence Based on Internal Structure

Validity evidence based on the internal structure of a test generally comes from statistical techniques such as factor analysis, structural equation modeling, and DIF analyses. These same techniques can be applied to the comparison of alternate test forms. Unlike the use of experts or the analysis of response processes, statistical analyses have been used heavily in the development of alternate test forms. The use of equivalent or alternative forms in testing can be linked to the classical testing literature on the estimation of reliability. In addition to test-retest procedures, one of the early methods to estimate reliability was to examine the correlation between two forms of test which were constructed using the same test specifications. In essence, the test developer attempts to create parallel versions of the same test, constructing them to be as similar as possible. According to Nitko (1994), when tests are carefully constructed using the same guidelines, "the reliability coefficient reflects both the equivalence of the assessment

techniques and the stability of students' performance" (p. 64). According to this logic, a high correlation coefficient between the two forms would support the interchangeability of the forms. Nitko does acknowledge that the use of the alternate-forms reliability coefficient does not take into account differences in the difficulty level of the test. However, the construct equivalence of the test forms is assumed. As Nitko notes,

> Ideally, scores from parallel forms of a test should a) have equal observed score means and standard deviations, b) measure students with equal accuracy (i.e. have equal standard errors of measurement, c) correlate equally with other measurements, and d) measure the same attribute in precisely the same way (p. 65).

However, these assumptions are quite lofty and cannot be assumed through simple correlations. As Suen (1990) points out, "Although the two versions are supposed to be equivalent, there is no guarantee that they do in fact meet the parallel tests assumptions (p. 33).

With the advance of more sophisticated methods of estimating reliability, few researchers today use the coefficient of equivalence when estimating reliability (Hogan, Benjamin, & Brezinski, 2003). However, researchers have still continued to make use of alternate test forms for a variety of reasons. As such, statistical procedures beyond simple correlations have been developed in order to examine the question of whether multiple forms of a test can be considered equivalent. Much of this literature can be drawn from research on equating. While equating is not performed in this research study, the background of equating and its purposes are pertinent to the study of construct equivalence and thus will be reviewed briefly.

Use of equating for alternate test forms.

As mentioned in Chapter 1, the equating process is used to create a correspondence of scores from one form of the test to the other. In other words, "[e]quating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 1995, p. 2). Lord (1980) discussed four conditions necessary in order to make the claim that two test forms are equivalent. These four conditions are that 1) the tests measure the same construct, 2) the distribution of scores on one form must be the same as the distribution of scores on the other form after the scores have been transformed, 3) equating transformations must be symmetric, and 4) equating transformations must be population invariant. The first condition that Lord suggests is generally assumed and is based on utilization of the test specifications to develop both forms.

A variety of research designs have been developed to collect data necessary for the equating process. All of these methods necessitate commonality in the data collected for the two test forms, either in providing the same items to different groups of people or in providing different items to the same groups of people (Suen, 1990). In single-group designs, individuals are presented with both forms of the test usually with counterbalancing in order to reduce any order effects that may occur. Because it is often difficult to administer both forms of the test to the same group of people and because of the potential for learning from the administration of one form to the next, usually data from separate groups are analyzed in the equating process. However, there needs to be some commonality in the items administered to both groups. A common item nonequivalent group design can be used in which separate groups receive only one form

of the test but with some common items constituting an anchor test. This type of design is quite common as it requires participants to only be tested one time (Kolen & Brennan, 1995).

Following the collection of data using one of the methods or a variation described in the previous paragraph, statistical methods are used to place the two forms of the test along the same scale. One method of equating is the three-parameter item response theory (IRT) approach, which provides estimates for ability and item parameters of item difficulty, discrimination, and pseudo-guessing. A variety of strategies are then available for placing the estimates from different test forms on the same scale. "Using the common items and one of several methods, the coefficients of the linear transformation relating the item parameters for the two tests can be determined. With knowledge of the linear transformation, the item and ability parameter estimates may be placed on a common scale" (Hambleton, Swaminathan, & Rogers, 1991, p. 142). Because the full 3-parameter IRT models require significant sample sizes, Rasch modeling, which only estimates the ability and item difficulty parameters, is a more practical option for many test developers as it requires a smaller sample size.

Besides the large sample size necessary to use IRT methods for calibration, there are also some assumptions that may be impractical for some test developers. Specifically, IRT has a strong implicit assumption that the scale measures a unidimensional construct. "What is required for the unidimensionality assumption to be met adequately by a set of test data is the presence of a 'dominant' component or factor that influences test performance" (Hambleton, et. al, 1991, p. 9). Typically, unidimensionality is demonstrated through the results of a nonlinear factor analysis

supporting that a dominant first factor exists, as based on a scree plot of the eigenvalues. If unidimensionality is not supported, multidimensional models may be used for more complex tasks with the requirement that the latent structure of the test be identified and understood.

Differential item functioning techniques for alternate test forms.

While IRT is often used in the equating process, it can also be used to simply compare the statistical properties of items on separate test forms, provided there is some commonality between either the items or test takers. If a researcher has the same test data from two test forms for the same group of people, concurrent calibration estimation can be used to compare the item parameters. This method would place the item parameters on the same scale and allow comparisons among items. The utilization of IRT in this comparison process is similar conceptually to how it is used in differential item functioning (DIF) examining theoretically equivalent items on alternate test forms. This type of analysis is often performed in the examination of translated test forms and is discussed further below. However, because these methods are most often used in the study of separate groups, a brief discussion of traditional DIF methods is appropriate.

In traditional DIF, the researcher is interested in comparisons for the same item for two separate groups (i.e. males versus females). In order to make a comparison of the two groups, ability estimates are plotted against the probability of the test taker correctly responding to the item (Camilli & Shepard, 1994). The S-shape curve which typically results is called an item characteristic curve (ICC). The curve is constructed based on the IRT estimates of difficulty, discrimination, and pseudo-guessing, if the three parameter model is utilized. In DIF, the ICCs for two groups of interest are separately

constructed. If the item functions the same between the two groups, the ICC will be identical. If the item is more difficult for one group than the other, in other words having distinct difficulty indices, the curves will be separate but parallel. In the terminology of DIF, these types of items display uniform DIF in that there is a "relative advantage for one group over the entire ability range" (Camilli & Shepard, 1994, p. 59). If the curves cross, this suggests that the item discriminates between high and low ability individuals in a different manner for one of the two groups. Items whose ICCs cross display nonuniform DIF, in that the group that has the higher probability of correctly responding depends on the range of the ability estimate. Camilli and Shepard detail various methods for how the differences in the ICCs can be measured, including model comparison measures. While IRT models can be used in this process, other methods that require smaller sample sizes and have fewer assumptions can be utilized and are described further below.

When comparing items from test different forms, the methods employed combine concepts from both traditional DIF and equating procedures. In traditional DIF methods, two groups of respondents complete the same test form. The unit of investigation is the individual item, comparing how this item functions between two separate groups. When considering how DIF can be applied to the analysis of two forms, the unit of investigation shifts from the examinee to the item, although the ability level of the examinee must still be controlled for in order to eliminate potential confounding of the differences in ability with the differences in item functioning. Because of this potential confounding effect, having a single group or equivalent groups complete both forms of the test will control for the ability level. As in the equating process, concurrent calibration can be used to

estimate the item parameters. These can then be used to construct the ICCs and compare the items in a manner analogous to what is performed in traditional DIF.

While the above application of DIF techniques to analyses of test versions may seem logical, the literature does not provide any examples of this application with the exception of developing translated test forms. Zumbo (2003) notes that tests are often translated for various reasons including cost savings from creating a brand new test, a desire to utilize the psychometric and validity evidence from the original test, the fairness associated with decision making, and for the purposes of conducting psychological or sociological research. The literature regarding translated test forms can provide interesting and useful information for the development of any type of alternate test form. In translation DIF, the object of investigation is the item, in the original form and the form for the translated language. DIF techniques attempt to identify potentially problematic items that function differently between the original and the translated form.

Sireci and Swaminathan (1996) describe the unique issues involved with translating one form of a test to another language. When developing translations of a test, many times it is not possible to use a single group design, for the obvious reason that examinees are generally only proficient in one language. Usually equivalent groups are used instead, which opens the possibility of confounding group differences with ability differences. Some researchers (i.e. Sireci and Berberoglu, 2000) have suggested using bilingual individuals to take both forms of the test. However, this method also may have some problems as it makes the unlikely assumption that bilingual individuals are similar to the population of individuals who speak only one of the languages. Sireci and Swaminathan emphasize that this group difference must be accounted for in some

manner, either through the use of a common bilingual group or in careful selection of two monolingual groups who are matched to be as equivalent as possible. This matching needs to be carefully performed to minimize the potential confound. As the researchers state, "Without accounting for group differences, differences between original and translated items cannot be made" (p. 1).

Sireci and Swaminathan also addressed the importance of the issue of unidimensionality in the selection of a method of detecting DIF. As mentioned above, standard IRT models assume that the test measures the same, unidimensional construct. Sireci and Swaminathan discuss how this issue applies to translated test forms:

> First, if unidimensional DIF detection procedures are to be used, the dimensionality of the instrument in each language must be assessed. Second, it must be demonstrated that the same unidimensional construct is measured by both language versions of the test. Additionally, if the construct is determined to be multidimensional, the common dimensions across languages must be identified. Therefore, to evaluate translation DIF, the dimensionality of the construct measured, and the degree of equivalence of the construct across languages, must be understood (underscore in original quotation) (p. 2).

This discussion can be carried over to the development of test versions of any sort. The dimensionality of the instrument in both forms must be examined to determine if the unidimensionality assumption holds. In addition, the researcher should attempt to show that the same unidimensional construct exists in both test forms. The dimensionality of the test should be studied using nonlinear factor analysis, appropriate for dichotomous data, using software programs such as TestFact (Wood, Wilson, Gibbons, Schilling,

Muraki, & Bock, 1997) or NOHARM (Fraser, 1983). If the test is determined to be multidimensional, this structure must be understood and should be equivalent across the test forms.

Methods of differential item functioning.

Because of the large sample sizes necessary for IRT and because of the strong implicit dimensionality assumptions of IRT modeling, statistical equivalence between translated test forms are often explored using other methods, including logistic regression, the delta plot, the standardization method, and contingency table methods which are described in further detail here.

If IRT methods cannot be used, logistic regression procedures are often used as the next-best recourse, as they require a smaller sample size and unidimensionality, while still assumed, is less strong. In logistic regression DIF techniques, the probability of correctly responding to the item is predicted by the independent variables of ability (as measured by total test score), group, and the interaction. As Camilli and Shepard (1994) note, the formula used in logistic regression for traditional DIF with separate groups is as follows:

$$P(u_i = 1) = \frac{\exp \Psi_i}{1 + \exp \Psi_i} \tag{1}$$

$$\Psi_i = \delta + \tau_1 G_i + \tau_2 X_i + \tau_3 G_i X_i \tag{2}$$

where i is the index for examinee, u_i is the examinee's item response scored with a 1 (correct) or 0 (incorrect), X_i is the examinee's total score, and G_i is the group membership of the examinee (coded as 1 for the focal group and 2 for the reference group). In translation DIF, this latter variable is replaced with a variable indicating the test form. The variable τ is a combined log odds ratio, similar to a regression weight. A step-wise

logistic regression procedure is performed entering first the total score (X_i), then the group variable (G_i), and finally the interaction term (G_iX_i). At each step, model fit is tested to see if each term significantly improves the fit to the data beyond that of the previous model. If the best fitting model contains all three terms, then the item is said to display nonuniform DIF. If the best fitting model contains only the score and group variable, then the item is said to display uniform DIF. If neither the group variable nor the interaction term is significant, then the item is not said to display DIF.

The logistic regression method has been well established for the use of identifying traditional DIF between groups as well as with translation DIF when equivalent groups are used. The problem with using this method in other construct equivalence studies is that the method requires separate groups of individuals. With a single-group of examinees, the logistic regression cannot be utilized due to the issue of dependency, as the same examinees have scores on both test forms. A more appropriate statistical procedure would perhaps involve generalized estimating equations (GEE), a method of estimating regression parameters when the outcome measure consists of multiple correlated discrete data points (Agresti, 2002). These procedures have not yet been utilized in the context of examining differential item functioning but may show promise for future research.

In the comparison of items from alternate test forms, if the IRT models are not appropriate due to the sample size and/or the dimensionality assumption and logistic regression procedures cannot be used due to issues with data dependency, other methods of detecting DIF must be used. The history of DIF analyses offers a variety of alternative statistical and graphical procedures, while less powerful than IRT or logistic regression

also make fewer assumptions. The delta plot method and the standardization methods are described below.

In the delta plot method, delta values are calculated for each item (Angoff, 1982; Muñiz, Hambleton, & Xing, 2001; Sireci & Allalouf, 2003). These values are transformations of the difficulty indices (proportion correct) for a specific item, normalizing the value and rescaling it to have a mean of 13 and a standard deviation of 4.[2] The delta values are then plotted for each item pair with the delta values for an item from one form on the horizontal axis and the values for the item from the other form on the vertical axis. On the scatterplot, a principal axis line is drawn through the origin with two parallel lines drawn around it. The rule of thumb that Angoff established is that these parallel lines are drawn 1.5 units from the principal axis line. Item pairs which fall outside of the parallel lines are flagged as potentially problematic. Examining the delta plot is an easy method of identifying items that display possible DIF although this method does have some limitations. As noted by Camilli and Shepard, the delta values are confounded with item discrimination indices. Items with very high discrimination indices may be erroneously flagged; in addition, items with very low discrimination indices which do contain item differences may be missed. Although this method may have some limitations, in a simulation study, Muñiz, Hambleton, and Xing (2001) found that this method was effective at identifying items exhibiting large DIF across languages, even when sample sizes were very small. Sireci and Allalouf suggest using the delta plot method as a preliminary investigation of DIF prior to other statistical analyses.

[2] These values are often referred to as "ETS Delta Values" after the organization that originally developed the scale.

Another method used to identify DIF is the standardization approach, introduced by Dorans and Kulick (1986). The standardization approach is a method of combining the differences in proportion correct at each individual score level, weighting the average by the frequency of individuals who score at each level. The standardized proportion correct difference, or "signed proportion difference controlling for the observed total score *X*", as defined by Camilli and Shepard (1994) is obtained by the following:

$$\text{SPD-}X = \frac{\sum_{j=1}^{s} n_{Fj}\Delta p_j}{\sum_{j=1}^{s} n_{Fj}} \tag{3}$$

where the product of the frequency of examinees in the focal group (n_{Fj}) and the difference in proportion correct (Δp_j) between the reference and focal group at each score level (j) is divided by the frequency of examinees in the focal group. Doran and Kulick suggest that items with an SPD value greater than or equal to 0.10 be flagged as being potentially problematic. Doran and Kulick also developed an unsigned proportion difference (UPD-*X*) measure, or root mean square standardized difference which is obtained by the following calculation:

$$\text{UPD-}X = \sqrt{\frac{\sum_{j=1}^{s} n_{Fj}(\Delta p_j)^2}{\sum_{j=1}^{s} n_{Fj}}} \tag{4}$$

According to Camilli and Shepard (1994), this unsigned index may be an indicator of non-uniform DIF while the SPD-*X* may be an indicator of uniform DIF. However, while Camilli and Shepard note that the SPD is thought to be relatively stable and is highly recommended in the detection of DIF, the unsigned index is thought to be more subject to sampling errors and should be used cautiously. Sireci, Fitzgerald, and Xing (1998) adapted the use of the SPD-*X* for use in translated test forms, with their analysis of an information technology certification examination being translated from English to three other languages. While the index has only been used for more than one group, it can be easily adapted to a single-group design although one needs to specify which form would act as the "focal" form in the equation.

Another measure of the size of DIF is the Mantel-Haenszel log odds ratio (Camilli & Shepard, 1994). The ratio, used to combine the odds ratio across score levels, is calculated through the frequencies obtained in a 2x2 contingency table, as represented by Table 2.

Table 2: 2x2 contingency table for traditional DIF

		Score on studied item		
		1	0	Total
Group	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	n_{Fj}
	Total	m_{1j}	m_{0j}	T_j

The log odds ratio is then obtained through the following formula:

$$\alpha_{MH} = \frac{\sum_{j=1}^{s} A_j D_j / T_j}{\sum_{j=1}^{s} B_j C_j / T_j} \quad (5)$$

where *j* is the score level. In order to aid interpretation, the natural logarithm is used to transform the index:

$$\beta_{MH} = \log_e(\alpha_{MH}) \quad (6)$$

Camilli and Shepard recommend the use of this index with the SPD-X for understanding the size of DIF for an item. The interpretation of α_{MH} is quite simple. For example, say that the index equaled 2.35. This would mean that the odds of answering the question correctly are 2.35 higher for the reference group than for the focal group. Negative values would mean that the item favors the focal group. Again, while this method has been used primarily for multiple group designs, the index can be adapted for the use of a single-group design in which all examinees take two test forms. In Table 2, the cells for group would be replaced by frequencies of individuals who correctly respond for each form.

Camilli and Shepard detail several methods of using inferential tests to detect DIF between groups on a single item. Most popular is the Mantel-Haenszel chi-square statistic, originally introduced for the use of DIF by Holland and Thayer (1988). Allalouf, Hambleton, and Sireci (1999) demonstrated how the Mantel-Haenszel chi-square can be used in the detection of translation DIF in their analysis of the Israeli Psychometric Entrance Test, which was translated from Hebrew to Russian. The disadvantage of the chi-square test is that it also requires independence among observations, similar to the logistic regression procedures. Thus, this test cannot be used

for a single-group of individuals multiple forms of a test. Rather, tests designed for dependent data, such as the McNemar test need to be utilized (Agresti, 2002).

Regardless of the method used to identify DIF on items on a translated versus the original version of a test, following the statistical procedures, judgmental methods are performed in order to explain the statistical results. For example, Gierl & Khaliq (2001) used DIF techniques (the SIBTEST) to identify biased items and then asked experts to hypothesize potential sources. The panel of experts (consisting of 11 testing specialists) identified the following sources of DIF: 1) omissions or additions that affect meaning (adding words, phrases, or expressions), 2) differences in words or expressions inherent to language or culture, 3) differences in words or expressions not inherent to language or culture, and 4) format differences (i.e. punctuation, capitalization, item structure, complexity, length etc.). Similarly, Allalouf, Hambleton, and Sireci (1999) asked a committee of translators and researchers to examine items flagged by the Mantel-Haenszel method. They identified potential causes of DIF as changes in difficulty of words or sentences, changes in content, changes in format, and differences in cultural relevance.

Conclusion

In conclusion, the process of examining alternate test forms to determine equivalence is rooted in studies of equating and DIF. Statistical methods of determining construct equivalence have some advantages over more judgmental methods in that they provide a relatively quick method of analyzing the patterns of responses. However, many of the statistical procedures, such as IRT, require large sample sizes and make strong assumptions about the structure of the test data. For many test developers, it may not be

feasible to gather such large samples. In addition, if the data structure is not unidimensional or if a clear multidimensional structure cannot be understood, the test developer must use other methods, such as logistic regression or contingency table methods. Yet these methods also have limitations in that they are not appropriate if a single-group of examinees takes both test forms. Procedures to examine statistical equivalence for this circumstance are not well developed, with the exception of IRT methods, which again may not be feasible.

In addition, and perhaps most importantly, the establishment of statistical equivalence does not ensure construct equivalence. Too often, the test specifications are relied upon in the conclusion that two test forms are measuring the same construct. Statistical procedures may provide an indication that the test forms are measuring something different, but do not provide information on what the problem may be nor on what types of revisions may result in improved items, which is why these methods tend to be paired with judgmental methods relying on experts.

Chapter 3

CONTEXT OF STUDY

As discussed in Chapter 1, the context through which the research questions will be explored is in the development of two forms of a test intended to measure legal case reading and reasoning among law students. This section of the literature review describes the construct of legal case reading and reasoning in relation to the traditional law school curriculum. In addition, the process of developing the two test forms is described.

Law School Curriculum

Entering law school must seem incredibly daunting to the new first-year student. The classroom environment and assessment system are dramatically different from what most students encountered during their undergraduate years. In the traditional law school curriculum, students are confronted with the professor's use of the Socratic method, in which individuals are called upon at will and grilled about details of cases. The assessment system in law school courses typically consists of only one exam at the end of the semester -- an experience not very common in the undergraduate curriculum. The introduction to this new schooling environment has been documented to be a very stressful time for students, often leading to psychological stress, substance abuse, and even suicide in some instances (Hess, 2002).

The traditional system of legal education utilizing solely the Socratic method and competitive one-shot assessment system has come under attack in recent years. In fact, a very recent publication by the Carnegie Foundation for the Advancement in Teaching (Sullivan, Colby, Wegner, Bond, and Shulman, 2007) stated that the traditional system of

legal education needs a vast reformation. The authors of this publication suggest that a newly revised curriculum needs to have more instances of simulations, real-world experience working with clients, and opportunities for formative assessment so students can receive feedback on how they are progressing.

One criticism of the traditional law curriculum is the overemphasis on summative assessment. As mentioned above, students' success is based on their performance on one examination in each course at the end of the semester. No formative assessment, other than the "drill and grill" method employed in the classroom, exists for students to receive feedback about how they are progressing in their courses. Gross (1972-1973), argues that the traditional law curriculum operates with a social-Darwinist view of the student in which failure occurs due to the students' personal deficiencies. Gross states, "Law school programs generally are designed to meet the needs of a supposed norm of student who requires minimal guidance in order to teach himself" (p. 262). Few opportunities exist for students to receive feedback, guidance, or support in their studies. Gross expands upon this perspective stating:

> One barrier to adequate law school instruction in these reading-writing-reasoning fundamentals is the 'social Darwinist' assumption that 'any law student worth his salt will pick these things up on his own.' Too often, however, even the better students do not 'pick up' adequate basic analytic skills in law school (p. 267).

Following this social-Darwinist assumption, many law professors assume that first-year law students begin their studies already possessing the reading and comprehension strategies necessary in order to succeed. According to Stratman (1990), many professors assume *skill-deployment*, which he describes as the following:

> ...[S]tudents who learn to think critically in their undergraduate years need merely deploy their ordinary critical thinking skills to legal reasoning and argument tasks, because the information processing required to perform these tasks is no different from, say, that required for debating a policy decision in an undergraduate world history class (p. 160).

This assumption fails to take into account the new challenges that novices face regarding the content of the law and the challenges with learning to read legal cases. Reading legal cases can be a very challenging task, as the format, structure, and vocabulary of these assignments are unfamiliar and difficult to understand for a novice student. Students are thrust into case reading immediately during their first semester, often times with no introduction or any guidance. As Lundeberg (1987) states, "Not only are legal texts largely incomprehensible to beginners, but students are rarely given instruction in case reading. Thus, case reading is often fraught with distress" (p. 409). While the Carnegie Foundation did not address the issue of case reading specifically, other legal education theorists have argued that law students are lacking in the skills to closely read and interpret legal cases. However, this skill is extremely important in the career of a lawyer. As Deegan (1995) notes, "Words are tools for lawyers, who must be able to forge words into consequential discourse. Learning to be a lawyer entails more than thinking like a lawyer; it necessitates being able to read and write like a lawyer" (p. 157).

The Discourse of Law

When training to be a lawyer, students need to become ensconced in the discourse of the law. As Gee (1991) states, "A Discourse is a sort of 'identity kit' which comes complete with the appropriate costume and instructions of how to act, talk, and often

write, so as to take on a particular social role that others will recognize" (p. 142). The discourse of law consists of the ability to think, read, and write like a lawyer. Part of the students' acquisition of the legal discourse includes the ability to read and understand cases. In order to better understand the challenges of learning how to read and reason through cases, one should understand the parts of cases and the basics of legal case reasoning.

Stratman (2002) provides a description of the basics of understanding legal cases. When presented with a new case, the court is presented with a question, or the *issue,* for which it needs to needs to make a decision based on legal rule. These legal rules are not clearly defined in many cases. Rather the court needs to infer legal rules from other precedent cases. Stratman describes this process below:

> Put simply, cases record and articulate decisions by courts in response to issues presented by opposing attorneys. As such, cases are also commonly referred to as *opinions.* In a case, a court interprets one or more legal rules by drawing on the facts and reasoning that other courts have used in previously decided cases concerned with the interpreting the same rule (p. 62, italics in original quotation).

In comparing the presented issue to precedent cases, the court needs to be aware of which cases are controlling. In other words, precedent cases from higher courts are binding or controlling for decisions in lower courts. Stratman continues to define this concept of controlling precedent cases:

> This determination as to which cases are authoritative and relevant precedents may look simple on the surface. However, it can become very complex. Even when the relative authority of given precedents are settled, courts must

> nevertheless take great pains to explain their understanding of the possible relevance of these precedents, that is, to justify their use or exclusion in reaching a decision. (p. 63).

Using deductive logic and analogic reasoning with these precedent cases, the court then presents the *holding*, the statement that interprets a legal rule and applies it to the situation presented in the case. Holdings not only provide a decision of the presented case, but also modify the law itself based on the specific details of the factual situation in the case.

According to Stratman (1990), each case consists of several canonical components. In other words, each case follows a certain pattern or "story line" in that certain pieces of information are expected, although unfamiliar to the novice. As he states, "Clearly, legal cases do confront novices with a host of unfamiliar text conventions and cues, at virtually every discourse level: legal terms, the larger canonical placement of information such as issues, facts, and rules, [etc.]…" (p. 174). Expert case readers are familiar with these expected or canonical parts of a case. These parts include the facts, issue, holding and reasoning, and dicta. The facts refer to the specific factual situation presented in the case. The issue is subtly different from the facts and rather is a statement of the question the court sees itself as addressing. The holding, described above, refers to the conclusion of the court and generally contains a synthesis of the court's reasoning. Dicta refers to "asides, commentary that the court may offer about the rule, its history, or possible application to the other like (or unlike) cases" (Stratman, 1990, p. 168).

As Stratman, 2002 notes, within each of the canonical parts of a case, students may encounter problems with understanding. For example, information may be missing from the facts which could be potentially helpful or harmful for a client's case. In the issue, the court may present its issue in a manner that is different from the manner presented by the litigants. Understanding the holdings of cases may be very difficult as they "can frequently be complex, containing exceptions, caveats, or indeterminate language that later legal readers (whether judges, lawyers, or law students) must reckon with" (Stratman, 2002, p. 63-64). In addition, students often have difficulty distinguishing the holding from the dicta (Stratman, 1990). The problems that students may face in interpreting the canonical parts of the case may occur when reading a single-case and also in the relationships among cases.

Research in Legal Case Reading and Reasoning

Several researchers have investigated the difficulties that students face in legal case reading and reasoning. For example, Fajans and Falk (1992-1993) discuss their experiences teaching legal writing courses in which they perceived students' lack of close reading ability. They argue that students are often not able to "read between the lines and to link texts to larger contexts" (p. 163). They continue, "Helping law students to get beyond purely denotative, case-briefing notions of reading is, however, no easy thing. In an age of reading comprehension tests, students are trained to read only for facts, for information" (p. 164). While they did not conduct an empirical study examining how training can improve legal case reading, Fajans and Falk felt that by pushing students to read closely, and to "read beyond what is explicit on the page" (p. 169) their students were better able to communicate through their writing and make certain arguments. "We

asked them to read the case for what is implicit there – literary style and jurisprudential or interpretive posture – and for what is not there at all – legal and historical context and omissions of fact or lapses in logic" (p. 169). Although their critique is interesting, Fajans and Falk performed no formal assessment or data analysis to determine how the training impacted students' abilities to read and reason through the cases. They conclude that instructional techniques are necessary in order to guide students towards this process of close reading. Their argument is compelling:

> We must develop techniques that help [law students] move past 'stage-one' legal reading (reading for what the court says it is saying) through 'stage-two' reading, that is, beyond unquestioning acceptance of textual authority and 'found' meaning to an open-ended process of unselfconscious response and self-aware reflection. Then students can move to 'stage-three' where purpose informs a 'final' reading, where readers take control over 'two or more opposite or antithetical ideas, images, or concepts simultaneously' activated by the text and thereby synthesize its proliferant meanings as fully as they can (p. 190).

This quote from Fajans and Falk suggests that research is necessary to determine what instructional interventions or techniques can be employed to help students to move beyond reading for the written word to a more active interaction with the text.

Lundeberg (1987) explored strategies that novice law students apply when confronted with legal texts. In her two-party study, Lundeberg first preformed a descriptive analysis examining the differences in the legal case reading strategies used by novices versus experts. The experts, primarily law professionals and lawyers, were found to engage in substantially different strategies than individuals who were less familiar with

legal case reading and the domain of law. Using think-aloud strategies, Lundeberg found that experts applied strategies such as using the context of the case (i.e. case headings, type of court, dates, judge names), overviewing (looking at the length, decision, marking the action, summarizing the facts), rereading analytically (looking for terms, facts, and the rule of the case), synthesizing (looking for cohesion among the information and considering hypotheticals), and evaluating the decision (approving or disapproving of the decision and showing a sophisticated view of jurisprudence). In some cases, novice readers displayed some of the same strategies albeit at a much lower frequency. In addition, novices were more likely to express confusion about legal terms, express confusion about English words having legal meanings, contextually define words, add incorrect information, and assign names to the plaintiff and the defendant.

Do strategies that students employ in reading legal cases make a difference in how they perform in their law school courses? Deegan (1995) added to the understanding of novices' strategy use in reading difficult legal texts. Her investigation centered on understanding the relationship between strategy use and performance as measured by course grades. Using methodologies similar to Lundeberg's study, Deegan collected think-aloud protocols from first-year students as they read a law review article. Deegan decided against having the students read an actual legal case in the study as she theorized that the unfamiliar format of cases might influence the strategies that students used, resulting in the use of more "algorithms rather than heuristics" (p. 158) The study uncovered a relationship between first-year grades and the strategies that students use. In particular, Deegan found that the use of *problematizing* or using purposeful strategies in which students either posed or solved a problem that the text created, was positively

related to higher performance in law school grades. In contrast, students who used *default* reading strategies, such as linearly progressing through the text, tended to be lower performing in their law courses. This study was replicated by Christensen (2006) who utilized legal cases and found that similar reading strategies emerged.

Both Deegan and Lundeberg's studies suggest that novice readers in the legal domain have difficulty getting beyond "Stage 1" reading as defined by Fajans and Falk. The problematizing strategy seems to be highly similar to the second and third stages of reading that Fajans and Falk discuss in which students made moves such as "'voicing confusion,' 'questioning,' 'drawing a tentative conclusion,' or 'noting an anomaly'…'hypothesizing,' 'predicting', and planning.'"(Deegan, 1995, p. 160). Similarly, Lundeberg's study suggests that expert law professionals are also utilizing strategies which require reading beyond the written word. Granted, one would expect experts with years of experience in reading and analyzing case reading to demonstrate marked differences in case reading strategies and skills. However, the question emerges as to whether novices can be taught to utilize these strategies and achieve a closer reading of the text.

In the second part of her 1987 study, Lundeberg addresses this question. Following the strategies that emerged in the descriptive study, the CORE (context, overview, rereading, and evaluating) method was introduced to students at varying stages in law school. Using a test she developed, Lundeberg compared test scores on the effectiveness of explicit training of the CORE model, written guidelines only, and no training or guidelines for students with no law school experience, two weeks of law school, two months of law school, and 1-2 years of law school. Lundeberg found that

particularly in the early stages of law school, both the explicit training and the written guidelines proved to be beneficial on law students' ability to comprehend legal cases.

Lundeberg's study is optimistic that, with sufficient guidance, first-year law students can improve their case reading skills. However, her study does have some limitations. The test that Lundeburg administered consisted of three parts: 1) a section on distinguishing relevant versus irrelevant information, 2) a multiple-choice section on identifying the issue, facts, and purpose of dicta, distinguishing similar from dissimilar situations, and using hypotheticals, and 3) a short-answer section on who won the case, the prior action of the case, the issue, the rule, and the rationale. Most of the questions on the test are those which students would be able to glean directly from the text, without requiring them to go beyond the written word and question. In addition, students were only tested on their ability to read and understand a single case, whereas the successful lawyer will need to be able to synthesize information across related cases.

Stratman (2002) provides some insight into how to encourage students to use strategies that will aid in understanding and reasoning through sets of related legal cases. Stratman hypothesized that the manner in which cases are used in most law school courses do not require students to go beyond the written word of legal cases. In order to test this hypothesis, Stratman asked a group of 56 students to think-aloud while reading a set of related legal cases. Each student was randomly assigned a different role to assume while reading the cases. Some students were asked to assume a class recitation role in which they were told to be prepared to explain the "significance" of the cases as they would for one of their courses. Other students were assigned more active task purposes, such as considering themselves to be a lawyer or an advocate for a particular person in

the case or that they would need to write a memo concerning the policy of the legal issue discussed in the cases. The think-aloud protocols were analyzed to determine the number of "textual and legal interpretation problems" (p. 57) students were able to identify within the cases. Students presented with the task of being an advocate for the client or writing a memo concerning policy were better able to detect problems in the text than students in the class recitation groups. In addition, students with the advocatory assignment were better able to identify problems that existed at the cross-case level than students in the policy and class recitation groups. Stratman concludes that the context that students undertake when reading a text can affect the ability to identify problems in legal cases.

The implications of Stratman's findings to the law curriculum are similar to those suggested by the Carnegie Foundation. Students need to have more hypothetical and real-world experiences, even in the seemingly routine task of reading legal cases. In addition, the construct of legal case reading and reasoning for inexperienced law students needs additional attention and research. The texts by Fajans and Falk, Deegan, and Lundeburg all support that novice law students are unable to tackle the tasks of reading complex legal texts in the sophisticated manner required to be a competent lawyer. While skill acquisition may potentially be developmental, requiring years of experience and practice, Lundeburg's study supports that with training and guidance, students may be able to improve these skills. This need for guidance and training in the area of legal case reading and reasoning also corresponds with the suggestions by the Carnegie Foundation to increase formative assessment in the law school curriculum.

Development of Test of Legal Case Reading and Reasoning

In order to determine whether interventions in the area of legal case reading and reasoning are successful, valid and reliable instruments are necessary. Such an instrument would need to measure students' abilities to navigate through a set of related legal cases, asking students to identify not only the information explicit in the text but also requiring students to read beyond the written word to areas of ambiguity that competent lawyers would be able to use to their advantage. With funding from the Law School Admissions Council (LSAC), Stratman, Evensen, and Oates (2005) initiated a project to develop such an instrument.

The first version of the test (TV1) was developed in 2004-05. The test consists of 14 multiple-choice items based on a series of three interrelated cases. Two of the cases were held at the appellate level and one at the trial court level in Pennsylvania. Each case concerns a Commonwealth of Pennsylvania statute related to court costs for arbitration hearings. The context of the cases is a topic that students would not likely have encountered during their law school courses. Stratman, whose 2002 article comprehensively reviews each of these three cases for use in a think-aloud study, describes the process involved with arbitration:

> A homeowner may enter into arbitration with a flooring company whom she feels overcharged her for a new kitchen floor in her home. The arbitrator listens to each party's attorney, then issues a judgment. If they wish, losing parties are required by a 1836 rule to first pay any record costs they owe to opposing counsel within 20 days of the decision. In each of the three cases used in this study, the losing parties failed to meet this prerequisite either in part or whole…(p. 69).

Upon reading the cases, students are asked to assume the advocatory role of their client, Mr. Mackey who lost an arbitration hearing in the hypothetical case of *Mackey Plumbing Company v. Pepper* (2002). Two supporting precedent cases are included in the task: *Meta v. Yellow Cab Company of Philadelphia* (1972) and *Black and Brown Inc. v. Home for the Accepted, Inc.* (1975). These cases are real-life cases, although they have been edited for the task. The *Mackey* case is fictional and was developed specifically for purposes of the task. As Stratman notes, "The *Mackey* case was deliberately constructed to contain incomplete and somewhat incoherent reasoning from precedent, presenting a subtly skewed application of the earlier *Meta* and *Black and Brown* cases to its facts" (p. 69).

In the *Mackey* case, the client, Mr. Mackey, was not allowed to proceed with an appeal from the original arbitration because he failed to pay the total portion of the record court costs. He had relied on communication from the plaintiff, juxtaposed two numbers, and sent in less than what was due to the court two days past the deadline. In the earlier precedent case of *Meta* (1972), the defendant (Yellow Cab Company) also lost an arbitration decision and paid an insufficient amount of the record costs before the deadline and was thus denied an appeal. The total amount paid by the defendant was short by only $7.75. The appeals court in this case ruled in the favor of the defendant, stating that the shortfall was a trifle matter or *de minimis*. "They base their position on arguments that the statute is hypertechnical, puts form before justice, and that it serves no real purpose" (Stratman, 2002, p. 70). In the *Black and Brown* case, the defendant (Home for the Accepted, Inc) also lost an arbitration hearing and was denied an appeal. However, unlike the *Meta* case, the defendant made no effort to pay the cost within the

20-day time limit. The *Black and Brown* court overturned the *Meta* case, stating that the statute did serve a purpose and denied the appeal for the defendant as Home made no "honest" attempt to make the payment. While the court does acknowledge that a minimal shortfall should not result in the harsh punishment of denial of an appeal, a party seeking an appeal still needs to put forward an effort to pay the costs. The holding of the court states that "a valid attempt to make…timely and full payment, coupled with substantial though incomplete compliance with the requirement should not result in the harsh finality of an order quashing an appeal from arbitration." The *Mackey* court relies on the holding of the *Black and Brown* case in its decision against Mr. Mackey. However, the court never specifies why the actions of Mr. Mackey do not constitute "substantial compliance" or a "valid attempt."

Stratman (2002) details possible textual and legal problems that students might identify upon reading the individual cases and when considering the relationship among the cases. Based on these problems and additional issues brought up in student think-aloud protocols and expert analysis, questions were constructed using two dimensions:

> First, students need to be able to construct representations of individual cases that accurately represent the discursive structure and content of those cases, and as well, be able to construct representations of multiple related cases and the different discursive relationships between them (Stratman, Evensen, & Oates, 2005, p.21).

Therefore, each item is first classified on whether it addresses one case (single-case) or if it addresses two or all three of the cases (cross-case).

The second dimension by which items are categorized concerns, "…the difference between, on the one hand, the ability to accurately represent the discursive structure and content of cases and, on the other hand, the ability to recognize indeterminacies (uncertainties) of interpretation…" (p. 22). Therefore, each item is also classified as being determinate or as being indeterminate in nature. The answers to the determinate items are explicit within the cases and are designed to measure students' ability to closely read the text. The indeterminate items go beyond the explicit information from the cases. They require students to go beyond the written word to identify potential ambiguities, lapses of logic, or unaddressed issues that might useful for the specified task of being an advocate for the client. The structure of these items often focuses on the students' ability to frame appropriate questions about the cases at hand. As such, the distractors actually appear in question format. As Stratman, Evensen and Oates state,

> …[T]he stems for these indeterminate items do not identify a specific indeterminacy in or among the three test cases students have read. Instead, the possible choices present different potential indeterminacies of interpretation, framed as questions, for students' consideration…[I]ndeterminate questions test whether students can recognize a specific indeterminacy of interpretation and then appropriately judge its relevance to their task situation (p. 25).

Based on these two dimensions, each item on the test is classified into one of four possible types: single-case determinate, cross-case determinate, single-case indeterminate, and cross-case indeterminate. Single-case determinate items focus on the ability to identify or summarize concepts such as the reasoning or legal issue within only one case. Cross-case determinate items require students to compare and contrast issues

and reasoning across at least two cases. For example, a question might ask students to compare the factual situations of the defendants in two cases. Another question might ask students to identify a statement which best summarizes the legal issue across all three cases. Single-case indeterminate items require students to identify and evaluate ambiguities that exist only within one case. For example, students might be asked to identify which of the listed questions describes an "unknown" which might be beneficial for the task of appealing the client's case. Cross-case indeterminate items are similar in that they ask students to identify useful ambiguities but at the cross-case level. For example, students might be asked to identify questions that opposing counsel might focus on in light of given precedent (from the two supporting cases) if the client's case is appealed.

Using this classification strategy, an initial subset of 20 items with justifications was created by the test developers for the set of cases described above. Following an internal review and revision of items, a group of 8 law school faculty and practicing appellate attorneys reviewed each item and justification independently. Three focus groups with the experts were held to discuss evaluations, leading to additional revisions of items and justifications. A total of 14 items remained after the review process (examples of which are shown in Appendix A), including 4 single-case, determinate items, 4 cross-case, determinate items, 3 single-case, determinate items, and 3 cross-case, indeterminate items. The test form was piloted in a sample of 161 first-year students from five laws schools, each of whom were paid $50 each. Approximately half (81) of the students took the test in their first semester. The other half (80) were administered

the test in their second semester. Students were allowed to use a legal dictionary, take notes at will, and to refer to the three cases as they completed the test questions.

The preliminary data collected from the TV1 form supported initial hypotheses about the test and the construct of legal case reading and reasoning. Items became more difficult as they moved from single-case to cross-case. In addition, as items moved from determinate to indeterminate, they became much more difficult. As expected, the correlation of the total test score with course grades was not high (0.14). The correlation was slightly higher for legal writing grades, but was also not very high. The average total score was 7.91 or 56.5% (standard deviation of 2.08). Students who took the test in their second semester did no better than students who took the test in their first semester.

Based on these interesting pilot studies on TV1, Evensen and Stratman wanted to continue the exploration into the construct of legal case reading and reasoning. In 2005, they submitted a second grant proposal to the LSAC with the intention of exploring whether the skills had a tendency to increase by the third year of law school. In response to suggestions from the grant reviewers, the proposed project was expanded to include the construction of an additional form of the test based on a set of different legal cases.

The cases selected for this second test version, called TV2, also concerned topics that students were unlikely to encounter during their courses. The three legal cases included in the test were *Bridell v. Ribier Clothiers, Inc.* (1995), *Alexander v. Primerica Holdings, Inc.* (1991), *Hamilton v. Air Jamaica, Ltd.* (1991). All three cases are real although modified slightly regarding length and content. The students are asked to assume the role of being an advocate for Ms. Joan Bridell. The *Bridell* case occurs at the district-level in the state of Texas. The *Alexander* and *Hamilton* cases are both held in

the higher appellate court for the Third Circuit District. All three cases concern a company's use of a reservation of rights clause in benefit plan documents and whether such usage is allowable under the Employee Retirement Income Security Act (ERISA). ERISA allows companies to protect their right to change employee benefit plans provided that this reservation of rights is disclosed to employees and such reservations are clear to the average plan participant. Joan Bridell, an AIDS patient, sued her former employer, Ribier Clothiers, after the health care provision for AIDS patients was drastically reduced from a lifetime benefit of $1,000,000 to just $5,000. Bridell loses the case at the district level as the company had a provision in the benefit plan documentation reserving the right to amend benefits at any time and there was no evidence that the company had discriminated against Bridell as an individual. The supporting cases also focus on this issue of the reservation of rights clauses and their legality under the ERISA statute.

In *Hamilton,* the plaintiff had previously won a case at the district level against his former employer concerning the amount of severance payment he was entitled to based on the company's summary plan document (SPD) describing the employee benefit program. The written SPD contained an error which the company had amended in a written notice to employees. Employees were notified of the change the day before Hamilton's employment was terminated. The district level court had found for the plaintiff, stating that Air Jamaica was bound by the policy written in the SPD and that that the written amendment one day before Hamilton's termination "would frustrate ERISA's policy of protecting employees' legitimate expectations." After Hamilton won his case at the district level, the company appealed to the circuit court. The circuit court reversed the decision of the district court, stating that the summary plan document

contained a reservation of rights clause that allowed the company to change the benefit plan without notice to the employee. Because the company had reserved the right to change the benefit plans without notice, the court reasoned that "employees are on notice that they have no guaranteed benefits."

The *Alexander v. Primerica Holdings, Inc.* case was also held at the circuit level after a decision in the district court found for Primerica. The case concerns a 10-fold increase in premium payments for the company retirees' medical plan. Primerica argued that they had reserved the right to change employee welfare benefit plans without notice. The district court case had found for the company, stating that the employees were notified that the benefits could change at any time and that lifetime benefits were not promised. The retirees appealed this decision to the circuit court, who remanded the case back to the district court as the "finder of facts." The circuit court reasoned that the reservation of rights clause could not be unambiguously construed. As they note, "Because the summary plan descriptions do not clearly reserve the right to reduce benefits, we will reverse the district court's grant of summary judgment dismissing the complaint."

The questions for this test form once again focus on the legal and textual problems that the students, assuming an advocacy role for Bridell, should be able to identify given the precedent court cases. Following the same specifications for the TV1 form, a set of questions based on these three cases was developed. Because of differences in the nature of the cases, creating a completely isomorphic test form was not feasible. In other words, while each test question was written along the dimensions

described above, differences exist between theoretically equivalent items due to the differences in the nature of the cases.

The expert review of the initial version of the form proceeded in a slightly different manner than during the development of TV1. A total of 9 expert reviewers were asked to take the test "as is" without being presented with justifications or keyed responses. The reviewers then completed phone interviews with the test developers who used their feedback to revise the test form. This revised test form then was presented to a panel of three experts who again provided feedback and suggestions for the test items. Following the expert review, a total of 16 questions remained on the test (examples available in Appendix B). This test form has been undergoing pilot testing during the past year, the data for which will be analyzed in this study.

When test developers create alternate forms of tests, a logical question concerns whether the two test forms are measuring the same thing. Even with careful adherence to test specifications, subtle differences in the test forms may emerge that may not be evident in expert reviews or through statistical analyses. Therefore, the purpose of this study is to explore alternate methods of gathering evidence to support the construct equivalence of test forms. Specifically, the two forms of the test of legal case reading and reasoning will be examined to determine if they can be considered construct equivalent.

Chapter 4:

METHODS

Three methods for comparing the equivalence of item pairs from the two forms of an instrument measuring legal case reading and reasoning were compared in the study. These three methods include a statistical technique, expert review, and collection of think-aloud protocols.

Instrument

As described more fully in Chapter 3, the instrument to be examined in the project is a test of critical reading and reasoning of legal cases, designed to be administered to students in law school. Two forms of the instrument, termed Test Version 1 (TV1) and Test Version 2 (TV2) were created following test specifications aligned with a theoretical framework of the structure of legal cases. Each test version consists of a set of three related cases and 14 multiple-choice items categorized along two dimensions. For the first dimension, each item asks questions about either one case (single-case) or multiple cases (cross-case). On the second dimension, items can be categorized as either referring to canonical information (determinate) or implicit information such as ambiguities, silences, or contradictions (indeterminate) in the case(s). Table 3 displays the frequency of items that fall within each category specified on the test. Within each test version, students are presented with a purpose of reading the cases; the student is asked to assume an advocacy role for a client who had previously lost a case at the trial level.

The development of the two test versions followed the same test specifications based on the case and determinacy dimensions. For this study, the object of investigation is the item pair, consisting of an item from each version of the test classified along the

same dimensions and theoretically measuring the same intended construct. The item pairs were identified based on their dimensions and the intentions of the item writers. A total of 14 item pairs were analyzed.[3] Appendix C lists the pairings for the items.

Table 3: Distribution of items by classification on each form

	Determinate Items	Indeterminate Items
Single Case Items	4	3
Cross Case Items	4	3

Data Collection and Analysis

Collection of Verbal Protocols: The methods of collecting the think-aloud data followed recommendations and practices by Ericcson and Simon (1999), Pressley and Afflerbach (1995) and Deegan (1995). A pilot test was conducted with one student in order to fine-tune the data collection techniques, to gather information about the appropriate number of items to administer, and to determine a baseline amount of time for task completion. After the pilot test, plans were made to collect the verbal protocols from a larger sample of students. All first-year law students from a large regional mid-Atlantic university were sent an e-mail asking them to participate in a study of legal case reading and reasoning. A compensation of $60 was offered to the students for their time and effort. A sample of 30 out of approximately 240 students responded to the recruitment e-mail and participated in the study.

Demographic data collected from the students included gender, LSAT scores, first semester law school GPA, and cumulative first year GPA. Of the 30 participants, 21

[3] The test developers had decided to remove two items from originally developed TV2 instrument for reasons relating to certain features of the items. Also, the items did not have a theoretically equivalent item in the TV1 version.

were male and 9 were female. The average LSAT scores equaled 159.83 (standard deviation = 3.94). The average fall semester GPA equaled 3.18 (standard deviation = 0.51) and the average cumulative GPA equaled 3.16 (standard deviation = 0.40).

Because the test is quite lengthy, students completed the verbal protocol for only a select number of items in order to reduce fatigue. In a pilot study, Stratman, Evensen, and Oates (2005) had asked a small sample of students to think aloud on the complete TV1 form. A very pronounced fatigue effect was evident as the think-aloud process took over three hours for students to complete. The data that emerged from this pilot study was not very useful due to the length of time required for students to complete the task. Therefore, the two main test versions were each divided into subforms of 8 items each. Because the TV1 form only consisted of 14 items, the form was split in half with 7 items each appearing on the two subforms. In addition, 2 items were replicated on each subform so that all students completed a total of 8 items. The TV2 form was split in half with each student receiving 8 items. Each subform contained a similar distribution of each item type. Students were randomly assigned which test subform they received. A total of 14 students completed the think-aloud on the TV1 test forms. A total of 16 students completed the think-aloud on the TV2 test forms.

Students were told that the purpose of the procedure was to collect information on how law students read and reason through legal cases. After introducing students to the study and obtaining their informed consent, students were given a practice passage and corresponding multiple-choice item from a practice version of the logical reasoning section of the Law School Admission Test. Each student was asked to read the short reading passage silently and then was asked to read aloud the multiple-choice item

voicing any thoughts and comments that came to mind. The proctor interrupted the student during this initial practice if he or she was not verbalizing anything beyond the printed text in order to give a general idea of how the student should proceed when presented with the actual task.

Following the practice, students were handed the task directions and the three cases associated with their assigned test form which they were asked to read silently. Students were provided with a legal dictionary if they encountered unfamiliar terminology. When finished reading, the students notified the proctor who then administered the 8-item test to which they were once again asked to read aloud each item, verbalizing what they were thinking as they selected what they perceived to be the best response. In order not to influence or bias responses, very few prompts were provided. Rather, only basic prompts such as "Keep talking" or "What are you thinking now?" were used if the student did not verbalize his or her thoughts after a 30-second period of silence. During both the silent reading and think-aloud phase, the proctor sat diagonally behind the student so that he or she would not be in the line of sight of the participant and would be less of a distraction. Following the verbal protocols, a short interview of the student was conducted in order to collect some retrospective information about the strategies used in answering the multiple-choice questions and elicit general ideas about case reading and reasoning. The practice think-aloud, the actual think-aloud task, and the interviews were audio-recorded and later transcribed.

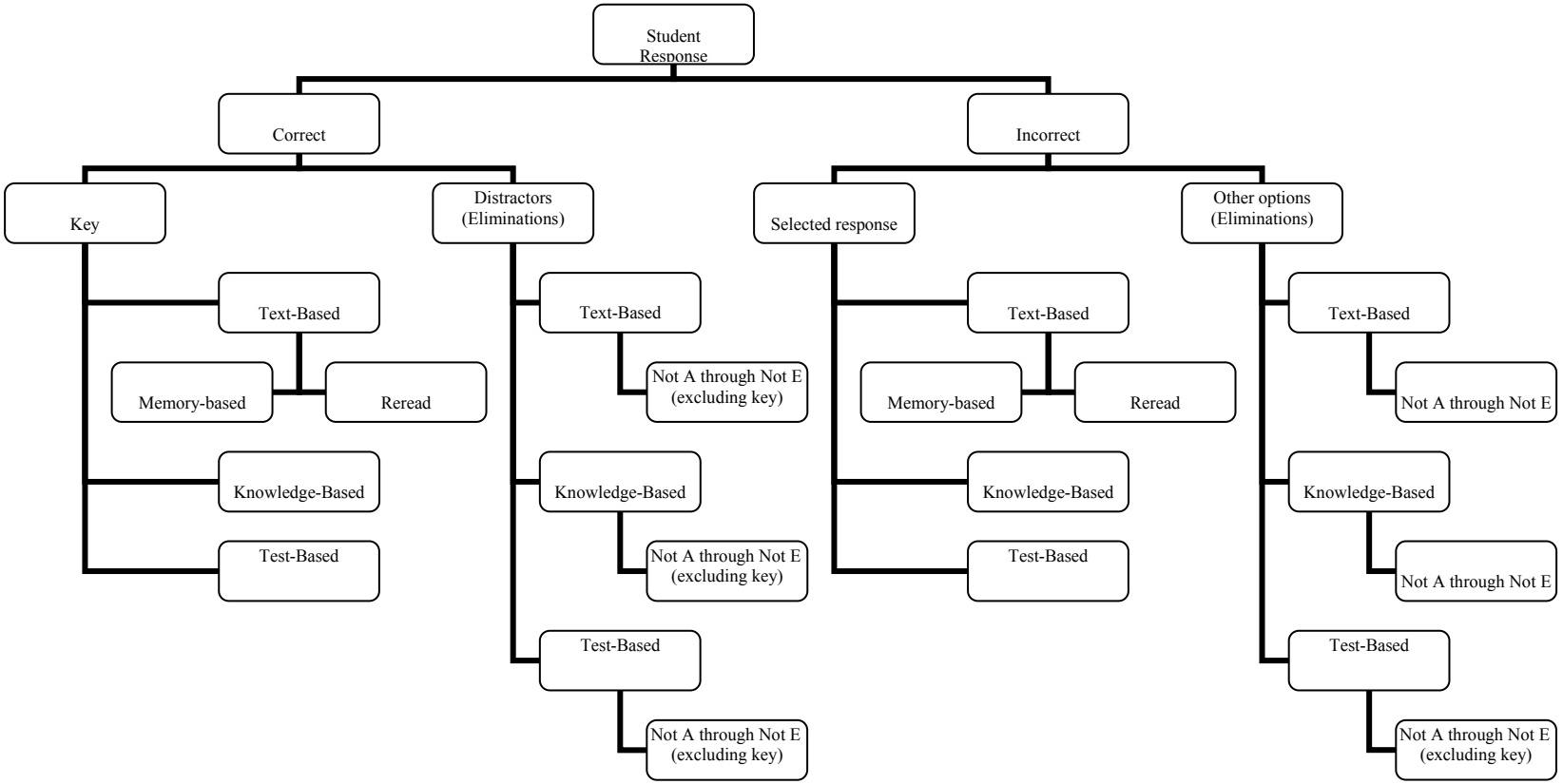
Analysis of Verbal Protocols: The data analysis occurred in three phases: *open coding*, *discourse analysis*, and *paired comparisons*. In Phase 1 of the data analysis, open coding was conducted on the item-level transcripts. Because the item is the unit of

measurement, all student transcripts for an individual item were coded together. One individual item was analyzed initially to develop a general coding scheme. This coding scheme was based on the strategies that participants employed when responding to the item. Analysis of the protocols was facilitated by the use of the N6 software program (QSR International, 2002). Figure 1 on the following page provides a graphic display of the nodes used during the initial open coding.

The coding strategy developed was based on both the analysis of an individual item and an inductive theory of the strategies that students would use when responding to the items. The researchers theorized that students would primarily depend on references to the case text, either by directly referring to the case or depending on their memory of the case. Some basic knowledge of the law might be required, such as definitions of "remand," and the concept of the appellate system, in order to correctly respond to the items. However, this requisite knowledge would be very basic, primarily including concepts that students learn during their first few weeks of their first semester in law school. No student should be able to correctly respond to an item based on prior knowledge. Rather, students should rely on references to the text as well as reasoning skills to correctly respond to the items.

In addition, the researchers were concerned with whether construct-irrelevant influences would lead students to the correct response without possessing the requisite reasoning skills. The coding strategy needed to be concerned with whether the students could incorrectly respond to an item by utilizing logical reasoning strategies, perhaps revealing flaws within the item key and distractors. Because of the nature of the multiple-choice test, it was expected that students would employ some test-taking

Figure 1: Nodes utilized in open coding of item-level transcripts



strategies to respond to items. However, of particular concern were potential instances of a student correctly responding to an item utilizing solely test-based strategies

All protocols for an individual item were coded as a group, since the item was the unit of analysis in this study. First, the response to the item was checked to determine if the student correctly or incorrectly answered the item. Each set of students' verbalizations corresponding to each option was analyzed according to a basic coding strategy of whether the student utilized text-based, knowledge-based, or test-based strategies to either accept or eliminate the option.

Text-based strategies included *memory-based*, in which the student relied primarily on their memory of the cases to make their decision, and *rereading*, in which the student directly referred back to a case and either quoted or summarized the text. *Knowledge-based* strategies referred to any strategy in which the individual referenced external information (i.e. definitions of legal terminology, knowledge of a particular statute, etc.). *Test-based* strategies referred to instances in which the students used test-taking strategies in selecting the correct option. These strategies included any strategy that might be related to testwiseness such as looking for patterns within the item or across the test or, more appropriately, weighing options against each other.

In the entire coding process, when the student correctly answered the item, both the discussion of the key and the discussion for each distractor were coded using these nodes. When the student incorrectly answered the item, the selected response and all other options were coded using the same nodes described above.

Phase 2 of the data analysis was primarily a discourse analysis of the functioning of items. In this intermediate phase, a set of guiding questions, listed in Table 4, helped

to summarize the information gathered during the open coding phase for each item pair. Much of this phase of the analysis depended on the item writers' justification for each item, which described reasons why the key was correct and why each distractor was considered incorrect. An example of the item writers' justification is available in Appendix D. The information in each node collected during first phase was summarized using these guiding questions. Items within a pair were analyzed sequentially.

Table 4: Guiding questions in summarization phase of think-aloud data analysis

Analysis	Guiding Question	Source of evidence
Analysis of ideal student responses	Are there students who are able to verbalize thought processes that correspond to the intent of the item writers, as indicated by the justification?	Students who correctly respond to the item
Analysis of distractors	Do students eliminate the distractors for reasons which mirror the reasons listed in the justifications?	Students who correctly respond and students who incorrectly respond but correctly eliminate the distractor
Analysis of incorrect responses	What are the reasons that students eliminate the key? Are these reasons construct-relevant or construct-irrelevant? What are the reasons that the selected response is attractive to the students? Are these reasons construct-relevant or construct-irrelevant?	Students who incorrectly respond to the item

First, the student responses for the item key were analyzed to determine if an ideal student response existed in the data. In other words, what evidence supported that

students were able to select the key for reasons that the item writers' intended? Next, the distractors were each individually analyzed to determine if there was evidence to support that students were able to eliminate these in the manner intended by the item writers. While the first question only included data from students who correctly responded to the item, this second question was also supported by students who may have incorrectly responded to the item but who may have correctly eliminated certain distractors. The final guiding questions concerned only the students who incorrectly responded to the item and investigated the reasons that they eliminated the key and the reasons that they selected their incorrect option. Based on the answers to the guiding questions, a tentative conclusion was drawn as to whether or not each item was measuring the intended construct, whether or not any revisions to the item were suggested, and whether the items in the pair were equivalent. The results to these guiding questions were written up in a summary document for each item pair.[4]

The information from the first two phases was critical for the final phase of determining whether each item pair could be considered equivalent. Because the original think-aloud transcripts were very lengthy, open coding and discourse analyses were necessary in order to identify and organize the data from the transcripts, as the original think-aloud transcripts were extremely lengthy. In Phase 3 of the think-aloud data analysis, termed pair comparisons, the summary documents compiled during Phase 2 were reviewed by two individuals trained in qualitative data analysis and familiar with the construct of legal case reading and reasoning. The individuals rated each item pair using the scale provided in Appendix E. The pairs were rated in terms of whether or not

[4] Due to the lengthy nature of these summary documents, they are not included in this paper, but are available upon request.

an ideal response was evident, whether or not the items were equivalent in terms of the types of strategies and reasoning employed, whether or not each distractor was eliminated in the intended manner by at least one student, and whether or not the types of errors that occurred in each item were generally similar. After completing this rating sheet, each rater assigned a dimension to the item pair, as described in Appendix F. The raters first independently selected a dimension for each pair, then discussed the ratings until a consensus was achieved for each pair. The dimension assigned to the pair depended upon two factors: 1) a perceived sense of equivalence between the items in terms of a matching the justification and the strategies and reasoning invoked in responding, and 2) whether or not revisions were suggested for one or both of the items in the pair. Item pairs were flagged within one of four dimensions: 1) Equivalent with no suggested revisions, 2) Equivalent with revisions suggested for one item in the pair, 3) Equivalent with revisions suggested for both items in the pair, or 4) Not equivalent.

Collection of Expert Data: In the fall of 2006, seven law professionals were asked to serve the role of legal experts in determining whether the two forms of the test were equivalent. These law professionals were professors from several law schools at various institutions including a large public urban university on the East coast, a large public rural mid-Atlantic university, and a midsized Catholic Midwestern university. Each law professional was mailed a packet of materials consisting of a set of instructions, the complete TV1 and TV2 tests including their respective cases and item sets, a set of rating sheets to rate the similarity of items, and a brief questionnaire.

In the set of instructions, the law professionals were asked to first skim over each test form including the cases and the items to get a general idea of what the task entailed.

Second, the professors were asked to complete a set of rating sheets consisting of a series of 25 item pairs. Because of the length of the items, only the stems were presented to the law professionals in the rating sheets. The distractors were not included in the rating sheets although the experts were exposed to these during their initial scanning of the tests.

Fourteen of these pairs coincided with the theoretically equivalent item pairs that were of interest in this study as presented in Appendix C. The categorization of these item pairs is as follows: 4 single-case, determinate pairs, 4 cross-case, determinate pairs, 3 single-case, indeterminate pairs, and 3 cross-case, indeterminate pairs. The remaining eleven of the item pairs consisted of randomly selected pairs of items that were not considered to be directly equivalent. The reason that these randomly selected pairs were included was to eliminate the possibility that the law professionals would realize that all the item pairs were designed to be theoretically equivalent and potentially bias their responses.

The methodology of pairing the items from test forms described above is a modified version of the method presented by Sireci (1998) who used multidimensional scaling to gather content validity information from experts. His methodology requires the experts to rate the similarity between all possible pairs of items. While Sireci's methodology is more complete, it was not feasible in this situation as the law professionals would have had to rate more than 180 item pairs. Given that the law professionals were voluntarily donating their time and effort, the task was shortened so that the law professionals were asked to rate a total of 25 pairs, the 14 theoretically equivalent item pairs and the additional 11 nonequivalent item pairs.

The experts were asked to rate each item pair on a scale from 1 to 10 with 1 being "Very Similar" and 10 being "Very Different." In order to not bias the experts' ratings, the directions provided on how they should judge the similarity of the items were scant. The experts were simply requested to "Rate the similarity of each item pair." No directions regarding what features or characteristics they should consider in their ratings were provided. Following the task of rating the item pairs, the law professionals were asked to complete a brief survey consisting of five questions asking them to describe the features of the items that influenced their ratings, whether they could detect the different types of items that were presented, and whether they thought that the items required reading or reasoning skills that were similar to what was taught in law school.

Regarding the eleven additional item pairs, it is important to note that some of these pairs would be considered more similar than others, based on the composition of the items being classified as cross-case or single-case and determinate or indeterminate. Table 5 provides a cross-tabulation of the item types for the additional eleven items.

Table 5: Cross-tabulation of additional item types presented to experts

		Determinate/Indeterminate Factor		
		Both determinate	One of each	Both indeterminate
Single-case/ Cross-case Factor	Both single-case	0	0	0
	One of each	2	5	1
	Both cross-case	0	2	1

One would expect theoretically that those items that would be categorized along the same dimension to be rated more similarly than those items that contain one of each item. For example, item pairs (even if not directly theoretically equivalent) that contain items that are both determinate or both indeterminate would be rated more similarly than those item

pairs that contain one of each type. Similarly, those pairs that contain both single-case or both cross-case items would be expected to receive more similar ratings than those pairs that contain one of each type.

Analysis of Expert Data: For each theoretically equivalent item pair, a measure of similarity was obtained by averaging the expert ratings of similarity. Item pairs with an average above 5.5 were flagged as being perceived as dissimilar among the experts and therefore potentially problematic. Items with an average rating below 5.5 were identified as being perceived as equivalent among the experts. The criterion of 5.5 was utilized because it is the midpoint of the scale presented to the experts. Although this midpoint is somewhat subjective, the items flagged through this method would reflect the most extreme ratings in terms of the experts' perceptions of nonequivalence.

In addition, the expert comments on the brief questionnaire were examined to determine other features that experts may attend to while completing the rating task.

Collection of Data for Statistical Analysis: A sample of 148 students from five different law schools completed both forms of the test. The schools included in the study include a large urban public university on the East coast, a large public university in the west, a midsized public university in the west, a large public university in the Mid-Atlantic States, and a midsized Jesuit Catholic university in the Pacific Northwest.

The data collection occurred at two different phases. During Phase 1 of the data collection, a sample of 66 students were administered the TV1 form during their first year of law school. These same 66 students completed the TV2 two years later during their third year of law school.[5] During Phase 2, an additional sample of 82 law students

[5] Note that during Phase 1, a total of 161 first-year students were administered the TV1 form. Of these 66, also completed the TV2 form during the next administration.

completed one form of the test during their first year of law school. These same students then completed the alternate form of the test during their second year of law school. For the students who participated during Phase 2, most students were randomly assigned which test form they would receive during the first-year administration with the alternate form being administered during the second year.[6]

During both phases of data collection, students were recruited using electronic bulletin boards at each location. In order to provide motivation for participation and to ensure a high quality of engagement with the task, a monetary incentive of $50 was provided to students for participating in each administration of the test.

Demographic variables collected from the sample included gender, school, age, LSAT scores, grade point average (GPA) during the first semester of law school, and cumulative GPA for the first year of law school. Within the sample, 76 were female and 48 were male (data on gender was missing for 24 students). Table 6 displays the rate of participation at each of the five schools.

Table 6: Rate of participation by school

School	**Sample size**
School A	24 (16.2%)
School B	28 (18.9%)
School C	28 (18.9%)
School D	38 (25.7%)
School E	30 (20.3%)

[6]During Phase 2, due to an administrative error, all 30 students from one law school in the sample were administered the TV1 test form during their first year of law school with the TV2 test form administered during the second year.

The average age of the students was 27 (standard deviation = 4.655). LSAT scores ranged from 143 to 174 with an average of 156.40 (standard deviation = 5.50, n=127). The average GPA during the first semester of law school equaled 2.97 (standard deviation = 0.467, n=90). The average cumulative GPA after the first year of law school equaled 2.93 (standard deviation = 0.534, n=82).

An independent sample t-test was conducted to compare LSAT and GPA scores from this larger sample to the sample of students who completed the think-aloud. Average LSAT scores were found to be significantly higher in the think-aloud sample (t=3.947, p=0.000). An effect size of the difference between the two groups is calculated by Cohen's *d* to be 0.717, defined by Cohen (1988) to be large. The average fall semester GPA and the cumulative GPA were also found to be significantly higher in the think-aloud sample (Fall GPA t = 2.141, p=0.034; Cumulative GPA t=2.12, p=0.036). An effect size of the difference between the two groups for the fall GPA equaled 0.429, defined as medium by Cohen. An effect size of the difference between the two groups for the cumulative GPA equaled 0.489, again defined as Cohen to be medium.

Analyses were performed to confirm that the year in law school in which the students completed the test forms does not pose a threat to the validity of the findings comparing the similarity of items across forms. Because it is hypothesized that the courses that students take in law school do not adequately train them in the construct being measured, it was not expected that students' test scores would change during their time in law school. Students were grouped using a variable called order, which used the following categorization: 1) students who took TV1 in their 1st year and TV2 in their 2nd year (n=32), 2) students who took TV2 in their 1st year and TV1 in their 2nd year (n=50),

and 3) students who took TV1 in their 1st year and TV2 in their 3rd year (n=66). An ANOVA using the TV1 score as the dependent variable and the order variable as the independent variable did not find a significant difference between average TV1 scores ($F_{2,145}$=1.367, p=0.258). This finding was also supported through an ANOVA using the TV2 test scores as the dependent variable and order as the independent variable ($F_{2,145}$=0.699, p=0.499).

Statistical Analysis of Data from Larger Sample: For the analysis of the test forms, first classical item analyses were performed on each version of the test separately to calculate difficulty and discrimination for each item and a reliability coefficient for each test form. The data was pooled across the phases of test administration. These analyses provided basic information about the functioning of the items and served as a starting point for examining differences for items within a pair. The software program Lertap aided in performing this analysis (Nelson, 2001). The difficulty index (termed *p*-value in the classic literature) is calculated in Lertap as the percentage of students who correctly respond to an item. The discrimination index is calculated in the program using a corrected point-biserial correlation in which the item score is correlated with the total test score minus the score on the item of interest. In addition, a factor analysis was performed using TestFact (Wood, Wilson, Gibbons, Schilling, Muraki, & Bock, 1997) on the combined dataset for all test items on both forms to determine whether the data could be considered unidimensional or multidimensional. The result of this factor analysis was important to determine the type of analysis that could be further performed. Although the results of these analyses are presented in the following chapter, the lack of unidimensionality found across both test forms dictated the further analysis. Specifically,

the lack of unidimensionality and the inability to understand the factor structure of the data prohibited the use of either unidimensional or multidimensional IRT models.

In order to determine whether the items from the TV1 form are equivalent to the items from the TV2 form, the item pairs were analyzed using a variation of conventional differential item functioning techniques. Typically in differential item functioning, the person, belonging to a particular group (i.e. reference or focal), is the unit of investigation. In these analyses, the item pair, consisting of theoretically equivalent items from each test form (TV1 and TV2), was considered the unit of investigation. Table 7 below details the differences between conventional DIF and how these methods may be applied to the analysis of test versions.

Table 7: Comparison of conventional DIF and DIF applied to analysis of forms

Conventional DIF	**DIF applied to analysis of forms**
One item, two examinee groups	One examinee group, two items
Unit of analysis is examinee	Unit of analysis is the paired item but ability of examinee is controlled for
Two groups of participants take the same test	One group of participants take two test forms
Independent variable: Group membership (i.e. black versus white, male versus female)	Independent variable: Test form
Estimate of ability: Total score on test (or IRT ability estimate)	Estimate of ability: Omnibus score obtained by summing scores from both test forms (or IRT ability estimate across both test forms)
Item is considered to exhibit DIF when group membership and/or interaction is significant predictor	Item is considered to have exhibited non-equivalence across forms when form and/or interaction is significant predictor

Due to the nature of the dataset, a combination of several established DIF methods as well as other statistical methods that have not yet been previously used for DIF were used in this study to examine the equivalence of paired items on the test. First, the delta plot method was used to graphically examine the item pairs. Following the guidelines established by Angoff (1982), difficulty indices were converted to a delta value, a normal deviate using a scale of the mean equal to 13 and the standard deviation of 4. Because one minus the difficulty index is used to calculate the delta value, higher delta values indicate more difficult items, whereas lower delta values indicate easier items. A delta plot was constructed with the delta values for the TV1 form on the vertical axis and the values for the TV2 form on the horizontal axis. A principal axis line was constructed through the origin of the graph. Item pairs with delta differences greater than 1.5 were flagged to be potentially problematic, following recommendations by Holland and Wainer (1993) and Muñiz, Hambleton, and Xing (2001). Two parallel lines were included on the delta plot to graphically display the effect size criterion. Those pairs which fell outside of the parallel lines were flagged for potential DIF.

Other graphical procedures were also performed in order to provide a visual understanding of the item pairs. Empirical item characteristic curves (ICC) were created for each item pair, plotting the total score against the observed proportion correct. Because the empirical ICCs can be difficult to interpret, additional graphs were created which collapsed the total score into four categories (0-10, 11-14, 15-18, and 19-28). These graphs provide a better understanding of the trends occurring to the proportion correct based on total score.

Next, two additional procedures were performed to establish indices for the size of DIF for each item pair. First, the signed proportion difference index, SPD-X as described in Chapter 2, was calculated for each pair. In the calculation of this index, the weighting used to account for the total scores (indicated by n_{Fj} in traditional DIF) was the number of individuals with each observed score on TV2, treating this second form as the focal group. When the signed area index is positive, the item is easier for the TV1 form. When the signed area index is negative, the item is easier for the TV2 form. Thus, the highest values represent "uniform" DIF meaning that the item is easier on one test form across the ability scales. Using the index suggested by Dorans and Kulick (1986), items with an SPD-X value greater than or equal to 0.10 were flagged for potential DIF.

Next, the Mantel-Haenszel log odds ratio was calculated to establish indices for DIF size. Again, the formula for this index is provided in Chapter 2. The natural logarithm was obtained for ease of interpretation. A positive value of the index provided evidence that the item in TV1 was easier than the paired TV2 item. A negative value of this index provided evidence that the item for TV2 was easier. Indices close to zero indicated that the items on both form had similar levels of difficulty across the range of total score values. The pairs with the absolute value log odds ratio greater than or equal to one were flagged for potential DIF.

Because of the issues with utilizing only one single-group for the analysis, rather than two groups as performed in traditional DIF, the standard techniques such as logistic regression, the summed chi-square, and the Mantel-Haenszel chi-square test could not be performed. Each of these standard techniques requires the use of independent observations. In addition, because of lack of unidimensionality and the difficulty

understanding the factor structure of the test, IRT methods could not be performed. Therefore, a new approach to examining the item functioning was performed using a modified version of the McNemar test, which is a test of the proportions for dependent samples. Specifically, the McNemar can examine the difference in observed frequency of individuals who shift from a positive response to a negative response within the item pair and the observed frequency of individuals who shift from a negative response to a positive response within the same item pair. As Seigel and Castellan (1988) notes, "…these are studies in which people could serve as their own controls and in which nominal or categorical measurement would be used to assess the 'before to after' change" (p. 75). Table 8 below signifies the table used in testing for the significance of the changes.

Table 8: Table for use in the testing of significance of changes.

		TV2 Item	
		Correct Response	Incorrect Response
TV1 Item	Correct Response	*A*	*B*
	Incorrect Response	*C*	*D*

Within Table 8 above, the cells of interest are *B* and *C*. One would expect under the null hypothesis that ½(*B*+*C*) changed in one direction (i.e. from correct to incorrect) and ½(*B*+*C*) changed in the other direction. In other words, "the null hypothesis is that the number of changes in each direction is equally likely" (Seigel and Castellan, 1988, p. 75). The formula used to calculate the sampling distribution under H_o is as follows:

(7)

$$\chi^2 = \frac{(B-C)^2}{B+C}$$

which is distributed as a chi-square with one degree of freedom. If the expected cell frequencies are small, the following corrected formula is used:

$$\chi^2 = \frac{(|B-C|-1)^2}{B+C} \quad (8)$$

which again is distributed as a chi-square with one degree of freedom. In this study, the McNemar test was first performed across all total test scores (without breaking down into smaller intervals) to test for the significance of changes. The null hypothesis is that the probability of students to correctly respond to the TV1 item but incorrectly respond to the TV2 item is equal to the probability of students to incorrectly respond to the TV2 item but correctly respond to the TV1 item. This test will then provide an indication of whether the item is functioning in the same manner between forms across all test score levels. Each student serves as his or her own control for the ability level, as each person completed each test form.

A follow-up test was then conducted for each item pair to more closely examine how the items were functioning between forms. Because the first test only examines the functioning across all total test score levels, this follow-up test attempts to gather information about whether the differential functioning varies based on the total score. Thus, total test scores were broken down into two score levels, low being from 0-14 And high being from 15-28. Attempts were made to break down the total scores into smaller intervals, but this resulted in too few observations per cell. The McNemar test was replicated for the frequencies within each score level, resulting in two chi-square indices:

one at the low score level and one at the high score level. These chi-squares were summed to obtain a test for uniform differences between the item, resulting in a chi-square with two degrees of freedom. The corresponding *p*-value was obtained. Those items with *p*-values of less than 0.05 were flagged as potentially being problematic. Because of the single-group design, the McNemar test cannot be a method of detecting DIF, which examines differences between two groups. However, the McNemar test could be viewed as a method of equating at the item level.

In order to flag pairs as potentially containing DIF, all of the above methods, the delta plot, the SPD-X, Mantel-Haenszel odds ratio, and the McNemar test, were used. Those pairs which were identified as being potentially problematic using all of the above indices were flagged as being nonequivalent.

Comparison of Methods (Hypotheses 1-3): The kappa coefficient of agreement was calculated to obtain a level of concordance between each pair of methods. The kappa coefficient is calculated by the following:

$$\kappa = \frac{P_A - P_E}{1 - P_E} \tag{9}$$

where $P(A)$ is the proportion of times that the methods agree and $P(E)$ is the proportion of times that the three methods would be expected to agree by chance. A K of 0 would indicate that there is no agreement among the methods, except that which may occur due to chance. A K of 1 would indicate complete agreement among the methods. In addition, a 2x2 contingency table was constructed in the form displayed in Table 9.

Table 9: 2x2 contingency table for test of comparison hypotheses

		Method 1	
		Equivalent	Nonequivalent
Method 2	Equivalent	Agreement between methods (*A*)	Disagreement between methods (*B*)
	Nonequivalent	Disagreement between methods(*C*)	Agreement between methods (*D)*

Comparison of methods by item type (Hypotheses 4-5): In order to compare the items flagged by each method by the type of item, a 2x3 contingency table was constructed in the manner displayed by Table 10. The cells contain the frequency of nonequivalent item pairs, displayed for each method by item type. The example in Table 10 displays the contingency table for the single-case versus cross-case dimension.

Table 10: 2x3 contingency table for Fisher exact probability test

	Think-aloud	Expert	Statistical DIF
Single-case items	*A*	*B*	*C*
Cross-case items	*D*	*E*	*F*

An additional table was constructed for determinate versus indeterminate item types. Because of the very small expected cell frequencies in this table, the Fisher's exact probability test for 2x3 contingency tables was performed to test whether the observed probability of the frequency distribution is what would be expected from chance. The null hypothesis tested is that there is no association between the dimension

of item type and the method used to flag nonequivalent items. Specifically, the Fisher test asks the following question: if any relationship between the item type and method were the result of nothing more than chance, how likely is it that we would end up with an observed result as extreme or more extreme? A one-tailed test was performed to determine the probability of the observed outcome or one that would be more extreme. The calculations were performed using the SAS program.

Chapter 5:

RESULTS OF THINK-ALOUD ANALYSES

The results of the analyses from the think-aloud protocols are discussed in this chapter. The analysis of item pairs within each classification is described below. Appendix G provides a chart with conclusions on all item pairs based on the rubric that was used to judge equivalence or non-equivalence.

Single-Case Determinate Item Pairs

Of the four single-case determinate item pairs, three were found to be equivalent based on the students' response processes. Pairs 1, 3, and 4 were rated to both contain response processes from students matching the item writers' intent. Pair 2 was judged to be nonequivalent. Summaries on each pair are provided below.

Pair 1

This pair, focusing on the summarization of reasoning in the main case, was rated to be equivalent by both raters. In the TV1 item, students needed to recognize that the *Mackey* court concluded that Mr. Mackey's attempts to pay his court costs were neither "valid" nor "honest," following the precedent case of *Black and Brown.* In addition, students needed to recognize the court's view that Mr. Mackey should have known the rules and acted accordingly. For the TV2 item, students are asked to recognize the statement that best reflects the reasoning used by the district court in awarding judgment to the defendants in *Bridell.* The students need to recognize the two reasons for their judgment: that Ribier did not promise lifetime benefits to its employees and that no evidence of prohibited conduct was presented that showed the company interfered with

benefits Bridell was entitled to under ERISA. Both items were hypothesized to be relatively easy because they focus on explicit information from only one of the three cases in the set and because students' attention was likely to be most focused on the primary case of their hypothetical client. All students in the think-aloud sample correctly responded to the TV1 item; seven of the eight students who received the TV2 item correctly responded.

Across both items, evidence supported that the items functioned consistently with the item writers' justification. Ideal student responses were easily found for both items as was evidence that the distractors functioned in the manner intended. Regarding the strategies utilized to respond to the items, some evidence did point to the fact that students were required to do more rereading of the text for the TV2 item in order to correctly respond to the answer. In contrast, students who received the TV1 item were much more likely to use strategies based on memory and did not have the need to refer back to the text as often. This difference in requisite strategies could potentially explain possible differences in difficulty levels for the TV2 item as compared to the TV1 item. However, this difference was judged to be relatively minor and did not impact the raters' categorization of the pair as equivalent.

As mentioned above, most students in the think-aloud sample correctly responded to the item. For the one student who incorrectly responded to the TV2 item, the type of error was an over-reliance on the test-based strategy of pattern finding. The student justified his selection of option D by stating, "That answer is different than all the others. All the others give reason 1 and 2, reason 1 and 2, reason 1 and 2." From the item writing guidelines suggested by Haladyna, Downing, and Rodriguez (2002), the item

writers may want to consider having all of the options follow a similar pattern to avoid student reliance on faulty test-based strategies as exemplified here. That being said, the student used a faulty test-taking strategy in looking for patterns within the item options.

Another minor suggested revision concerned the key for the TV1 item. The key described the reasoning described above, that the court concluded that Mr. Mackey's attempts to pay were neither valid nor honest and that Mr. Mackey should have known the rules. Included in the key is the following statement, "…but otherwise offers little reasoning concerning why Mackey's attempts to learn his costs and the correct due date are neither 'valid' nor 'honest.'" This statement does not actually summarize the court's reasoning but is an inference not explicitly garnered from the text itself. Given that this item is intended to be determinate, relying on explicit textual references, this inference may be inappropriate and would be better suited for an indeterminate item. The student evidence from the protocols did not support that the inference highly impacted response, but in order to make the item conform to the intended manner, this phrase should be removed.

Although the protocols suggested that both items needed to be revised, the raters considered these changes to be minor and did not impact equivalence of the items. Rather, because ideal responses were elicited and because the strategies necessary to respond to the item were similar, the raters judged the pair to be equivalent.

Pair 2

Of the single-case determinate items, Pair 2 was found to be most problematic in terms of construct equivalence. The two items in this pair both focus on legal issues in the cases. The TV1 item focuses on the legal issue within the *Mackey* court while the

TV2 item focuses on the legal issue that emerged in the *Alexander* case. The TV2 item varies slightly in that it asks the student to consider this legal issue in light of the decision that will be brought back to the district court. As mentioned in Chapter 2, the decision of the *Alexander* court was to remand the case back to the district court, as the finder of facts. These differences in the question stem are relatively minor and can be attributed to the differences presented in the individual cases in that the *Mackey* case resulted in a clear decision whereas the *Alexander* case resulted in a remand back to the district court. These two items were not hypothesized to be very difficult, requiring only basic textual references to a single-case. Some knowledge of the definition of a legal issue was required, but this was considered to be something that a law student should learn very early in the first semester of law school. In the think-aloud sample, four students correctly responded to the TV1 item whereas only two students correctly responded to the TV2 item.

The differences that emerged on analysis concerned the nature of the options. The TV1 item required students to identify which question best expressed the legal issue and presented options that varied in their level of specificity. The best expressed option contained a level of specificity that was evident from the cases while also presenting a legal issue that can be used as a rule for other cases. An example of an incorrect option concerned the very specific factual situation of Mr. Mackey, listing exact details of the case such as the fact that Mr. Mackey "paid $237 of the $327 due for record costs two days late." Such a response would be incorrect in that it is too specific to be generalizable to other situations. Another example of an incorrect option is one that would be too general, such as the option that stated, "Under Pennsylvania rules, what

reasonable conditions may a court impose upon the right of appeal from arbitration?" This example is too general as the *Mackey* case focused on only one specific aspect of the conditions for the right of appeal from arbitration, namely the record cost issue. The correct answer was specific enough to focus on the issue of nonpayment of record costs, yet general enough to be generalizable to other legal cases and also contained language that was closest to that used by the court itself. Students who correctly responded to the item seemed to recognize the language of the court and some considered the level of specificity. For example, one student, when reading the very factually specific option stated, "It doesn't best express the legal issue because it's so factually specific to this. And you're not going to be able to abstract any sort of legal concept from that. So you have to make it a little more general." The verbal protocols provided evidence that this item, in terms of the reasoning, supported the intent of the item writers.

In the TV2 item, once again students were supposed to consider the legal issue presented, but in the *Alexander* case. The options presented to the students were of a very different nature to those presented in the theoretically equivalent item. In this item, students were not asked to consider which option is correct in light of the generality/specificity of legal issues. Rather, the student is asked in this item to discern which statement best summarizes the issue amidst options that present "issues" that either are not discussed in the case or are incomplete. The correct answer for this item follows:

> The district court must decide whether American Can's prior behavior in raising the cost of plaintiffs' medical insurance premiums supports the company's claim that it unambiguously reserved the right to reduce plan benefits and never promised lifetime, fixed-cost benefits in its benefit plan documents.

According to this option, the circuit court remanded for two reasons: 1) the district court needs to determine whether the reservation of rights disclaimer in the defendant's SPD can be unambiguously construed and 2) the district court needs to determine whether the defendant's prior behavior elucidates the reservation of rights that employees were never promised lifetime benefits. The distractors for the item either point to issues that were not discussed in the case or only present one of the two reasons for the circuit courts' decision to remand. For example, one distractor that was attractive to the students stated the following:

> The district court must decide whether the benefit plan documents provided by American Can unambiguously reserve the right to alter participants' benefits or whether these documents unambiguously promise lifetime benefits that cannot be altered.

This statement is actually true about the circuit court's reasoning for remanding the case. However, the option is incomplete in that it misses the other reason provided by the court: determining whether the prior behavior of the company elucidates the promise regarding lifetime benefits. Not surprisingly, this option seems to be very attractive to some test takers.

The options in TV2 are functioning differently than those in TV1 in that the student needs to read very closely within the text to gather the two reasons the court provides for remanding the case. For the TV1 item, students need to have basic knowledge of what constitutes a legal issue and its purpose of being able to generalize to other cases. On face value, these two items may have been considered equivalent. However, further analysis of the strategies and reasoning elicited through the verbal

protocols suggests that these items are asking different questions. The suggestion to the item writers is that an additional item be added to each form. For TV1, an item could be added which specifically measures the ability of the student to discern the legal issue. The distractors would focus on issues that either do not emerge in the case or perhaps only focus on one aspect of the legal issue presented by the court. For TV2, an item could be added that requires students to identify the option which most appropriately summarizes the legal issue. The distractors would include the very specific factual situation in the given case as well as very general statements of legal issue that do not succinctly summarize the case at hand. By adding these two additional items, the equivalence of the test forms could be better supported.

Individually, the items were examined to see if they are functioning as intended, even though equivalence is not supported. For both items, an ideal response emerged from students who correctly responded to the items. The distractors also functioned generally as intended, although minor revisions were suggested for choice A in TV2.

For the TV1 item, several students eliminated the correct option due to their interpretation of the wording of "attempt to pay" and "nonpayment." The court considered Mr. Mackey's attempts to pay the court costs two days late as nonpayment. Some students seem to consider a nonlegal definition of the term "nonpayment" and felt that Mr. Mackey did pay the costs, albeit late. As one student eliminated the key "primarily because they call them an attempt to pay them two days late in C. And it wasn't an attempt to pay them two days. It was successfully paying them two days late." Another student stated, "Well, this wouldn't be the issue because the defendant made some payment and it was just late, so he did try beforehand." Yet another student stated,

"It isn't a nonpayment." This is in direct contradiction with the judge's language that Mr. Mackey engaged in "nonpayment of record costs." Thus, this error is related to the construct of interest and is related to student shortcomings rather than item shortcomings. One other error type that emerged was related to a student's use of test-based strategies, matching terminology from this item to another item he[7] had already completed.

An interesting construct-relevant error emerged in the analysis of the TV2 item in Pair 2. Option B was an attractive response in the think-aloud sample. This option is partly true in that it discusses whether the court unambiguously reserved the right to change benefits but is incomplete in that it is missing whether the company's prior behavior supports the claim. In order to eliminate option B as being incorrect, students need to be aware of this important second fact regarding prior behavior. Students who selected option B focused only on the one clause and didn't feel that "prior behavior was a big factor in *Alexander*." When considering why students missed the idea of "prior behavior," a close examination of the case revealed that the answer to this item is located in two separate places in the text. The students who incorrectly responded focused on the following statement which appeared at the very end of the case:

> Because the summary plan documents do not clearly reserve the right to reduce benefits, we will reverse the distract court's grant of summary judgment dismissing the complaint. We will remand the for the district court's interpretation of the summary plan descriptions in light of *all relevant evidence* and for the district court's further consideration of the retirees' claims. (Italics added for emphasis.)

[7] Note that because gender was not of particular concern in this study, the pronoun of "he" versus "she" was randomly assigned when discussing specific students' comments.

The italicized phrase above acts as a referent to evidence related to prior behavior that was discussed several pages previously. Students who only focused on this phrase may have missed the court's position on prior behavior. This error is related to students' abilities to closely read the text and thus is considered construct-relevant.

In conclusion, individually, when not examining the correspondence between test forms, the two items seem to function as intended. However, Pair 2 needs to be carefully considered by the test developers as the evidence presented from the students' think aloud protocols suggests that different questions are being asked.

Pair 3

Pair 3 contains items that ask students to summarize the court's reasoning in one of the precedent cases. In the TV1 item, students are asked to summarize the reasoning in *Black and Brown.* In order to correctly respond to the item, students need to recognize that the court quashed the attempts at appeal as the defendant made no attempt to pay and overruled *Meta* saying that if this decision was allowed to stand, it "would allow defendants to simply ignore the prepayment requirement altogether." In the TV2 item, students are asked to summarize the reasoning of the *Alexander v. Primerica* case. The students need to recognize that the company's reservation of rights clause was judged to be ambiguous and specifically fails to state whether the right is unlimited or limited. Additionally, the students need to be aware of the court's view that the prior behavior of the company did not clarify the meaning of the reservation of rights clause. Once again, these items were hypothesized to be relatively easy, as they focus on only one case and on information that is explicit from the text. In the think-aloud sample, only one student incorrectly responded to the TV1 item; all students correctly responded to the TV2 item.

The evidence collected in the think-aloud protocols supported that the keys and distractors for both items were functioning as intended, with students who correctly responded to the items demonstrating reasoning consistent with the item writers' justifications. As hypothesized, students used a combination of memory and rereading text-based strategies at an equivalent rate when responding to these items. Both items required basic understanding of a legal concept: the TV1 item requiring knowledge of the appeal process and the definition of "quash" and the TV2 item requiring knowledge of the definition of "remand." These are very basic definitions that students would have encountered very early in their courses or could have looked up in the legal dictionary which was available for use during test administration.

The student who erroneously responded to the TV1 item did so for reasons related to the construct of legal case reading and reasoning. This student did not have an understanding of the terminology *de minimis,* which is a Latin phrase meaning "trifling" or "insignificant." Again, although the legal dictionary was available for use, the student did not use it in order to clarify her understanding. She treated the terminology as if it were a law that was meant to be followed, matched the language from the option to the case, and thus followed the wrong path to his conclusion. She failed to consider the full holding of the court and the reasoning attached to that holding. The raters considered these errors to be relevant to the construct of legal case reading and reasoning.

Because the ideal responses and the reasoning behind the elimination of the distractors supported the item justifications, no suggestions were made regarding revision of the item. The items within the pair were rated as equivalent.

Pair 4

Pair 4, also rated as equivalent, requires students to summarize the reasoning within a single case that is less central to the main reasoning of the court. The TV1 item focuses on the reasoning of the minority opinion or dissent in the *Meta* case whereas the TV2 item focuses on the court's reasoning on a second issue involved in the *Hamilton* case. This pair represents items that are less isomorphic in nature, but still would be considered theoretically equivalent as they are single-case and focus on determinate, explicit textual information related to less central themes in the cases. In the TV1 item, students are required to identify that the minority opinion reasoning that the majority did not provide reasons why the 20-day limit for paying the record costs was unreasonable. In addition, the minority felt that the defendant in the case had the responsibility to find out the costs owed through the prothonotary's office. The TV2 item requires students to recognize the court's reasoning regarding a second issue in the case, namely that Air Jamaica's use of a disclaimer in their benefit handbook is part of the promise made to employees regarding their benefits and that this disclaimer provides information to the employees "that they have no guaranteed benefits." Both items were hypothesized to be relatively easy, but more difficult than the other single-case determinate items as they focus on a more specific aspect of the case. Each item should require the students to refer back and closely reread sections of the case in order to correctly respond. In the think-aloud sample, only three of seven students correctly responded to the TV1 item; five students of eight correctly responded to the TV2 item.

Once again, evidence in the think-aloud protocols demonstrates that the items are functioning as intended with ideal student responses supporting that students who

correctly respond to the item do so for reasons listed in the justifications. Also, evidence supported instances of the distractors being eliminated for reasons intended by the item writers.

Analysis of the incorrect responses revealed that a phrase in the key of TV1 may be confusing to the students. The key includes the following in its summarization of the minority reasoning: "It argues that, in order to vacate the statute, the majority should have made (but failed to make) an argument that the record cost statute is unreasonable…" The term "vacate the statute" may possibly be causing students to eliminate the option. As one student stated, "They're not trying to vacate the statute." Other students dismissed the option quickly with phrases such as, "From what I see that isn't involved" and "That's not in the dissent at all." While the latter phrases are not very informative, the raters hypothesized that the "vacate" phrase may not be appropriate. A decision was made to raise this question to a legal expert to determine if the phrasing should be changed. In the meantime, the phrase was not judged to be critical to the equivalence of the item pair. Another minor suggestion discussed by the raters is that option B for the TV1 item should be rephrased slightly to make it "more wrong" and thus potentially less attractive to the students. The option was somewhat correct, although not completely. This option was selected by all students who incorrectly responded to the item. These two suggested minor revisions may make the item function more closely to the intent of the item writers. While the confusing phrase and the attractiveness of option B were related to students' incorrect responses, an additional construct-relevant error concerned a failure to return to the text and an over-reliance on memory of the case. Several of the students who incorrectly responded did not return to the text at all or

employed selective reading strategies of looking in the case to find a quote that supported their option choice. Given that this item focused on a less central aspect of the case, the strategy of returning to the case was necessary in order to correctly respond.

No revisions were suggested for the TV2 item in the pair. In this item, the construct-relevant errors that emerged in the protocol appeared to be related to student shortcomings regarding the construct being measured. One student selected his choice without logical reasoning, simply using verbiage and throwing around terms to justify his decision. As he stated while reviewing the correctness of the key, "That's some reasoning but it's not really using it by the decision. It's sort of a counter analysis in some sense. It's not really the reasoning of their decision on the second issue." Another student incorrectly responded because he over-relied on memory, attempting to match the language of the option to that of the case without going back. This error is demonstrated by the following statement, "That language is a little different than explicit language that I saw in this case." These two errors are related to the construct of legal case reading and reasoning. Another student made an error in that he lost track of the question being asked in the stem and answered a different question, focusing on the first issue that the court discussed. This error, while not directly related to the student's ability to read and reason through legal cases, may be more indicative of his test-taking strategies.

Based on the evidence supporting ideal responses and distractors being eliminated as intended, the raters judged these items to be equivalent with minor revisions necessary for the TV1 item, which may need some changes to the wording of the key and an overly attractive but incorrect response. With the exceptions for TV1 described above, all other

errors that emerged are more indicative of the students' abilities on the targeted construct and their test-taking skills and do not reflect any shortcomings with the items.

Cross-Case Determinate Item Pairs

Several problems emerged among the cross-case determinate item pairs. Pairs 5 and 7 were judged to be equivalent although revisions were suggested for at least one item within each pair. Pairs 6 and 8 were deemed nonequivalent, also with suggested revisions to one or both items within each pair.

Pair 5

Pair 5 asks the students to summarize the legal issue that emerged across all three related cases. For the TV1 item, students need to recognize that the commonality among the cases concerns the ways in which parties may not completely comply with the procedural statute requiring losing parties in arbitration to pay all court costs within a 20-day period. For the TV2 item, students need to recognize that the commonality among the cases concerns the conditions by which a company's reservation of rights clause satisfy ERISA's oftentimes competing goals of both informing employees of their rights and also protecting the company's rights to offer or not offer benefit plans. These items were hypothesized to be quite difficult for the students as they require synthesis across all three of the cases, which is an important skill for lawyers to possess. For the TV1 item, only two students correctly responded to the item; for the TV2 item, three students correctly responded.

This pair was judged to be relatively unproblematic regarding the strategies and reasoning elicited during the think-aloud process. Both items yielded ideal responses and reasoning from at least one student. In addition, students who correctly responded to the

item showed the hypothesized strategies of synthesizing the information across the three cases and recalling or going back to the text. While the raters judged these to be equivalent, there was some evidence that students in the TV2 item were not synthesizing as well as they should have been able to do. However, because an ideal response was still elicited, this item was not judged to be problematic regarding equivalence in terms of strategies or ideal responses.

The main problem within this pair concerned the wording of the key in the TV1 item. The correct response to this item summarizing the *Mackey*, *Meta*, and *Black and Brown* cases stated:

> In what ways, if any, may a party fail to comply with the procedural statute requiring a losing party in arbitration to 'pay all costs that have accrued on such suit or action…within 20 days after the entry of the award of the arbitrators on the docket' and yet proceed with an appeal?

An issue emerged in the think-aloud protocols that was not intended by the test developers. Specifically, students interpreted the phrase "in what ways" to signify intent on the part of the individual. This assignment of intent is supported by students' quotes such as the following, "I don't think that they're looking for ways to find out to what extent you could go to not comply with the statute." Another student stated the following:

> I don't think [the option] summarizes the issue because it's not phrased correctly…I mean it's more looking for the end result. So how can they fail to comply? I think it's what they can't do is more the issue. The issue isn't what they can do to fail to comply and still get [an appeal].

These comments from the students indicate that the item key is not functioning as intended with the key wording introducing a source of construct-irrelevant error influencing students' responses. The suggestion to the item writers is to rephrase "in what ways" to something that implies less intent such as "under what circumstances." With this minor change, the item should function more closely to the original intent of the item writers. The raters also judged Option B to be lacking in evidence to support its functioning. However, this was judged to be relatively minor in light of the problems with the key described above. Other than the problem with the key, student errors on this item were related to the construct of legal case reading and reasoning, such as vocabulary failures with the court's application of the terminology *de minimis*.

The TV2 item functioned as hypothesized eliciting an ideal response and reasons for eliminating distractors that matched the justification of the item writers. The errors that emerged in this item included a failure to refer back to all three cases or relying on only one case, misunderstanding the definition of a legal issue (i.e. "None of the cases said what the legal issue was."), over-reliance on intuition rather than reasoning (i.e. "…[T]his one sounds like the most legal of all the issues."), utilization of legal terminology in a commonplace manner (i.e. the court's definition of "welfare benefits" referring to employee benefits), and a failure to read item options carefully. All of these were deemed to be construct-relevant, indicating student rather than item shortcomings.

In conclusion, this item pair was judged to be equivalent with a revision required in the wording of the key to the TV1 item. The TV2 item did not have any suggested revisions. Both items, with the exception of the key in TV1, had evidence that the items were measuring the intended construct and functioning as intended.

Pair 6

Pair 6 asked students to summarize the similarities and differences between the facts presented in the instant or main case under dispute (*Mackey* or *Bridell*) to those presented in one of the precedent cases. The TV1 item asks the students to compare the facts in *Mackey* to those in *Meta* and select the statement which would best summarize the similarities and differences in light of the task of submitting a motion to the court to reconsider its decision against Mr. Mackey. In order to correctly respond to this item, students need to recognize that while both defendants failed to pay the complete amount owed, the important difference between the cases concerns the fact that Mackey paid 2-days late while the defendant in *Meta* paid on time. The TV2 item asks the students to compare the similarities between the facts *Bridell* to those in *Alexander* in light of the decision to appeal the Ms. Bridell's case. In order to correctly respond to the item, students need to recognize that the main factual similarity is the use of the phrase "in conformity with applicable legislation" in *Alexander* and the use of the phrase "pursuant to changes in federal legislation" in *Bridell.* These items were hypothesized to be more difficult than all the single-case determinate items, but less difficult than other cross-case determinate items, as they only require the comparison of two cases. In the think-aloud sample, a total of three students correctly responded to the TV1 item; a total of seven students correctly responded to the TV2 item within the pair.

In the TV1 item, while an ideal response was elicited and each the distractors had evidence of being eliminated for the intended reasons by some of the students, a problem emerged from the think-aloud protocols regarding the item options and the key. The

options for the item are listed below in the order presented to the students in the think-aloud sample:

a) Except for the different amounts of record costs involved, the fact patterns in these two cases are identical: both defendants tried and failed to contact the Prothonotary's office; both paid their record costs late; and both failed to pay the full amount that they owed.

b) The only factual similarity of any importance is that both defendants failed to comply with the record cost statute. The key difference is that the defendant in Meta was allowed to proceed with an appeal.

c) The *Mackey* case is distinguishable from *Meta* because Mr. Mackey paid about 72% of what he owed, whereas Yellow Cab paid only 57% the amount owed. The cases are similar in that both defendants argued that they had to deal with apparently deliberate mis-statements from the plaintiffs when trying to learn the true amount of costs owed.

d) The relevant similarity between the facts of these cases is that both defendants violated the statute because both defendants failed to pay the full amount, while the relevant difference is the large gap between what Mackey owed ($90) and what Yellow Cab owed ($7.50).

e) The relevant factual similarity is that the payments of both defendants fell short of the amount owed. However, a comparison also shows an important difference in that Mackey paid most ($237 of $327) of the full amount that he owed two days late, whereas Yellow Cab paid just over half ($10 of $7.50) of the amount owed on time.

The correct response to the item is choice E. Three of the options listed, C, D, and E, all mention the costs of the payments albeit in different formats (i.e. percentages versus dollar amounts). The important feature about choice E is that is the only option that includes the important time difference for Mr. Mackey paying "two days late" and Yellow Cab, the defendant in *Meta*, paying "on time." By the time many students got to reading choice E, they seemed primed by choices C and D to focus on the amounts specified and skip over the critical features of choice E. For example, one student stated, "[The option] talks about how Mackey paid most of the full amount whereas Yellow Cab paid just over half. I don't think that's so important either." The time issue is "buried" within the option causing students to miss this important distinction. Another student noted, "I think D and E are actually pretty similar. They just switched the parties [of who paid more of the amount]." There were several students who correctly noticed the time differences and selected E as their response choice. However, the fact that several of the students within the think-aloud sample seemed to focus on the amounts rather on the time issue as listed in the option, suggests that the key to this item needs to be revised. Another revision that is necessary for this item is the change of the word "identical" in choice A. One student used test-based strategies to eliminate this option because of this word. He stated that, "Identical is a dangerous word in multiple-choice exams." A less absolute phrase should be substituted within this option.

The TV2 item functioned closer to the intent of the item writers. As in TV1, an ideal response for selecting the correct response was identified within the protocols. In addition, the distractors seemed to function as intended. No suggestions were made regarding the revisions for this item. One type of error that emerged in the item was

losing track of the question asked in the stem. This is evident by one student who focuses on the similarities between the facts in *Bridell* and the facts in *Alexander* without considering which would be most useful to the defense of Bridell. This error was considered to be related to the construct.

While both items elicited ideal responses, a comparison of the TV1 and TV2 items within Pair 6 suggested that they are not measuring the same construct. Both items in the pair asked the students to consider the similarities between the two cases in light of a given task of appealing the case for the client. However, the TV1 item focuses specifically on the comparison of the factual similarities and differences between the main and precedent case whereas the TV2 item focuses more on the usefulness of each option. This difference is important when considering the equivalence of the items, particularly considering that one of the types of errors that emerged in the TV2 item concerned the student losing sight of the given task of appealing the *Bridell* case.

These items also elicited different strategies and reasoning. Selecting the key and eliminating the distractors in the TV1 item rely heavily on the use of text-based references. In other words in order to correctly respond to the item, students need only remember the specific facts of each case and identify that the main similarity is that they both failed to pay the complete record cost statute and the main difference being that Mackey paid the costs two days late. The distractors either contained information that was erroneous or left out the distinction of Mr. Mackey paying the fees too late. On the other hand, the TV2 item focuses much more on the usefulness of the options rather than on their factual correctness or incorrectness as in TV1. In order to correctly respond to the item, students need not only to consider the factual similarities between the two cases,

they also need to consider what would be most useful in the task of appealing the client's case. The question in TV1 seems to imply that the task of asking the court to reconsider the client's case should influence the students' choice of options. However, the analysis of the students' reasoning suggests that the TV1 item requires more textual references and fewer instances of logical reasoning and evaluating the options to correctly respond.

In conclusion, these two items were rated to be nonequivalent with significant revisions required for the TV1 item. Given the need for these substantial revisions, the item writers may want to consider the inclusion of the construct of usefulness, in order to ensure that the items are measuring the same thing.

Pair 7

Pair 7 was judged to be equivalent by the raters although revisions were suggested for both items. The items within this pair ask the students to compare the courts' reasoning for two trials. The TV1 item asks the student to summarize the similarities and the differences between the reasoning in *Meta* with that in *Black and Brown.* In order to correctly respond to the item, students need recognize that the reasoning between the two courts differs in that the *Meta* case follows the sufficiency of substantial compliance invoked in earlier cases and that the record court requirement is deemed a trifle matter in comparison to the amount of money involved in arbitration decisions. In contrast, the *Black and Brown* court does not focus on the disparity of amounts of money. The TV2 item asks students to summarize the differences in the reasoning in the district court case to that of the circuit court case as evident in the *Hamilton* case.[8] Students would need to recognize that the district court felt that Air Jamaica's use of the reservation of rights

[8] Note that the students did not receive a copy of the circuit court case but rather had to rely on the references by the district court in the *Hamilton* case to respond to the item.

disclaimer went against ERISA's policy of protecting employees' expectations. In contrast, the circuit court felt that Air Jamaica's reservation of rights clause put employees on notice that they had no guaranteed benefits, consistent with ERISA's policy. These two items were hypothesized to be moderately difficult as students need to be able to synthesize the court's reasoning for both cases and to make evaluations on the similarities or differences between the two. Within the think-aloud sample, only two students correctly responded to the TV1 item and three students responded correctly to the TV2 item.

At face value, the raters noticed that the TV2 item only focuses on the differences in reasoning whereas the TV1 item asks the student for both similarities and differences. Restricting the TV1 item to requesting only a summary of the differences was suggested in order to make the questions posed more similar. However, this surface feature was not the source of the problem with the items.

A close inspection of the think-aloud protocols for the TV1 suggested that this item is not functioning as intended. The question stem for this item states, "Which one of the following statements best summarizes the similarities and differences between the court's reasoning in *Meta* and the court's reasoning in *Black and Brown*?" The correct answer to this stem follows:

> The reason used by the two courts differs: although both courts refer to the principle of the sufficiency of substantial compliance in their decisions, the *Meta* court further reasons that the record cost requirement is a matter *de minimis* because in comparison with the amount of money in controversy in arbitration

decisions, record costs are usually quite small. The *Black and Brown* court does not discuss this disparity in amounts involved.

The issue that emerged in this item is that the option does not really summarize the similarities and differences between the two cases. Rather, the question seems to ask the student to summarize the similarities and differences in the courts' reasoning in light of the *Meta* case. The other differences in the *Black and Brown* reasoning do not appear in the key. Students who correctly respond to the item note that that something is missing from the stem. For example, student 29 states, "This may be my answer. I may have to go back and look…But they're not really saying. I like this answer but the further reasoning is more that [the *Meta court]* makes [the record cost statute] directory rather than mandatory." Both students who correctly respond utilized the test-based strategy of process of elimination in order to select the key. In addition, the students who incorrectly eliminate the option seem to reason that the option does not sufficiently describe the similarities and differences, as evidenced by the following statement: "...I didn't pick [the option]. Although it says that the two courts differ, I don't think it really gets at what they're differing over." Another student stated that the option was true, although he felt like it was not complete. He stated, "Well, that's true. That's certainly true, I think. The *Black and Brown* court really does not discuss the disparity." However, he follows with "But does that really summarize [the similarities and differences regarding the reasoning]?"

Another reason that emerged in the protocols is that while the key notes that the *Black and Brown* court doesn't discuss the disparity of record cost amounts, several students noted that the court does indeed talk about this disparity, albeit briefly and in the

context of the *Meta* case. This is evident by the student's statement that, "…[T]he *Black and Brown* court does discuss this disparity…[I]t states that if the *Meta* court is followed, then no payments should be paid in general." These two issues that emerged in the protocols are evidence that the item needs major revisions. Additionally, options A and E do not have complete evidence to support that they are functioning as intended. However, this may be a consequence of the problems with the item key.

The TV2 item for Pair 7 does not seem to have any problems. The item better focuses on a true summary of the differences between the two cases rather than differentiating a case in the context of the other case. Ideal responses emerged for students correctly responding to the item. The distractors also seemed to function as intended. A minor suggestion for revision emerged in the analysis of the students who incorrectly responded to the option. The last portion of the key focuses on the benefit plan disclaimer in the cases, whose purpose is to let employees know what to expect and that they have no guaranteed benefits. The last portion of the key states that the circuit court reasoned that the disclaimer serves the purpose of putting employees "…'on notice that they have no guaranteed benefits' and that, consistent with ERISA, they 'know exactly' where they stand." The phrase "know exactly" comes from the *Hamilton* case, in a quote citing a different precedent case. However, some of the students who correctly responded to this option felt that this phrase seemed too strong and didn't feel that the circuit court would state that the participants would have such strong knowledge of the benefits. As one student noted, "I don't think they ever used language that would amount to 'know exactly where they stand' because in that case, they didn't know where they stood because documents were inconsistent." A minor suggestion is to change this

phrase to "know what to expect" which still maintains the intended meaning but eliminates a potentially construct-irrelevant source of error. Other than this minor change, the item seemed to function as intended.

While the TV1 item needed major revisions and the TV2 item needed minor revisions, the raters both felt that the items were eliciting the intended reasoning and thus were judged equivalent.

Pair 8

Pair 8 was judged to be nonequivalent by the two raters who suggested that the TV2 item needed to be revised. These two items focused on the issues of legality and constitutionality. The TV1 item asks students to summarize whether or not the record cost statute could be considered constitutional based on all three cases. The correct answer stated that none of the three cases presented directly challenge the constitutionality of the statute, although one of the cases mentions that the statute had been challenged on constitutional grounds in the past. The TV2 item asks students to summarize the view of the courts regarding the legality of the reservation of rights disclaimer under the ERISA statute. The correct answer states that,

> The *Hamilton* case in its reasoning implies that a reservation of rights disclaimer is legal under ERISA as long as it is published within a summary plan document and not merely communicated orally to employees; however, the *Bridell* and *Alexander* cases in their reasoning each imply that a reservation of rights disclaimer may be illegal, hence unenforceable, under ERISA if its meaning is demonstrably ambiguous.

These items were hypothesized to be quite difficult within the cross-case determinate item set, as they require synthesis across all three cases in light of a concept (legality/constitutionality) that students may or may not have considered during their initial reading of the cases. A total of five students in the think-aloud sample correctly responded to the TV1 item; a total of four students correctly responded to the TV2 item within this pair.

In the TV1 item, less synthesis than hypothesized was found to be necessary to answer the item. Students needed to merely scan the text of the cases to see if any of the cases directly challenged the constitutionality of the statute. Ideally, the item would have required students to synthesize or combine information from across all three cases in order to correctly respond. Because none of the cases challenges the constitutionality and the *Hamilton* case only discusses it indirectly, a "cross-case" analysis is almost forced in this item calling for only a quick scan of all three cases without requiring more advanced synthesis. This item needs to be revised to try to make it more of a cross-case item requiring synthesis. An ideal response was elicited from students for this item, in that students are able to select the key and eliminate the distractors for the reasons intended. The errors that emerged in this item were related to the construct including failure to differentiate the court's holding (or decision) from dicta (opinions of the court that are not considered law), not reading the option or case carefully, and illogical reasoning.

The TV2 item also needed some revisions regarding the options. Three of the four distractors contain statements that are truthful, yet are not complete, which may introduce another source of item difficulty. Because the TV1 item is not written in this manner, this could also pose a source of nonequivalence for the item pair. While the

distractors in the TV2 item are incorrect, because of the length of the item and the options, students have difficulty navigating through the options, eliminating incorrect choices and selecting the correct option. The distractors in TV1 are much more clear in terms of their correctness and incorrectness from the test-takers' perspective and do not contain portions that are true. This issue is particularly evident within Option A, which states,

> It is clear from *Hamilton* that employers are free to use, phrase, and locate their disclaimers in any manner they choose in order to deny the provision of welfare plan benefit plans that may be described elsewhere in their plan documents. At the same time, the other two cases, *Bridell* and *Alexander*, make it clear that the use of reservation of rights disclaimers is not illegal in principle under ERISA.

The second sentence within this option regarding *Bridell* and *Alexander* is correct. The first sentence is incorrect in that *Hamilton* did not go so far regarding the reservation of rights disclaimer. This option was very attractive to the students in the think-aloud sample. The item writers are encouraged to modify this option to make it less overlapping with the correct option. Other than this option, the distractors and the key functioned as intended eliciting ideal responses. The errors that students made are related to the construct and to personal shortcomings such as illogical reasoning and failure to fully read the item option, such as the student who stated as he read the key, "The *Alexander* decision clearly offers a direct challenge to such blah, blah, blah, blah…"

In this pair, both items focus on issues of legality and constitutionality. The test writers may want to focus only on legality to make these more similar in nature. In addition, in order to make the items more equivalent across the two test forms, the

options should function in a similar manner and require similar strategies for correctly responding. The options for TV2 need to be less overlapping. The options for TV1 need to require more synthesis across the cases rather than requiring a quick scan of the text. Therefore, both of these items are suggested to be revised. The conclusion for this pair is that they are nonequivalent with major revisions required for the TV1 item and minor revisions required for the TV2 item.

Single-Case Indeterminate Item Pairs

Fewer problems were identified with the single-case indeterminate item pairs. All items within this category were rated to be equivalent by both raters. Revisions were suggested for one item within Pair 9. Minor suggests were made for both items in Pair 11.

Pair 9

The items constituting Pair 9 ask the student to identify an ambiguity that is presented in the one of the precedent cases. Specifically, in the TV1 question, students are asked to identify an ambiguity that is present in the *Black and Brown* case. In this item, students need to recognize that a very relevant ambiguity in the *Black and Brown* case is that the court does not discuss whether the 20-day time limit for paying record court costs is a fixed time limit or if it is subject to the same considerations regarding substantial compliance as is the amount paid. This issue is important to the *Mackey* case because of the factual situation that Mr. Mackey paid only part of his record costs two days late. The TV2 question asks students to identify an unanswered question from the *Hamilton* case. Both questions are asked within the context of considering an argument or an appeal for their client in each respective test form. In the TV2 question, the student

needs to identify that the relevant ambiguity presented in *Hamilton* is the extent to which employers are held liable for ensuring that employees are cognizant of meanings in the reservation of rights clause. This question is very applicable to the *Bridell* case in that if employers are held liable for ensuring that employees understand the reservation of rights clause, Ms. Bridell's case might be construed within a different light.

With the shift from determinate to indeterminate items, this pair is hypothesized to be moderately difficult. While students will need to still rely on textual references, either utilizing memory or rereading, much of their response will be dependent on their reasoning strategies in determining which ambiguity would be most helpful in light of the advocatory role for their client. For the TV1 item within Pair 9, a total of four of the seven students provided a correct response. For the TV2 item, a total of three of the eight students correctly responded.

Both the TV1 and TV2 items within Pair 9 elicited responses consistent with the intent of the item writers. In addition, the strategies used to respond to each item were similar between forms. The primary problem that emerged in the analysis of the student responses concerned the distractors in the TV1 item. The TV1 item was quite lengthy, requiring students to read each option carefully and often more than once, in order to grasp the general idea of the option. However, upon close analysis of the options, choices A, C, and E are quite similar in nature. One student who correctly responded to the item was able to eliminate all three choices together using test-based strategies, rather than relying on understanding of the case or reasoning skills. She stated the following:

> The nice thing. I do love it when multiple-choice tests do this. A, C, and E all pretty much say the same thing…It would be almost impossible to pick A, C, or E and one over the other. So right off the bat, I can just eliminate them.

This same student was able to identify the correctness of the key for reasons listed in the justification, although she was able to eliminate other options for reasons unrelated to the construct. The item writers are encouraged to revisit the options for this item and to make A, C, and E less overlapping so that students are not able to eliminate them in one fell swoop.

An additional change that is necessary for this item is to drastically reduce the length. The item takes up an entire page and is very wordy. Some students lost sight of the task at hand because of the item length, which could place large demands on working memory. As one student stated, "I'm getting lost in all of this."

The three people who incorrectly responded to the item relied on a combination of both text-based and knowledge-based strategies. One student relied heavily on intuition, stating that, "I didn't even look at E that much but it seems to be the best." One student incorrectly selected the option because of his perception of how much burden would be placed on the opposing counsel, rather than considering the option in terms of the ambiguity. He stated when selecting his response, "Yeah, that makes a lot of sense because that could be annoying if you have to or if you have to prove two mitigating factors." Later on he considers this selection again because of a perception that it best matches the reasoning related to the holding in the case stating, "But I would pick C because it deals with the holding of the case which is what you're going to use with your argument." While these errors made by this student could be considered construct-

relevant, it is also possible that he simply lost sight of the task due to the lengthy nature of the item.

The TV2 item functioned more closely to its intended purpose. While there was not an exact, clear ideal student response, the students who correctly responded to the item provided indications that they were utilizing the reasoning and strategies that would be appropriate for the response. Options A and E did not have clear evidence to support the item writers' intent, although no apparent problems were identified with either option. The students who incorrectly responded to the item over-relied on memory or lost sight of the question in the stem. No major revisions were suggested for the TV2 item within Pair 9.

The consensus between the raters was that the two items within the pair functioned similarly, eliciting analogous strategies and reasoning skills, although significant revisions were necessary for the TV1 item. The item pair was judged to be equivalent.

Pair 10

Pair 10 was also judged to be equivalent by the raters. Both items asked the student to identify an "unknown" that was absent in their client's case but that could be factually helpful in the next step in arguing for the client (asking the court to reconsider its decision for *Mackey* and submitting an appeal for *Bridell*). For the TV1 item, students needed to recognize an important question that would provide a key strategy for submitting a motion to reconsider. The correct answer focused on whether the plaintiff's communication was unintentional or deliberate in misleading Mr. Mackey. Intentional deceit on the part of the plaintiff would likely be favorable to Mr. Mackey's case,

supporting a claim that Mr. Mackey's attempts to pay were both valid and honest. The correct answer for the TV2 item focuses on whether company records exist that would support an argument for employee confusion over the reservation of rights disclaimer and whether the company responded unclearly to these. This unknown would support that the employees were not on notice regarding the limitation of rights and that the company may have either intentionally withheld information or that they themselves were unclear as to the company policy. This unknown would go against the policy maintained by ERISA and would likely help Bridell's case.

In order to correctly respond to each item, students would need to evaluate each option in determining whether the "unknown" would be a helpful avenue to pursue in client's case. This item was hypothesized to be of moderate difficulty, requiring some textual references. However, the majority of the information involved with option selection should be determined by student reasoning about the helpfulness of the unknown in light of the task at hand. In the think-aloud sample, a total of three of seven students correctly responded to the TV1 item. For the TV2 item, a total of six of eight students correctly responded.

No problems were found for either item within the pair. For the TV1 item, an ideal response was clearly elicited from the students who correctly responded. Evidence was also obtained that the distractors were eliminated with the reasons and rationale that were intended. The students who incorrectly responded to the item made construct-relevant errors. One student failed to limit the reasoning to the question asked in the stem. Another student used faulty logic to select C, stating that "If the records show that there are 90 different cases where they lost an arbitration hearing and paid the costs 90

times, that could speak well on their behalf." This line of logic would be unlikely to help Mackey's defense.

The TV2 item also functioned as intended, with evidence supporting the reasoning of the selection of the key and the elimination of the distractors. A student who correctly responded to the item provided the following justification for her selection:

> That one would be helpful because it would show that people did not understand the SPD and if the company personnel responded evasively or unclearly, for us it means that they didn't know the answer to begin with. Or if they did, they were trying to avoid the employee's knowing that their benefits could be affected by that.

No suggested revisions were found necessary for this item. Students who incorrectly responded to the item also did so for construct-relevant reasons including a failure to consider certain information provided in the cases. Specifically, several students were attracted to options that focused on the notion of discrimination, particularly B and D, which students may first view as discriminatory towards Bridell. However, these students failed to notice that ERISA's policy does not proscribe discrimination, a fact that is provided within the text of the case. Under ERISA, companies are allowed to discriminate against individuals or classes of individuals regarding their benefits, provided that actions are not retaliatory against an individual. The students who erroneously selected these options do not seem to have grasped the important ERISA definition regarding discrimination. This error is related to the construct of legal case reading and reasoning.

These two items both functioned as intended, eliciting similar strategies from the students to respond. Thus, the raters agreed that the items were equivalent. No revisions were suggested for either item.

Pair 11

Pair 11 was the final single-case indeterminate item set. These two items asked the student to identify the information that would be most helpful in a new case for the client. On the surface, these items appear to be more dissimilar than some of the other item pairs, although both focus on the same ultimate goal of helping the client in a new case. The TV1 item asks students to identify an ambiguity within the *Mackey* case that would be relevant to the task of appealing the case to the Superior Court. The correct answer focuses on the attempts of Mr. Mackey to pay the record costs and why these attempts are not discussed by the *Mackey* court in terms of being an "honest effort" or "valid attempt." The TV2 question asks the students to identify the question that would most likely thwart the opponent's position in the new case. The option that is correct for this item focuses on an example that would support than Ribier was retaliating against Bridell for filing a large claim, an action that would be considered an impermissible discrimination under the ERISA policy. In order to correctly respond to both items, students would need to use a combination of text-based strategies (referencing the notion of an honest effort/valid attempt in TV1 and the ERISA policy on discrimination on TV2) and reasoning strategies of weighing the helpfulness of each option for the client's case. The raters noted that as the items moved from determinate to indeterminate, that it became more difficult to write items that were isomorphic from one test form to the other, given the specific details of the individual cases.

These two items were hypothesized to be moderately difficult because they require students themselves to speculate how helpful each option would be for the client's case. Fourteen students received the TV1 item. Of these, a total of nine students correctly responded. Five of the eight students who received the TV2 item correctly responded.

The strategies and reasoning utilized to respond to the items were fairly equivalent. In addition, ideal responses emerged within the verbal protocols demonstrating the intended reasoning for the selection of the key within both items. For example, one student who correctly responded to the TV1 item stated the following:

> I think that's going to be the one most relevant to our task because that can certainly be interpreted differently by a court. Or by a Superior Court here especially given the other case law where it seems like an honest attempt was made because he did pay some of the amount…It seems like they want to give the full picture, but if the Superior Court sees that and given some of the case law, I think that there's a reasonable effort."

The ideal response to the TV2 item can be demonstrated by the following student reasoning:

> Their case [the opponent's case] rests on the fact that there was no discrimination …The court says that the purpose of the change was to protect the plan and they made the changes because of financial losses. So any proof that the reduction of benefits was enacted to retaliate against her, that might be against the purpose of the act.

The primary differences elicited in the strategies of these items concerns the perspective requested in the stem. Students in the TV1 item consider which option would be most beneficial to the client's case. Students in the TV2 item consider which option would be most harmful to the opponent's case, and by inference, most helpful to the client's case. Even with the differences in wording in the stem regarding the perspective students are asked to take, the strategies and reasoning elicited appear to be similar.

Within the TV1 item, some minor issues in the distractors emerged. Clear evidence in the verbalizations was not found to support the item writers' intent. However, no problems that required substantial revisions were apparent. Minor wording changes were suggested for Option D in the TV1 item as the wording of "Why doesn't the *Mackey* court reprimand the plaintiff…" On close inspection, the raters felt that this was not an appropriate answer to the question posed in the stem regarding "…which one of the following questions presented by the *Mackey* opinion" given that what the *Mackey* court doesn't do could not be presented in the opinion. The distractors for the TV2 item functioned as intended with one exception regarding option C which stated, "Can Bridell provide evidence that she had not received reimbursements for the claims she filed with National American Life Insurance prior to July 1988." The item writers had intended the students to weigh its option in terms of it's helpfulness for the client's case. However, one student noted that, "They already said that she did receive reimbursements before the claim" and quoted the phrase from the text that "'She was not denied any compensation for expenses related to treatment for AIDS during 1987 when the plan paid up to $1,000,000 for AIDS patients, nor was she denied medical compensation for insurance claims.' So that doesn't work." This quote from the text was overlooked by the test

writers and thus resulted in the distractor functioning differently than intended, although in a manner considered to be minor by the raters.

The types of errors that students made for each item appeared to be related to the concept. Students who incorrectly responded to the items did so for reasons related to the construct including not understanding the legal definition of a commonplace term (i.e. "nonpayment") and misunderstanding a concept (discrimination) described within the content of the case. These same errors had previously been spotted in other item pairs.

While these items were not necessarily isomorphic, the raters did feel that the reasoning elicited was quite similar and thus an equivalent rating was assigned. Minor revisions were suggested for one distractor in each of the items for Pair 11.

Cross-Case Indeterminate Item Pairs

All pairs within this set were rated as equivalent with significant revisions deemed necessary for the TV2 item of Pair 12. In addition, minor revisions were necessary for the TV1 item in Pair 12, and both items in Pair 13, and the TV1 item in Pair 14.

Pair 12

The items within Pair 12 asked the students to consider aspects of opposing counsel's likely argument strategy on appeal, based on the precedent cases students were asked to read. In the TV1 item, students need to consider the threat that the *Black and Brown* situation would have on their case for Mr. Mackey. Opposing counsel would most likely compare the factual situation in *Mackey* to that in *Black and Brown.* In this controlling precedent case, the opposing counsel would want to show that Mackey, like the defendant in *Black and Brown*, had complied with all the requirements for appeal with the exception of the payment of the record costs in full and on time. Thus Mr.

Mackey should receive the same penalty of not being allowed an appeal. In order to correctly respond to the item, students need to recognize that the similarities between Mackey and that of the defendant in *Black and Brown* would be detrimental to the appeal of the *Mackey* case. In the TV2 item, students are also asked to consider the most relevant argument that opposing counsel would be likely to present. The correct answer focuses on the comparison of the Ms. Bridell's factual situation with that in *Hamilton.* *Hamilton* makes it clear that even individual discrimination is permissible under ERISA. Hence, any argument to build a theory of Bridell's case around discrimination would be futile.

These two items were hypothesized to be quite difficult as they require a shift in the mindset from considering strategies that would help the client to considering how opposing counsel would argue against the client. Although this cognitive shift is likely to be difficult, this is a skill necessary in order to be a competent lawyer (and a successful law student). In the think-aloud sample, of the 14 students who received the TV1 item, a total of 7 students correctly responded. For the TV2 item, only two of the eight students who received the item correctly responded.

Several ideal responses from students were clearly elicited in the think-aloud for the TV1 item. For example, as she was evaluating the correctness of the key, one student stated, "I think the opposing counsel comparing Mackey's situation with the defendant's situation in *Black and Brown* would be important because that is the current law…" Another student stated, "Yeah, I would say that because they're going to use *Black and Brown*. They're going to want to analogize Mackey's situation with the defendant's situation in *Black and Brown* because they want the *Black and Brown* decision." All of

the distractors were found to function as intended, with evidence that students were able to eliminate them due to the reasons described within the justification. Analysis of the incorrect responses yielded information that suggested some possible revisions to the item stem. All but one of the students who incorrectly responded selected the same option, which focused on one specific aspect of the *Black and Brown* case that was not relevant to the *Mackey* situation. The option states the following:

> Given that the *Black and* Brown court stressed that the appellant in that case had 'express notice' that record costs be paid within the 20 day limit, how likely is it that opposing counsel will argue that Mackey similarly had 'express notice that he had to pay his record costs?

On the surface, this option may seem correct, but a closer analysis of the cases would reveal that the issue in the option, "express notice" is not discussed within *Mackey* and is only a very minor aspect within *Black and Brown*. Students who incorrectly responded to this item focused on the fact that this could help opposing counsel's argument without considering the facts in the cases themselves, an error related to the construct being measured. However, the reasons why these same students rejected the key suggested that the item needs to be revised. Rather than focusing on the factual similarities between Mackey and the *Black and Brown* defendant, the students seemed to fixate on their perception that opposing counsel would be unable to make a successful analogy between the cases. For example, one student stated, "I don't think [the choice] is relevant because I don't see how they're going to bring the *Black and Brown* defendant and *Mackey* defendant together. It seems extremely different as far as the facts or the actual situation." Another student stated, "Well, in that case, in *Brown,* the appellant just

completely disregarded the notice and at least in *Mackey*. It can't be [the choice] because Mackey actually made significant attempts to get the money to Pepper." These students show that they are more focusing on the case from their perspective, as the advocate for Mr. Mackey rather than on how the opposing counsel would be able to portray the factual similarities between the two cases. While the raters did feel that this error was related to the construct being measured, a suggestion was made to revise the stem to clarify the question being asked, which may result in a decreased frequency of this error. Specifically, the stem would be clarified to state the context within which to consider opposing counsel's strategy, perhaps with the following wording: "Which one of the following questions about opposing counsel's approach to the *Mackey* decision's relation to precedent is important for you to consider as you prepare your case for Mackey?" In addition, the key should be revised to clarify the term "situation" to "factual situation," which may increase the likelihood for students to consider the similarities in the case rather than focusing on whether or not the opposing counsel would be able to successfully argue the similarities between the two cases. The conclusion on this item that while it did function as intended regarding the strategies elicited and having evidence that the key and distractors were functioning as intended, the analysis of the incorrect responses revealed a potential problem that could be mediated by slight revisions to the wording of the stem and the key.

Because only two students correctly responded to the TV2 item, there was unclear evidence for an ideal response. While the two students appear to be getting close to the intended justification, neither one verbalizes it completely. The distractors seemed to function as intended although there was unclear evidence for choice A. The primary

issue with the item may be the wording of the key. The key specifically states, "To what extent will opposing counsel attempt to show that the case by case justification put forward in *Hamilton* rightfully extends to Bridell's case." The wording of "case by case justification" was designed to lead students to the idea of discrimination. However, neither the students who correctly respond nor those who eliminate the option appear to be making this connection. The conclusion on this item is that there is not evidence for the key to function as intended, based on the analysis of both the students who correctly responded and those students who had eliminated the option. Thus, revisions to the key are suggested with the wording of "case by case justification" changed to something that explicitly includes the term "discrimination."

The raters agreed that the two items within this pair are equivalent in terms of the strategies and reasoning skills elicited. However, minor revisions are required for the stem in the TV1 item. More substantial revisions for the key in TV2 item are required which may change how the item functions.

Pair 13

The items within Pair 13 required students to compare the client's case to both precedent cases regarding the most relevant or important lines of questioning. The TV1 item required students to decide "which one of the following questions about the relationship between the *Mackey* decision and the other two Superior Court decisions…is most important for you to think about?" In order to correctly respond to the item, the student would need to read each option carefully and weigh its importance for the task of appealing Mr. Mackey's case. Students need to recognize once again that the time issue for record cost payment is important for Mackey's case. The correct answer states, "Can

the *Mackey* court, without qualification, properly conclude from the decisions in either *Meta* or *Black and Brown* that record costs must be paid in twenty days?" Students should note that the *Mackey* court does not quote the entire holding of *Black and Brown* and leaves out the caveat directing courts to examine to see if "a valid attempt to make…a *timely* and full payment, coupled with *substantial though incomplete compliance with the requirement*," has been made (italics added for emphasis). The *Mackey* court discusses the attempts by Mr. Mackey but never discusses why they would not be considered "honest" or "valid" and never discussed whether his actions could possibly demonstrate "substantial compliance." Therefore this question would be highly important for an advocate for Mackey to consider. The TV2 item within Pair 13 also functioned generally as expected. The question asks students to consider if the line of questioning related to Bridell's defeat at the district level would be best supported by the cases. The correct response to the item includes the question, "Did counsel for Bridell consider obtaining expert evidence concerning the visibility and understanding of the footnoted disclaimer and its relationship to the phrase in the main part of the SPD referring to 'lifetime medical benefits'?" The question focuses on the students' ability to recognize a strong comparison between the *Bridell* and the *Alexander* case regarding ambiguity of the reservation of rights clause. The *Alexander* case used experts in the form of empirical research about how disclaimers are understood. This comparison would show support for Bridell's advocate to seek out expert testimony as to the understandability of the disclaimer regarding its location in a footnote as compared to the main body of the text.

Both items were hypothesized to be quite difficult as they require a thorough understanding of all three cases and the ability to synthesize the information from all of them in order to determine which question would be most relevant as an advocate for the client. Significant textual references as well as reasoning and evaluating the helpfulness of each option would be required to correctly respond. On a superficial level, the item might seem to function slightly differently in that it does not reference the precedent cases directly as done in the TV1 item. While correctly responding to the item still requires the synthesis and understanding across the cases, the options do not explicitly mention the precedent cases, which may lead to differential difficulty. Rather, the reference is made within the item stem, not within the options as in the TV1 item. A total of four of seven students correctly responded to the TV1 item while three of eight students correctly responded to the TV2 item.

Evidence was found that students taking the TV1 item were able to provide reasoning which matched the item writers' justification. For example, one student states, "*Mackey* is just concluding that the 20 days is absolutely binding. It's basically, *Mackey* is concluding that the substantial compliance does not apply to the 20 days." The distractors functioned as intended with reasons for elimination matching the justification with the exception of B which was unclear. In general, the item functioned as intended with the students who incorrectly answered doing so for construct-relevant reasons. The primary errors made were not reading the item carefully or relying on "intuition" about an item rather than logic, such as the student who stated, "I just like the wording in E. It feels more important than the one in D." One minor suggestion for the item was to slightly change the stem. Several students who incorrectly responded to the item did not

seem to focus on the task of appealing the *Mackey* decision to the next higher court. Even though this was stated in the first sentence of the stem, some students seemed to have lost sight of this task. Therefore, an additional clause was suggested for the end of the stem, changing it to "Which one of the following questions…is most important for you to think about *as you prepare your case for Mr. Mackey*?" This minor change may help students to focus on the task at hand as they consider the correctness of the options.

An ideal student response for the TV2 item was easily identified in the verbal protocols, as evidenced by student quotes such as the following:

> The *Alexander* case talks about the ambiguity in the provisions and the *Air Jamaica* case has a very straightforward plan. So what a reasonable person would think when reading the plan would have to do with whether or not it's ambiguous. And it has to be written in a manner to be understood by the average plan participant. So expert evidence would help to support or refute the question of whether or not the average person would be able to read it.

This student demonstrates analogical reasoning in his rationale for selecting the key. All the distractors functioned as intended and the errors that emerged were construct-relevant including the use of illogical reasoning. For example, one student was convinced that expert testimony would not help determine what the average person could understand, as suggested by his statement: "Expert testimony referring to something that should be understood by an average plan participant so an expert's opinion wouldn't really be that helpful." The reasons that students incorrectly responded to the item appeared to be related to shortcomings in their ability to reason through the cases.

There was some evidence that students had some difficulty understanding the question being asked. One student repeatedly read the item stem indicating he was having difficulty understanding the question posed. He finally noted that, "So I guess they're trying to find textual support for Bridell's defeat." This evidence supports some suggested revision to the stem. The raters considered the phrase within the stem of "best supported" be changed to "might logically follow."

Based on the verbal protocols, the raters judged the items in this pair to be equivalent although minor changes were necessary to the stems of both items.

Pair 14

Pair 14 was also judged to be equivalent between the two raters. The questions focused on comparing the client's case to one other precedent case regarding the most relevant questioning lines on appeal. The TV1 item asked students to consider the line of questioning based on the *Mackey* use of *Black and Brown* that would be most relevant to the task of appealing. The correct response focuses on the use of *Mackey*'s reliance on *Black and Brown* regarding the definition of valid attempt and substantial compliance. The *Mackey* court only quoted a part of the holding from the *Black and Brown* case which left out a part that might be helpful for Mr. Mackey's case. Responding correctly to the question requires students to find the exact phrase in the *Black and Brown* case, determine what has been left out, and realizing the benefit to the client's situation. Other options were considered incorrect because they have no relevance to the *Mackey* case, are too narrow in scope and do not address the very important issue of time, or are less persuasive in helping Mr. Mackey's case. The TV2 item asks the students to consider the line of questioning that would present the strongest challenge to the *Bridell* court's

analogical reasoning with the *Hamilton* case. The correct option compares Ribier's reservation of rights disclaimer to the one examined in the *Hamilton* case. Specifically, the option asks if the position of the disclaimer in a footnote of the summary plan document renders that court's analogy to the disclaimer in the *Hamilton* case as suspect.

On the surface, these two items may not appear to be measuring the same thing. However, both ask the students to consider the line of questioning that would be most relevant to their task of appealing the client's case. The TV1 question asks for which line would be most relevant for the task of appeal. The TV2 question asks for which line would be most challenging to the reasoning presented by the *Bridell* court. One suggestion for the test developers is to use more consistent language between the two items, either by asking which one is most relevant or by asking which one best challenges the main court's reasoning. Both of these items were hypothesized to be quite difficult due to their cross-case indeterminate nature. A total of four students of seven correctly responded to the TV1 item; a total of three students of eight in the think-aloud correctly responded to the TV2 item.

An ideal student response was clearly elicited for the selection of the key in the TV1 item. One student went back to the *Black and Brown* case and found the quote listed in the option. He stated, "They seemed to skip the whole second half of the sentence where they make the exception for sufficient compliance or something like that." Another student stated that the court "shortcut a lot of the analysis. I would bring up the difference in language and try to distinguish because I think there's a clear distinction." Clear evidence for the functioning of the distractors was also obtained with the exception of option C. The justification describes this option as the second best

answer given that it does have some relevance to the *Mackey* case but is too narrow and does not address the important time issue. The reasons that student who correctly responded to the item eliminated this option were somewhat unclear or not well verbalized. This option was a draw for the students who incorrectly responded as all of them selected it. Because it was so attractive for students and because the students who correctly responded do not seem to have a clear rationale for eliminating it, this option may need to be reconsidered. Because the line of questioning that is more useful can be somewhat subjective, the option should be made "more incorrect" rather than having some elements of correctness. Other than this change to option C, no other suggestions are apparent for this item. The errors associated with students selection of C seem to be related to a lack of careful reading of the cases, a construct-relevant error.

A clear ideal response from a student was elicited for the TV2 item. One student stated the following after reading the option:

> I think that's substantial. You're reading a document. There's something in the text as opposed to a footnote. I think it's really easy to miss a footnote and if that's the disclaimer that it says, yeah, we can change your policy. I think that's important.

Another student stated the following:

> It's [the reservation of rights clause] not as noticeable to the employee as it would be in the *Hamilton* case. So to use the *Hamilton* case as precedent does seem suspect in light of the fact that it wasn't in the text itself.

All distractors functioned as intended with reasons for elimination matching the justifications written by the item writers. The reasons that students incorrectly responded

to the item were primarily construct-relevant although one student relied primarily on test-based strategies in making his decision, believing that option A was a "trick" question. She also searched for patterns in the order of the responses across the test items. Another reason that emerged for incorrectly responding to the item was an over-reliance on the specific textual features of the cases. One student rejected the key because he did not "remember the court talking about it at all." Another student stated, "[The case] talks about the disclaimer itself not necessarily the fact that it was in a footenote." It is true that the court did not talk about how the reservation of rights disclaimer appeared in the footnote. However, this student appears to have lost sight of the question stem which asks the examinee to consider what might best challenge the *Bridell* court's reasoning. It would be likely that challenges to the court's reasoning would NOT appear in the case itself. The raters agreed that this is a construct-relevant error in that students are failing to go beyond the cases themselves to a general reasoning strategy outside of the text.

That being said, the items still elicited similar strategies in that students need to consider what would be helpful or harmful for the client's case. They need to evaluate each option in light of the task for appeal and consider the argument that would most advance their position. The evidence does support that students are utilizing similar strategies to respond to the items. Ideal student responses for selecting the key and for eliminating supported that the items were functioning as intended with some exceptions described above. Therefore, the raters judged these two items to be equivalent.

Conclusion

A total of three item pairs were identified by the raters as being potentially nonequivalent. These item pairs were 2, 6, and 8.

Chapter 6

RESULTS OF EXPERT JUDGMENT, STATISTICAL ANALYSES, AND METHOD COMPARISONS

The purpose of this chapter is to provide the results of the expert judgment and the statistical DIF analyses. In addition, the results of the comparisons among the three methods (expert, statistical, and think-aloud) will be described.

Results of Analysis of Expert Data

Table 11 provides the data on the average and standard deviation for each of theoretically equivalent item pairs. The higher the average rating (on the scale of 1-10), the less similar experts perceived the items in the pair to be. The lower the average rating, the more similar the experts perceived the items to be. Three pairs had average expert ratings that were higher than 5.5 and thus were perceived by the experts to contain items that were least similar. These items, ranked from highest to lowest mean are Pairs 13, 11, and 14. It is important to note that for several of the item pairs, the standard deviation among expert ratings is quite large and thus suggests a lack of agreement regarding the similarity of the items. The most similar item pairs were Pairs 3, 5, and 10.

When asked what features of the items influenced their ratings, most mentioned something similar to the wording of the items. For example, one law professional listed "similarity of language." Another expert stated, "Wordiness. Was the question asked directly or somewhat lost in the wording." Two individuals focused on the specific question asked in the stem. For example, one individual stated, "As I went through, I began to focus first on the…end of the question and I'd skim through the rest."

Table 11: Average expert ratings (n=7, scale of 1-10) and standard deviation for item pairs

Item Pair	Average Rating	Standard Deviation
1	3.86	1.46
2	5.14	3.44
3	1.14	0.38
4	4.86	2.91
5	2.00	1.15
6	4.86	3.29
7	3.14	1.68
8	4.14	2.48
9	3.14	1.07
10	1.86	0.69
11	5.86	3.13
12	2.71	2.14
13	7.86	1.21
14	5.71	2.29

In addition to surface features of the items, experts also considered aspects related to the construct. One expert wrote that she considered whether the item concerned "questions about facts versus questions about law" and "questions about reliance on precedent." Similarly, another expert stated, that he looked for "Whether the cue of the question asked for the same thing – e.g. reasoning, issue, question." He considered the

nature of the question first and then considered "the specificity of the question" and "the legal context of the question – appeal vs. motion to reconsider."

When asked whether they could detect the different types of items that were in the test, two wrote that they didn't understand what was meant by "types of items." Three said yes and listed the following comments:

- "Yes, if by this you mean that some questions asked for reasoning, some for the evaluation of questions, some for synthesis of cases, etc."
- "Yes. Some questions focused on the court's reasoning while others focused more on issue-spotting. Also, some questions were reflective while others required imagining new applications of law."
- "Yes. The questions illustrated very practical differences regarding purpose. Was the purpose to appeal? Did the lawyer want to rebut an argument?"

Finally, the experts were asked whether they felt that the type of reasoning necessary to answer the items is taught in law school. All the experts stated that there was some similarity between the test and the curriculum. Some of their comments are listed here:

- "Yes – much of the close case reading requires a good grasp of case synthesis."
- "[Some of the items] struck me as fundamentally different from legal education. The rest are similar in varying degrees."
- "Yes. Issue spotting, analysis of reasoning, and application of precedent are all fundamentals of legal education."

- "Yes. It's actually more useful than what's taught in law school. The questions seemed to have a purpose. Sometimes we do not focus our students on that and instead teach concepts abstractly. I liked that it was very concrete."

Results of Statistical Analyses

The factor analysis conducting using TestFact supported that the test forms were not unidimensional, with a chi-square of model fit for a one-factor model equal to 3407.67 with 91 degrees of freedom and $p<0.000$. Table 12 displays the eigenvalue associated with each component in the factor analysis. Figure 2 displays the scree plot, again supporting that the data across both test forms are not unidimensional. Cronbach's alpha was calculated to be 0.3636 for the TV1 scale. The standard error of measurement was calculated to be 1.66. For the TV2 scale, Cronbach's alpha was calculated to be 0.4634 with a standard error of measurement to be 1.68. The relatively low values of Cronbach's alpha provided additional support for a lack of unidimensionality.

The average TV1 score equaled 8.05 (57.5%) with a standard deviation of 2.08 (14.8%). The average score on TV2 equaled 7.38 or 52.7% correct with a standard deviation of 2.30 (16.4%). Figure 3 displays a histogram for the TV1 scores; figure 4 displays a histogram for the TV2 scores.

Table 12: Eigenvalues from factor analysis on combined data from both test forms

Component	Eigenvalue	Component	Eigenvalue
1	3.350	14	0.918
2	2.811	15	0.803
3	2.453	16	0.708
4	2.051	17	0.615
5	1.913	18	0.569
6	1.781	19	0.485
7	1.546	20	0.401
8	1.502	21	0.375
9	1.459	22	0.224
10	1.216	23	0.165
11	1.172	24	0.112
12	1.106	25	0.009
13	1.029		

Figure 2: Scree plot from factor analysis on combined data from both test forms

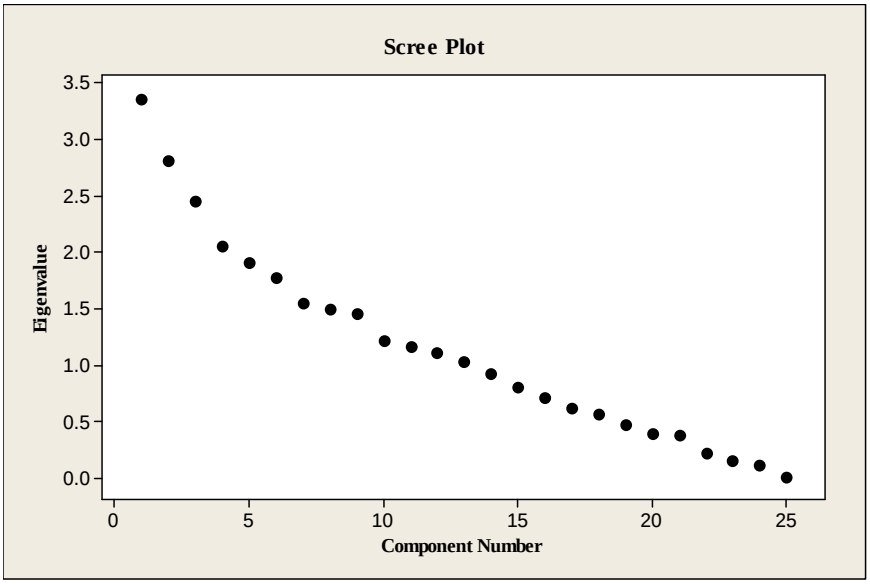


Figure 3: Histogram of scores from TV1

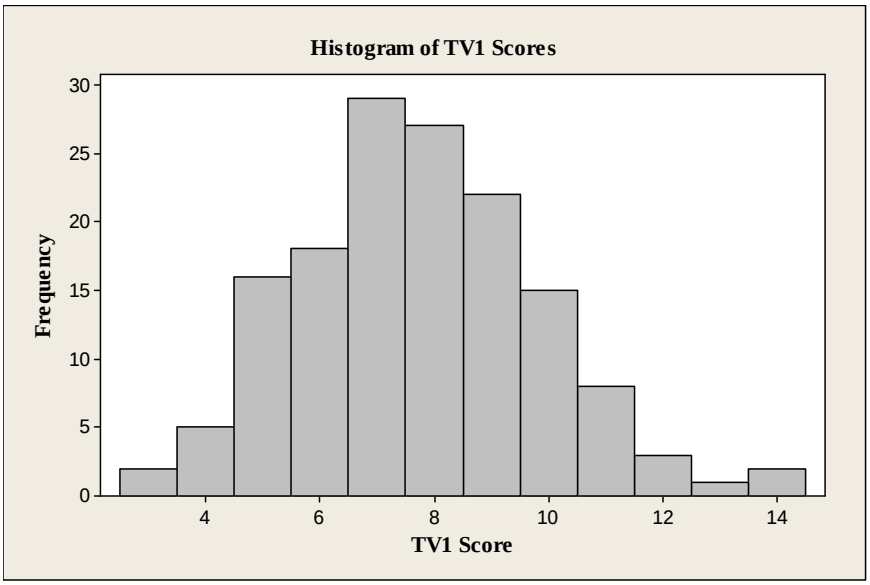
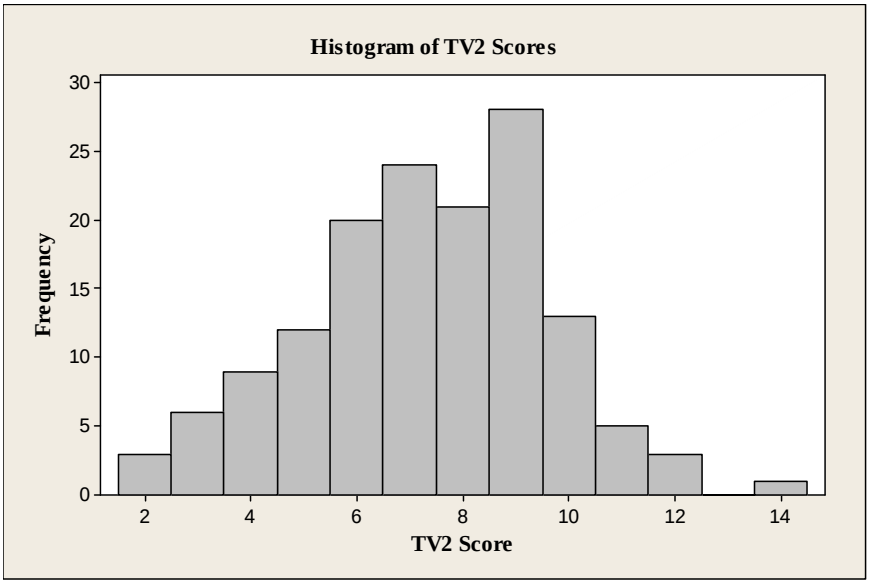


Figure 4: Histogram of scores from TV2



The results of the classical item analysis appear in Appendix H. For TV1, the easiest items occurred within Pairs 1 and 3. The most difficult items occurred within Pairs 6 and 12. For TV2, the easiest item occurred within Pair 3. The most difficulty items occurred within Pair 12 and 14. Across both test forms, the item discrimination indices were quite low.

Figure 5 displays the delta plot which graphically compares the difficulty levels of the items, in the form of the transformed difficulty index. Higher delta values indicate more difficult items. Items which fall outside of the parallel lines, thus having differences in delta values greater than 1.5, were Pairs 1, 4, 7, 8, 11, and 14.

Figure 5: Delta plot for TV1 by TV2

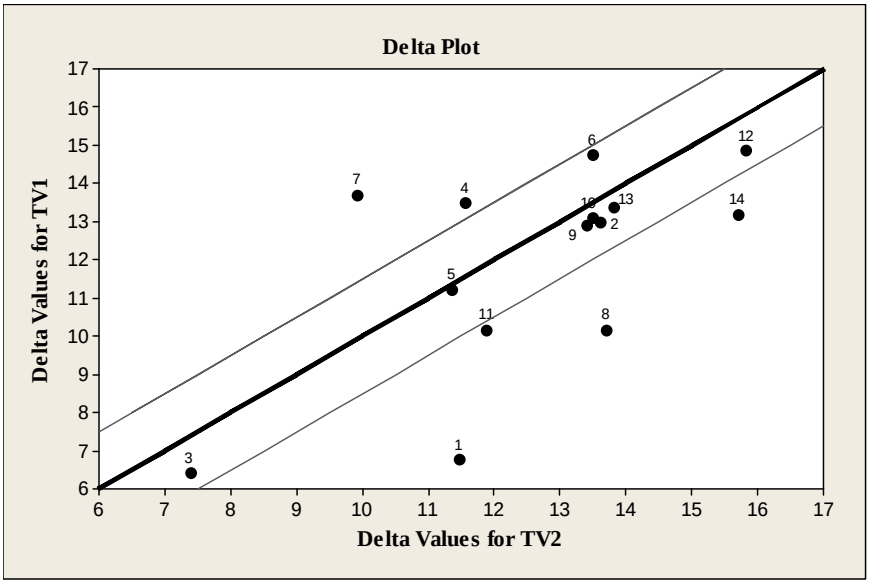


Table 13 displays the indices of DIF size including the signed proportion difference index (SPD-X) and the Mantel-Haenszel log odds ratio. Pairs which were calculated to have an absolute values of the SPD-X higher than 0.10 included Pairs 1, 4, 6, 7, 8, 11, and 14. The highest values of the absolute value Mantel-Haenszel log odds, ratio were Pairs 1, 7, 8, and 14, each of which had an absolute value greater than or equal to one. Of these three pairs, Pairs 1, 8, and 14 were all found to be easier for the TV1 item. Only Pair 7 was found to be easier for the TV2 item.

Table 14 displays the chi-square and corresponding *p*-value for each pair for the McNemar test across all test score levels. Because of the low expected cell frequencies for Pair 3, the McNemar test could not be performed and instead a binomial test for the difference in proportion correct was obtained. The pairs for which the null hypothesis was rejected are Pairs 1, 4, 7, 8, and 11.

Table 13: Measures of DIF size for each item pair

Pair Number	Signed Proportion Difference Index (SPD-X)	Mantel-Haenszel Log Odds Ratio
1	0.245	2.18
2	0.025	0.255
3	0.024	0.497
4	-0.231	-0.825
5	-0.039	0.069
6	-0.207	-0.518
7	-0.405	-1.643
8	0.261	1.517
9	-0.002	0.296
10	-0.077	0.202
11	0.134	0.796
12	0.044	0.457
13	-0.046	0.166
14	0.167	1.122

Table 14: McNemar Test across all score levels

Pair Number	Chi-Square	*p*-value
1	34.241	0.000
2	0.877	0.349
3*	-	0.424
4	11.042	0.001
5	0.017	0.897
6	3.413	0.065
7	34.256	0.000
8	29.280	0.000
9	1.125	0.289
10	0.444	0.505
11	7.779	0.005
12	2.526	0.112
13	0.368	0.544
14	17.016	0.000

* Because the expected cell frequencies were less than 5 for one cell in the 2x2 contingency table for Pair 3, the chi-square across all score levels could not be calculated. Thus, the p-value associated with this pair is from the binomial distribution instead.

Table 15 displays the results of the McNemar test when the data was collapsed into two score intervals. Because of the low expected cell frequencies, the chi-square for the low score level could not be calculated for Pairs 3 and Pair 12. The chi-square for the high score level could not be obtained for Pairs 1, 3, 5, 11, and 12. Because of this problem, this modification of the McNemar test was only performed for the nine remaining pairs. Of the remaining pairs, potential problems were found for Pairs 4, 7, 8, and 14.

Based on the above findings, Table 16 displays the level of agreement across these various methods of detecting DIF. Pairs 7 and 8 were identified by all methods as displaying DIF. In addition, Pair 1 was identified by all four methods except for the McNemar test with the two score intervals, which was unable to be performed. Pair 14 was identified by four methods with the exception of the McNemar overall test. Because of these patterns of agreement, these four items pairs, 1, 7, 8, and 14 were flagged as potentially containing DIF. Other pairs that are potentially problematic, but perhaps less severely, include Pair 4, which was flagged by all except the Mantel-Haenszel log odds ratio and Pair 11 which was also flagged by all with the exception of the Mantel-Haenszel log odds ratio and the McNemar collapsed test which could not be performed.

Table 15: McNemar Test with two score intervals

Pair Number	**Chi-Square (Low)**	**Chi-Square (High)**	**Summed Chi-Square**	***p*-value (df=2)**
1	19.314	-	-	-
2	0.108	3.361	3.469	0.176
3	-	-	-	-
4	5.921	4.364	10.285	0.006
5	0.878	-	-	-
6	1.029	2.025	3.054	0.217
7	14.205	19.184	33.389	0.000
8	20.021	8.500	28.521	0.000
9	0.625	0.281	0.906	0.636
10	5.625	1.561	7.186	0.028
11	4.356	-	-	-
12	-	3.559	-	-
13	1.730	0.129	1.859	0.395
14	8.500	7.500	16.000	0.000

Table 16: Agreement among DIF detection methods

Pair	Delta-plot	SPD-X	M-H log odds ratio	McNemar overall test	McNemar collapsed
1	Yes	Yes	Yes	Yes	-
2	No	No	No	No	No
3	No	No	No	-	-
4	Yes	Yes	No	Yes	Yes
5	No	No	No	No	-
6	No	Yes	No	No	No
7	Yes	Yes	Yes	Yes	Yes
8	Yes	Yes	Yes	Yes	Yes
9	No	No	No	No	No
10	No	No	No	No	No
11	Yes	Yes	No	Yes	-
12	No	No	No	No	-
13	No	No	No	No	No
14	Yes	Yes	Yes	No	Yes

An investigation of the empirical ICCs (in Appendix I) and the empirical ICCs collapsed by score level (Appendix J) provides some insight into what is happening within these item pairs that are flagged as being potentially problematic. The ICCs, while not specifically used as criteria for detecting nonequivalence, were rather used as a way to describe possible trends in the item pairs across the levels of the total score. For Pair

1, the item in TV1 is quite easy across all levels of the total score. The TV2 item within this pair functions closer to what would be expected. Students with lower total test scores are less likely to correctly respond to the item than those at the higher score levels. A test developer would likely suggest that the TV1 item needs attention as it would not discriminate between those students who have the knowledge or skills and those who are lacking in this area.

For Pair 7, the TV2 item appears to be easier across all levels of the ability spectrum. In addition, there does seem to be potential nonuniform DIF, as the lines in the empirical ICC with collapsed score levels are not parallel.

Pair 8 also displays an unusual graphical pattern in the empirical ICCs with collapsed score levels. Students at the highest test score levels tend to perform similarly on the TV1 and TV2 item. However, at all other test score levels, the TV1 item is much easier than the TV2 item. The TV1 item does not appear to discriminate as well within the middle score levels.

The curves for Pair 14 appear to be parallel, with the item on TV1 consistently easier than the item on TV2 at all levels of the ability spectrum. This suggests uniform DIF for this item pair.

While Pair 4 was not identified by all methods, it could potentially be problematic. A closer investigation at its ICC suggests that it could potentially contain nonuniform DIF. The items within the pair tend to function similarly at extreme ends of the total score scale. However, at the middle score levels, the TV2 item seems to be much easier. These items may need some further investigation. Pair 11 also was flagged by several of the methods as being potentially problematic. A closer look at the empirical

ICCs suggests that the TV2 item is more difficult in general. However, a closer examination suggests that the items tend to function more similarly at the highest score levels. The TV2 item has a relatively flat curve although there is a general trend towards an increase in proportion correct as the test score level increases. While the TV1 item is also relatively flat, there seems to be a sharper division between the two lower score level groups and the two higher score level groups.

As mentioned above, the McNemar test could not be performed on Pair 3. A glance at the empirical ICC for this pair gives indication on why this would happen. Neither item in TV1 nor TV2 is well discriminating across all score levels. The items, while they function similarly, do not add to the discriminating power of the test. While they do not display differences, a test developer would likely examine these closely to see if they could be revised to provide better discrimination across the score levels.

Based on the empirical ICCs, test developers would also likely be concerned with the pattern of data demonstrated within Pair 12, particularly the TV2 item which seems to increase in difficulty as the total test score increases. While both items in the pair are difficult across the ability range, the curve demonstrated by the TV2 item is not ideal.

The graphical and statistical indices utilized in this study converge with a relatively clear identification of several item pairs that function differently across the forms. While there are several other pairs that display non-ideal patterns of responses, the focus of this research paper is on those that display DIF. Therefore, the items flagged to be potentially nonequivalent are Pairs 1, 7, 8, and 14.

Comparison of Methods

Hypothesis 1

Appendix K displays a chart comparing the items flagged across all three methods of item comparison. The first hypothesis of this study concerned the agreement between the verbal protocols and expert ratings. The kappa coefficient of agreement between these two methods was calculated to be -0.273, which is less than what would be expected by chance. The 95% confidence interval ranges from -0.483 to -0.063. Table 17 provides a 2x2 contingency table of the agreement between the expert and think-aloud methods.

Table 17: Agreement between expert judgment and verbal protocols

		Verbal Protocols	
		Equivalent	Nonequivalent
Expert Judgment	Equivalent	8	3
	Nonequivalent	3	0

The two methods did not identify any of the same item pairs as being nonequivalent. Expert judgment flagged Pairs 11, 13, and 14, which were all judged equivalent by the think-aloud protocol, although minor revisions were suggested for at least one item within each of these pairs. The think-aloud method flagged Pairs 2, 6 and 8, which were flagged as equivalent by the expert method by using the criteria of the mean greater than or equal to 5.5. However, it is important to note that these two pairs also had expert rating means that were quite high with Pair 2 equaling 5.14, Pair 6 equaling 4.86 and Pair 8 as 4.14.

Given the findings of the kappa coefficient, the first hypothesis, that the pairs flagged by the expert ratings will be different than those flagged by the verbal protocols, is supported.

Hypothesis 2

The second hypothesis of this study concerned the agreement between the statistical analyses and the verbal protocols. The kappa coefficient between these two methods was calculated to be 0.054 with observed agreement approximately what would be expected by chance alone. The 95% confidence interval for the kappa coefficient ranges from -0.477 to 0.585. Table 18 displays a 2x2 contingency table of the agreement between the statistical and think-aloud methods.

Table 18: Agreement between statistical DIF and verbal protocols

		Verbal Protocols	
		Equivalent	Nonequivalent
Statistical	Equivalent	8	2
	Nonequivalent	3	1

Both the verbal protocols and statistical methods flagged Pair 8. The statistical methods also flagged Pairs 1, 7, and 14. The think-aloud methods flagged 2 and 6, which were not flagged by the statistical DIF methods.

Once again, given the findings of the kappa coefficient, the second hypothesis, that the pairs flagged by the statistical DIF procedures will be different than those flagged by the verbal protocols, is supported.

Hypothesis 3

The third hypothesis of this study concerned the agreement between the statistical DIF and the expert judgment methods. Table 19 displays the agreement between these two methods. The distribution in this table is identical to that displayed for the agreement between the statistical and think-aloud methods. Once again, the kappa coefficient was calculated to be 0.054 with observed agreement approximately what would be expected by chance. The 95% confidence interval for the kappa coefficient ranges from -0.477 to 0.585. The methods agreed that Pair 14 is nonequivalent.

Table 19: Agreement between statistical DIF and expert judgment

		Expert	
		Equivalent	Nonequivalent
Statistical	Equivalent	8	2
	Nonequivalent	3	1

Given the findings of the Kappa coefficient, the third hypothesis, that the pairs flagged by the statistical DIF procedures will be different than those flagged by expert judgment, is supported.

Comparison of Methods by Item Type

Hypothesis 4

The fourth hypothesis stated that the think-aloud procedure would identify more of the cross-case items as being nonequivalent as compared to the other methods. Table 20 displays the 2x3 contingency table for method by item type on the case dimension. The results of the Fisher exact test state that if the null hypothesis were true, the exact

probability of finding a positive association between X and Y as large as the one observed would be 1.000. Therefore, there does not seem to be a clear relationship between the method of identifying nonequivalence and item type on the case dimension. Therefore, the fourth hypothesis is not supported by the results of this study.

Table 20: 2x3 contingency table for method by item type on case dimension

	Think-aloud	Expert	Statistical DIF
Single-case items	1	1	1
Cross-case items	2	2	3

Hypothesis 5

The fifth hypothesis of the study stated that the think-aloud procedure would identify more indeterminate items as nonequivalent. Table 21 displays the 2x3 contingency table for the method of detecting item nonequivalence by the item type on the determinate/indeterminate dimension. The results of the Fisher exact test state that if the null hypothesis were true, the exact probability of finding a positive association between X and Y as large as the one observed would be 0.071. In other words, there is approximately a 7% probability that a more extreme pattern than the one observed would be possible given chance alone. Based on the contingency table, there appears to be a pattern that the think-aloud and statistical methods are more likely to flag determinate items within the sample data. The experts are more likely to flag indeterminate items.

Table 21: 2x3 contingency table for method by item type on determinate/indeterminate dimension

	Think-aloud	Expert	Statistical DIF
Determinate items	3	0	3
Indeterminate items	0	3	1

Conclusion

Based on the above results, the first three hypotheses of this research study are supported. Specifically, the item pairs identified through the verbal protocols as being nonequivalent across the test forms were different from those identified from both expert judgment and statistical DIF methods. In addition, the item pairs identified through expert judgment were different from those identified through statistical DIF methods.

The fourth hypothesis, which compares the method of identifying nonequivalence based on the case dimension, is also not supported. Specifically, the Fisher exact test does not support that the think-aloud procedure is more likely to flag cross-case items as being nonequivalent as compared to other methods. The fifth hypothesis, which theorized that the think-aloud method would be more likely to identify more of the indeterminate items as being nonequivalent, is also not supported. While the Fisher exact test did support that the observed frequency distribution of item types by method is rather extreme, the think-aloud method seems more likely to flag determinate items as nonequivalent while the expert judgment method seems to flag more indeterminate items as nonequivalent.

Chapter 7

DISCUSSION

Previous literature has supported that information from response process data can be helpful in the test development process (i.e. Haladyna, 2004; Leighton & Gokiert, 2005a, Leighton, 2004). Think-aloud protocols can be used to identify items that contain ambiguities, are interpreted in a different manner than intended, and elicit different cognitive strategies than anticipated. Response process data adds information about test items that are not obtained through traditional test development procedures. This process is vital in considering the validity argument for an instrument.

Until this study, response process information had not been used in the development of multiple test forms, which are frequently utilized in practice but are not often studied, with the exception of the application of advanced statistical and psychometric analyses such as equating. This study supports that response process information can help identify items between test forms that are not considered construct equivalent and to spot potential item problems suggesting possible revisions.

This research project set out to test five hypotheses. The first three of these hypotheses compared the results of three methods of identifying item nonequivalence across forms: think-aloud, expert judgment, and statistical DIF. Specifically, Hypotheses 1 stated that the think-aloud methods would identify a different set of items as nonequivalent than would expert judgment. Hypotheses 2 stated that the think-aloud methods would identify a different set of items as nonequivalent than statistical DIF methods. Hypotheses 3 stated that the expert judgment would identify a different set of

items than would the statistical DIF methods. All three of these hypotheses were confirmed through this study. The think-aloud data may identify potential issues in items related to nonequivalence that do not necessarily match with the results of statistical or expert judgment methods.

The second set of two hypotheses tested in this study concerned whether the think-aloud method would be more likely to identify certain types of items as being nonequivalent. Specifically, the think-aloud method was hypothesized to identify more cross-case items as nonequivalent and more indeterminate items as nonequivalent as compared to the other two methods. These two hypotheses were also not supported by the results of this study.

The findings and limitations for each method used in this study are discussed below followed by a discussion of the implications of the study's findings for test developers.

Discussion of Think-Aloud Protocols

As mentioned above, the think-aloud protocols yielded valuable information that was not obtained by the other methods. The protocols demonstrated the complexity of the test items, showing how and why students would consider each option and ultimately make their selection. It is important to note that the sample who participated in this portion of the study was very adept at verbalizing their thoughts. The multiple-choice items used in this study are extremely intricate, requiring the examinee to possess close reading skills and considerable reasoning ability (closely associated with thinking like a lawyer). These same features made the items ideal for the task of verbalization thought processes.

The cognitive processes identified in the think-aloud protocols by the raters did not conform to traditional taxonomies, such as that developed by Bloom (1984). While Bloom's taxonomy might be appropriate to use in the analysis of think-aloud for some instruments, the test of legal case reading and reasoning was designed intentionally to measure higher-order cognitive skills. Using Bloom's taxonomy to identify cognitive processes utilized in responding to items would not likely have yielded the kind of information that would be deemed helpful for test development.

The think-aloud protocols helped to define instances of individual items functioning in a different manner than intended, due to construct irrelevant influences on student response such as item wording, ambiguities, item-writer oversight, reviewer oversight, and failures of the item to conform to item writing guidelines.

The verbal protocols revealed interesting examples where a seemingly minor phrase or word choice in one option could unduly influence the way the item functioned in a manner not anticipated. For example, the TV1 item in Pair 5 used a simple phrase of "in what ways" which seemed relatively innocuous to the test developers and to the panel of experts who had reviewed the instrument. However, this simple phrase was interpreted by students in such a manner that lead them down a different path than intended in their reasoning skills. Such a seemingly minor flaw in the item may not seem to be a huge problem from the test developers' perspective. However, if the test developer only relied on statistical information or expert review for this particular item, it is possible that this flaw would be overlooked.

Another instance where minor wording problems posed a potential problem is the TV1 item in Pair 6, which demonstrated an interesting example of the complexity of

multiple-choice items. The important part of the key to the item was placed within the middle of the phrase and was often skipped over by the students. It seemed as if students were primed by the similarities of the key to the other options in the set that they failed to read or grasp the significance of the change to the wording in the key relating to the time issue in the *Mackey* case. While this error may seem to be related to the construct of interest in that students are not closely reading the items themselves, it may also be considered a construct-irrelevant influence on students' response as the important part of the option is buried within the key. The item writers may want to consider whether these findings should result in a revision of the item. Again, the expert judgment and the DIF methods did not flag this pair to be nonequivalent. While all item discrimination levels in both test forms were quite low, the values for both items in Pair 6 were among the highest.

Another finding that emerged in the think-aloud protocols included oversights by the item writers regarding the functioning of certain distractors. Students sometimes eliminated distractors for different reasons than listed in the justification, as exemplified in the TV2 item in Pair 12. While the item writers had expected students to weigh this option in light of its usefulness for the advocacy task, an unexpected response by a student revealed a different reason that the distractor should be rejected in that the question posed had already been determined within the text of the case itself. While this example seemed relatively minor within the functioning of the item itself, it elucidates a benefit of the think-aloud protocols in identifying instances related to possible item writer oversight. Once again, the expert and statistical methods deemed the items within the pair as equivalent. The item discrimination levels were quite low for the items in the

pair, particularly for the TV2 item which was a negative value. However, as mentioned above, the discrimination levels for most items were quite low.

Some of the problems that emerged in the think-aloud may have been avoided if multiple-choice test writing guidelines were more carefully followed (Haladyna, Downing, & Rodriguez, 2002). For example, within several items including the TV1 item in Pair 9, the options overlapped to some extent, causing students with testwise abilities to be able to eliminate options for reasons not related to the targeted construct. The multiple-choice item writing guidelines proposed by Haladyna, Downing, and Rodriguez state that options should be independent as much as possible to reduce overlap. This item flaw became apparent by a student being able to eliminate three of the five options without even considering the correctness of each. Another guideline that was violated is keeping the content of options homogeneous, as shown in the TV2 item for Pair 1. In this item, a student incorrectly selected his or her response because he noticed that one option stood out from the others and followed a different pattern. While this student certainly employed poor test-taking strategies in responding, this item does violate the multiple-choice item writing guideline and has introduced a potential source of construct-irrelevant error. A final violation of the guidelines that emerged in the think-aloud protocols concerned the length of some items. One of the items was very lengthy and the protocols clearly showed a fatigue effect with students verbalizing feelings of being lost in the verbiage. While often times very complex multiple-choice items do not follow the item writing guidelines, in some cases, failing to follow certain rules may introduce sources of construct-irrelevant error. The think-aloud protocols helped to

identify these instances where students' utilization of test-based strategies interfered with either the item functioning or with the students' performance.

The initial stage of the think-aloud data analysis concerned this identification of potential item flaws. These findings will result in revisions to many of the items on each form of the test. While this information is important, the ultimate purpose of this research was to examine the feasibility and benefit of using think-aloud information in the examination of construct equivalence of different forms. The task of identifying nonequivalence in some ways was hindered by the revelation of item flaws. Many of the items, while they were identified as equivalent when compared to their counterparts in the alternate form, contained potential flaws and had suggestions for revision. These flaws likely lead to random errors that would affect both the think-aloud and DIF methods in the approximately manner, thus not changing the results of the study. However, at times during the item analysis, it seemed that the task of identifying construct equivalence was premature and that the task of identifying item flaws within individual forms should have preceded the form comparison. In test development, perhaps the collection and analysis of the think-aloud data should be performed in an iterative fashion, first collecting and analyzing data for each individual form. From this information, the items could be revised to minimize the construct-irrelevant sources of error. Then additional think-aloud data could be collected for the task of comparing the two forms. Likely, additional sources of construct-irrelevant error would emerge, but hopefully the major sources would be eliminated. Because of the high frequency of suggested revisions to items, whether or not an item needed to be revised was not an issue that was generally considered when making the determination of equivalence versus

nonequivalence. Rather, the strategy for identifying nonequivalence in item pairs depended primarily on whether the strategies and reasoning elicited were similar. However, the problems identified through the think-aloud strategies may likely influence the statistical results for the classical item analysis and the DIF analyses, which could influence the comparison of results across the three methods in this study. A future research study on using response process information in construct equivalence studies might try using this iterative approach to collecting and analyzing the think-aloud data.

The analysis of the think-aloud protocols identified three item pairs as nonequivalent. What were the overall reasons that nonequivalence was assigned? For Pair 2, the think-aloud protocols suggested that students responding to one item were answering a different question than the students who received the other item. On the surface, the item stems appeared to be rather similar. However, the options themselves functioned in a different manner and elicited different reasoning strategies from the students. Within Pair 6, the two items also elicited different strategies and reasoning. The TV1 item required relatively unsophisticated textual references of comparing one case against the other. The TV2 item required students to also compare the similarities of two cases, but with the consideration of the usefulness to the client's case. For Pair 8, a close inspection of what students had to actually do in order to respond to the item revealed that the items were not measuring the same thing. The TV1 item did not require the students to do much more than a cursory examination of the text whereas the TV2 item required a greater synthesis of the cases at a higher level. All of these three item pairs either asked different questions of the students or required different processes in order to correctly respond.

One might argue that if the items were written to be more isomorphic in nature that the problems in construct equivalence for these three item pairs could have been avoided. Indeed, this may be true for Pair 6, where the term "useful" actually appears in the stem of TV2 but not in the TV1 item and perhaps the differences in strategies for responding could easily have been predicted. However, the item writers noted that given the differences in the cases themselves for each test form, it was not always feasible to write completely isomorphic items. In addition, isomorphism was not necessarily a requirement in order to have two items elicit similar strategies, as evidenced by the analysis of the indeterminate item pairs.

Against the hypotheses of this study, the think-aloud data was not more likely to identify items that were considered cross-case or indeterminate. The results of the Fisher exact test suggested that the case dimension did not influence the likelihood of the methods to flag the items as nonequivalent. However, the Fisher exact test did suggest that the determinate/indeterminate dimension had some impact in identifying nonequivalence. Contrary to the hypothesis posed, the think-aloud methodology seemed to be more likely to flag determinate item pairs while the expert judgment seemed to flag indeterminate item pairs. In fact, of the three pairs flagged as nonequivalent, all three were determinate items. One of these pairs was single-case and the other pairs were cross-case. The original hypothesis is that the think-aloud procedure would be more sensitive to the differences in more complex items, those that are indeterminate and those that are cross-case. It is not entirely clear why the think-aloud method did not detect more complex items as being nonequivalent. Perhaps the difference occurred because the raters were less focused on the surface features of the items and rather were interested in

the cognitive processes that emerged through students' verbalizations. When analyzing the verbalizations in depth, while the surface features of items may differ, the advanced indeterminate item types tended to elicit similar strategies and reasoning. However, alternate reasons are also possible. For example, one possible reason is that the raters of the transcripts could have possibly been more lenient in their analyses of the indeterminate items given their less isomorphic nature.

Limitations of think-aloud collection and analysis

The data collection of the think-aloud data itself had some limitations. First the sample used in this study may be potentially different from the sample utilized for the DIF analyses. The students were from a different school and had slightly higher averages for LSAT scores and first-year GPA. These differences in the samples could potentially impact the results of the think-aloud analyses as students with a higher GPA might utilize different strategies when completing these items. However, even though the students were enrolled at a different school, almost all law schools follow the same traditional curriculum so students at each school would likely have been introduced to similar topics and taught using a similar approach.

Another limitation of this portion of the study concerns the number of items completed by each student in the think-aloud sample. Because of the potential fatigue effect, each student only completed a segment of each test form. Therefore, think-aloud data for most items on each form was only collected from seven or eight students. However, even though a statement generalizing the conclusions from the items cannot be made, the findings are still very valuable and can help to minimize potential construct-irrelevant sources of error on responses. The small sample will not identify every

possible source of error, but the findings from the think-aloud data can help minimize these sources.

Implications for future research

Several ideas for future research emerged during the analysis of the think-aloud protocols. First, the think-aloud protocols provided insight into the intended construct of legal case reading and reasoning. In the data analysis, the raters identified the types of errors that students tended to make when responding to the items. The purpose of this data analysis was to identify sources of construct-irrelevant error. However, this data also provides insight into the common types of construct-relevant errors that students tend to make when attempting to perform a task of this nature. The errors that students tended to make during the task consisted of two types: strategy errors and skill errors. Strategy errors are those types of mistakes related to the strategies for responding to the items and included over-relying on memory, rereading selectively, and referring only to one case for cross-case items. Skill errors are those mistakes that reflect students' skills at reading and reasoning through the cases. These included errors such as the failing to consider full holding of the court, vocabulary failures, utilizing legal terminology in a commonplace manner, using everyday ideas of justice rather than relying on legal principles, and utilizing faulty logic. A future research study may explore these types of errors more in depth and develop interventions to help students employ better strategies and to enhance their skills in the construct of legal case reading and reasoning. In addition, the think-aloud process itself may be used as a form of intervention. Students can be asked to think-aloud the items then discuss the process, helping students to be more aware of the types of errors they make in reading and reasoning through cases.

Another area for future research with the think-aloud data is in better understanding test-taking strategies that students use. Considering that this test is intended for the students seeking an advanced degree, it was surprising that many students resorted to using test-based strategies in selecting their response. Test-based strategies that were observed in the think-aloud protocols included pattern seeking in options or across items, looking for clues in other items, and thinking that certain options were "trick" questions. While certain strategies were found to be detrimental to performance, other students used strategies that might prove to be beneficial. For example, one of the interesting findings in the protocols is the strategies that students utilize after reading the question stem. Many students read the stem and then jumped directly into reading the answer choices, eliminating and weighing options against each other. Other students took a more reflective approach, attempting to answer the question posed in their own words before moving ahead to reading the options. These students used a matching strategy in which they searched for an option that most closely matched their own constructed answer. A very cursory analysis of these students' protocols suggests that this strategy may be beneficial to performance.

One of the limitations of the think-aloud methodology concerns the time-consuming nature of data collection and analysis. Each student must individually complete the test whereas group administration was performed for the collection of the larger sample for the statistical analyses. Each session took approximately 1-2 hours of time. The transcripts from the think-alouds must be transcribed, which is also very time-consuming and tedious. In addition, because of the vast amount of information collected, the data analysis in this study took months. Many test developers do not have this time or

the finances to gather and analyze this data. However, one must weigh the benefits that emerge from this data regarding test validity. This study supported that the benefits for test development vastly outweighed these costs. Future research should examine modified methods of collecting response process evidence, such as using computerized collection of verbalizations, in order to reduce the time and cost associated with this type of data.

The think-aloud data in this study proved to be very valuable in several ways. The analysis of the data provided insight into the construct of interest, in how students confront tasks of legal case reading and reasoning. The data provided information about both positive and negative types of test-taking behaviors. Most importantly, the think-aloud protocols provided information about how individual items function and the complex relationship between the item stem, key, and distractors. While there are some limitations with the study, think-aloud methodology shows potential for the use of demonstrating construct-equivalence across test forms.

Discussion of Expert Judgment Analyses

The data from the experts supports that at least on some level, experts were aware of the categorization used to guide the construction of the test of legal case reading and reasoning. The item pairs that the experts identified as being nonequivalent were Pairs 11, 13, and 14, all of which consist of indeterminate items. The stems of these four pairs do seem to be the least isomorphic in nature. Although the items within the pairs have the same purpose, the items were often approached in a different manner which could explain why the experts rated these to be the least similar.

While not the focus of this study and thus not presented in detail, some additional analyses were performed to better understand the process by which experts selected their ratings. A series of analyses were performed on the expert ratings of all the presented item pairs, including the 14 pairs considered theoretically equivalent and those considered theoretically nonequivalent. The analyses support that the pairs that contained theoretically equivalent items were rated as more similar than those that were not considered theoretically equivalent. In addition, the experts were found to rate item pairs that contained items the same along the case dimension (both single-case or both cross-case) more similarly than pairs that contained one of each type. This finding was also found when analyzing the expert ratings of item pairs based on the determinate/indeterminate dimension. Experts rated item pairs that contained items the same along the determinate/indeterminate dimension more similarly than those pairs that contained one of each type. While they did not explicitly state that the case factor or the determinate/indeterminate factor influenced their ratings in their comments, theses supplemental analyses suggest that the experts were able to glean a basic understanding of the categories represented in the test.

The expert comments, the analyses of their rating based on case dimension, and an examination of the pairs with the least similar ratings support that experts' ratings of similarity is influenced by both the theoretical structure of the test and the surface features.

Limitations of expert data collection and analysis

This study did have some limitations regarding the collection of expert similarity ratings. Because the experts voluntarily donated their time and effort for the project, they

were only asked to rate the similarity of 25 pairs rather than the more complete paired comparison procedure introduced by Sireci (1998). This procedure would have yielded more information about the features of items that experts attended to in this task. Because only 25 pairs were rated, the means of the items were used to flag the pairs perceived to be most similar. The means for such a small sample size of experts could be potentially misleading. However, it was not possible to gather a larger sample of experts for this task.

An additional limitation of this portion of the study is that experts were only presented with the item stems and not the options in the rating task. Although experts had access to the item options, the rating sheets themselves only contained the stems for each item in the pair. The reason for this decision is that the items are quite lengthy and again, because the experts were volunteers, the task needed to be as simple as possible. However, some of the differences in items between the forms could be attributed to the way the key and the distractors function rather than differences in the stem itself. Indeed, the think-aloud data supported the complexity of item functioning and included examples where the item options completely changed the way the item worked. Therefore, the data for this portion of the study primarily consists of the experts' perception of the similarities of the stems, rather than the complete items, and thus has to be interpreted cautiously.

Implications for future research

Surprisingly, there was not much information in the literature on how experts could be used to judge construct equivalence of alternate test forms. In addition, the literature was limited on the characteristics of items that experts attend to while

completing rating tasks such as these. Additional research should be conducted to examine how experts decide to rate similarity of items, such as in the paired comparison procedure or other rating tasks such as these. Another possible area of research would be to extend the methodology of the think-aloud to the experts to examine what cognitive processes they utilize while completing the rating tasks. This may provide valuable insight into how expert judgment differs from the students' actual test-taking experience. However, while this method might prove interesting and value, it is also likely to be expensive and time-consuming for the test development process.

While the results of the expert judgment method in this study provided some interesting results, the information was not very helpful for test development, partly because expert judgment had already been collected previously during earlier development stage. The information collected in this portion of the study is not likely to lead to any test revisions. The item writers did express some frustration that potential item flaws had not been previously identified through the previous expert reviews. However, expert judgment is still very valuable as experts are needed to ensure correctness of content. Based on the results of this study, the expert judgment does not seem to be sufficient in identifying item flaws or in determining the construct equivalence of alternate test forms.

Discussion of Statistical Analyses

The statistical information collected on the test forms suggested that some revisions were necessary to many of the items. The statistical analyses identified Pairs 1, 7, 8, and 14 as being nonequivalent. Most items had low discrimination, suggesting that the items were unable to differentiate between students who possessed skills relating to

the construct and those who were lacking in these skills. In traditional test construction, test developers would likely examine the items closely to see how the discrimination indices could be improved. In comparing the discrimination values between the equivalent and nonequivalent item pairs as identified by the DIF methods, no discernible differences are evident. Therefore, it is not likely that the conclusions of this study are influenced by the poor discrimination values.

Limitations of statistical DIF methods and analysis

The data used for the statistical analyses in this project had some limitations. First, over half of the sample was not randomly assigned the form that they would complete first. The majority of the sample received the TV1 form followed by the TV2 form. Although the statistical tests supported that the total scores were not significantly different based on the order that students received the test, the patterns of responses among items may still have potentially influenced the results. A second limitation of the data concerns the time during which students received each test form. A portion of the sample were administered the forms during their first and third years of law school whereas the remaining sample received the forms during their first and second years. Once again, the statistical tests supported that there was no difference on total score based on the year during which students had received the forms. In addition, research supports that the current traditional law school curriculum does not sufficiently teach the skills and knowledge measured by the test, particularly the ability to do cross-case analyses and to identify indeterminacies within a case.

The analyses performed for the DIF procedures also had some limitations due to the nature of the data in that the analyses that were possible given the dimensionality and

dependency issues were limited. The advantage of using one sample for this project was that each person served as his or her own control for ability. However, this unexpectedly resulted in a severe limitation in the types of analyses that could be performed. Logistic regression procedures and classical DIF analyses using chi-square tests were not possible as they assume that observations are independent. If the dimensionality of the instrument was confirmed to be unidimensional or if a clear factor structure could be identified, Rasch procedures could have been attempted to estimate the item parameters. However, the data did not support the assumption of unidimensionality and subsequent factor analyses not reported in this paper could not specify a clear factor structure. Therefore, primarily descriptive procedures were used in the form of graphical and DIF size indices.

The McNemar test was performed as an alternative to the traditional DIF methods for the reasons described above. This statistical test along with generalized estimating equation (GEE) methods reflect potential areas for future research in using single-group designs to explore construct equivalence between forms as both methods allow for dependencies in the data. The McNemar test was used in this study in two ways. First, the test was used to compare the differences in changes across the range of all total scores. When utilized in this fashion, this test looks to see if there was a general pattern in differences in responding to the items within the pairs, for all students regardless of their total omnibus test score summed across both forms. The second way that the McNemar test was used to determine whether the differences in difficulty between items in a pair were differential based on the omnibus total score. Thus, two score levels were used to obtain information on low-scoring versus high-scoring students. The very coarse division between only low versus high scoring students was necessary due to the small

sample size of the study. When more intervals were used in the test, certain cell frequencies were too small to perform the McNemar test. While the information gathered from the McNemar test seemed to converge with other indices calculated in this study, this application needs to be further explored for the comparing theoretically equivalent items from test forms for single-group designs.

One limitation of all the detection methods used in this study is the reliance on the total test score as a criterion for detection. As is the case with all DIF studies that rely on total score, if there is a widespread difference between groups, or in this case forms, the total score may not be an appropriate criterion for detection. As Camilli and Shepard (1994) note, "The fact that item bias procedures rely on an internal criterion means that none of the various indices – old or new – is sensitive to constant or pervasive bias" (p. 24). If the two test forms are not measuring the same construct or are measuring the construct in a systematically different manner, the items detected by DIF could be affected. One advantage in this study that would not be possible in conventional DIF studies using groups is that the total omnibus test score was utilized as the internal criterion which would lessen the impact of systematic bias.

Conclusion

While this study does have limitations regarding each of the three methods compared in demonstrating construct equivalence, the think-aloud procedure demonstrated an invaluable source of information regarding item functioning. Response process information is a necessary component of the validity argument. This type of information cannot be duplicated through either expert judgment or through statistical procedures. Test developers need to expand their repertoire of tools in test construction

to include think-aloud procedures. The utilization of think-aloud procedures has many potential uses in test development, from the process described here of ensuring construct equivalence, to simply making sure that examinees understand the items, and to gaining a better understanding of test-taking strategies. Although the think-aloud method is time-consuming and expensive when done right, so is the use of expert panels. In addition, the information gleaned from the examinees themselves, demonstrating the processes actually utilized when responding to items, is extremely important to ensure that a test is functioning in the manner intended.

REFERENCES

Afflerblach, M. & Afflerbach, P. (1995). *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading.* Hillsdale, NJ: Lawrence Erlbaum.

Agresti, A. (2002). *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons.

American Educational Research Association, American Psychological Association, National Council for Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*. *16*, 55-73.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*. *36*, 185-198.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In Berk, R. A. (Ed.) *Handbook of methods for detecting item bias.* Baltimore, MD: Johns Hopkins University Press, 96-116.

Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! *Applied Measurement in Education, 8,* 99-109.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., Freidebach, M. (2001). Read-aloud accommodation: Effects on multiple-choice reading and math items. Technical Report. Report based on a paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Bloom, B. (1984). *Taxonomy of Educational Objectives.* Boston, MA: Allyn and Bacon.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage Publications.

Christensen, Leah M. (2006). Legal Reading and Success in Law School: An Empirical Study. U of St. Thomas Legal Studies Research Paper No. 06-29 Available at SSRN: http://ssrn.com/abstract=924650.

Cizek, G. J. (1999). *Cheating on Tests: How to Do It, Detect It, and Prevent It.* Mahwah, NJ: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum.

Czaja, R., & Blair, J. (1996). *Designing Surveys.* Thousand Oaks, CA: Sage Publications.

Deegan, D. H. (1995). Exploring individual differences among novice reading in a specific domain: The case of law. *Reading Research Quarterly. 30*, 154-170.

Ding, C. S. & Hershberger, S. L. (2002). Assessing content validity and construct equivalence using structural equation modeling. *Structural Equation Modeling. 9*, 283-297.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement. 23,* 355-368.

Englehard, G., Davis, M., Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education. 12,* 199-210.

Enright, M. K. & Bejar, I. I. (1989). An analysis of test writers' expertise: Modeling analogy item difficulty. Report ETS-RR-89-35. Princeton, NJ: Educational Testing Services.

Enright, M. K., Tucker, C. B., & Katz, I. R. (1995). A cognitive analysis of solutions for verbal, informal, and formal-deductive reasoning problems. GRE Board Professional Report No. 90-04P. Princeton, NJ: Educational Testing Services.

Ercikan, K., Law, D., Arim, R., Domene, J., Lacroix, S., & Gagnon, F. (2004). Identifying sources of DIF using think-aloud protocols: Comparing thought processes of examinees taking tests in English versus in French. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Ericcson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press.

Ericcson, K. A., & Simon, H. A. (1999). *Protocol Analysis: Verbal Reports as Data (2nd Edition).* Cambridge, MA: MIT Press.

Fajans, E. & Falk, M. R. (1992-1993). Against the tyranny of paraphrase: Talking back to texts. *Cornell Law Review, 78,* 163-205.

Ferrara, S., Duncan, T. G., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (April, 2004). Examining test score

validity by examining item construct validity. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ferdous, A. A. & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education. 18,* 257-267.

Fraser, C. (1983). *Noharm II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, University of New England, Centre for Behavioral Studies.

Frederiksen, N., Mislevy, R. J., & Bejar (1993). *Test Theory for a New Generation of Tests.* Hillsdale, NJ: Lawrence Erlbaum.

Gee, J. P. (1991). *Social Linguistics and Literacies: Ideology in Discourses.* London: Falmer Press.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *Journal of Educational Research. 91*, 26-32.

Gierl, M. J. & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement. 38,* 164-187.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement. 42,* 351-373.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice. 25,* 21-35.

Graham, M. & Anderson, B. (2000-2001). Law students' undergraduate major: Implications for law school Academic Support Programs (ASPs). *UMKC Law Review. 69,* 533-556.

Gross, P. W. (1972-73). On law school training in analytic skill. *Journal of Legal Education. 25,* 261-311.

Gusky, T. R. (1997). *Implementing Mastery Learning.* Belmont, CA: Wadsworth Publishing Company.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items.* Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education. 15,* 309-34.

Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). Anchor-based methods for judgmentally estimating item difficulty parameters. Computerized Testing Report 98-05 for the Law School Admission Counsel. Newtown, PA: Law School Admission Counsel.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Haney, W. & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In Freedle, R. O. (Ed.) *Cognitive and Linguistic Analyses of Test Performance.* Norwood, NJ: Ablex Publishing Corporation.

Hess, G. F. (2002). Heads and hearts: The teaching and learning environment in law school. *Journal of Legal Education, 52,* 75-111.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2003). Reliability methods: A note on the frequency of use of various types. In Thompson, B. (Ed.) *Score Reliability: Contemporary Thinking on Reliability Issues.* Thousand Oaks, CA.

Holland, P.W. & Rubin, D. B. (1982). Introduction: Research on test equating sponsored by Educational Testing Service, 1978-1980. In Holland, P.W. & Rubin, D. B. (Eds). *Test Equating,* p. 1-6. New York, NY: Academic Press.

Holland, P.W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (p 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P.W. & Wainer, H. (Eds). (1993) *Differential Item Functioning: Theory and Practice.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Jabine, T. B., Straf, M. L., Tanure, J. M., & Tourangeau, R. (1984). Cognitive aspects of survey methodology: Building a bridge between disciplines. A report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. Committee on National Statistics.

Jensen, A. R. (1980). *Bias in Mental Testing.* New York: Free Press.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Kobrin, J. L. & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education. 16*, 115-140.

Kolen, M. J. & Brennan, R. L. (1995). *Test Equating, Scaling, and Linking: Methods and Practices.* New York, NY: Springer.

Leighton, J. P. (2004). Avoiding misconception, missed use, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice. 23*, 6-15.

Leighton, J. P. & Gokiert, R. J. (April, 2005a). Investigating test items designed to measure higher-order reasoning using think-aloud methods: Implications for construct validity and alignment. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Leighton, J. P. & Gokiert, R. J. (April, 2005b). The cognitive effects of test item features: Informing item generation by identifying construct irrelevant variance. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher. 20*, 15-21.

Loehlin, J. C. (1998). *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis.* Mahwah, NJ: Lawrence Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lundeberg, M. A. (1987). Metacognitive aspects of reading comprehension: Understanding in legal case analysis. *Reading Research Quarterly, 22*, 408-432.

Martinez, M. E. & Katz, I. R. (1995). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational Assessment. 3*, 83-98.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd Edition). New York: Macmillan.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement. 30,* 55-78.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing. 1,* 115-135.

Nelson, L. R. (2001). Item analysis for tests and surveys using Lertap 5. Perth, Australia: Curtin University of Technology.

Nitko, A. J. (2004). *Educational Assessment of Students.* Upper Saddle River, NJ: Pearson Education, Inc.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement. 27,* 41-58.

Norris, S. P. (1992). A demonstration of the use of verbal reports of thinking in multiple-choice critical thinking test design. *The Alberta Journal of Educational Research. 38,* 155-176.

O'Neil, T., Sireci, S. G., & Huff, K. L. (2003-2004). Evaluating the consistency of test content across two successive administrations of a state-mandated science assessment. *Educational Assessment. 9,* 129-151.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education. 24*, 307-353.

Plake, B. S. & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment. 7,* 87-97.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education. 5,* 285-301.

Pressley, M., & Afflerbach, P. (1995). *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading.* Hillsdale NJ: Erlbaum.

QSR International (2002). *N6.* Victoria, Australia.

Rovinelli, R. J., & Hambleton, R. K. (April, 1976). On the use of content specialists in the assessment of criterion-referenced test item validity. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Rzasa, S. E. (April, 2003). Item analysis on a developmental rating scale using both statistical and qualitative methods. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Savage, I. R. (1982). General Remarks. In Holland, P.W. & Rubin, D. B. (Eds). *Test Equating,* p. 343-344. New York, NY: Academic Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing. 18*(1), 55-88.

Schwarz, R. D., Rich, C., Arenson, E., Podrabsky, T., & Cook, G. (2002). An analysis of differential item functioning based on calculator type. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Schwarz, R. D., Rich, C., Podrabsky, T. (2003). A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Siegel, S. & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* Boston, MA: McGraw Hill.

Sireci, S. G. (1998). Gathering and evaluating content validity data. *Educational Assessment. 5*(4), 299-321.

Sireci, S. G.& Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing. 20*(2), 148-166.

Sireci, S. G. & Berberoglu (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education. 13,* 229-248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). Adapting credentialing examinations for international uses. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Sireci, S. G. & Swaminathan (October, 1996). Evaluating translation equivalence: So what's the big DIF? Paper presented at the Annual Meeting of the Northeastern Educational Research Association, Ellenville, NY.

Skorupski, W. P. & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education. 18*, 233-256.

Snow, R. E. & Lohman, D. F. (1993). Cognitive psychology, new test design and new test theory: An introduction. In Frederksen, N., Mislevy, R. J., & Bejar, I. I. (Eds). *Test Theory for a New Generation of Tests.* Hillsdale, NJ: Lawrence Erlbaum.

Stratman, J. F. (1990). The emergence of legal composition as a field of inquiry: Evaluating the prospects. *Review of Educational Research, 60*(2), 153-235.

Stratman, J. F. (2002). When law students read cases: Exploring relations between professional legal reasoning roles and problem detection. *Discourse Processes, 34*(1), 57-90.

Stratman, J. F., Evensen, D. H., & Oates, L. C. (June 2005). Developing an assessment of 1st year law students' critical case reading and reasoning ability. Report to Law School Admissions Council.

Strauss, A. & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage Publications.

Suen, H. K. (1990). *Principles of Test Theories.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Sullivan, W. M., Colby, A., Wegner, J. W., Bond, L., & Shulman, L. S. (2007). *Educating lawyers: Preparation for the Profession of Law.* Stanford, CA: The Carnegie Foundation for the Advancement of Teaching.

Thissen, D. & Orlando, M. (2001) Item response theory for items scored in two categories. In Thissen, D. & Wainer, H. (Eds.) *Test Scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In *Test Equating.* In Holland, P.W. & Rubin, D. B. (Eds). *Test Equating,* p. 309-317. New York, NY: Academic Press.

Turner, R. C. & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing, 3,* 163-171.

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing. 22*(2), 211-234.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (1997). *TestFact 4*. Lincolnwood, IL: Scientific Software International.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing. 20*(2), 136-147.

APPENDIX A

Sample TV1 Items

Single-case level, determinate item

Pair 1 item:

Which one of the following statements best summarizes the court's reasoning in Mackey?

a) *Citing Black and Brown, the court concludes that Mackey's attempts to learn what he owed and to make his payment on time are neither "valid" nor "honest." The court states that Mackey should have known the procedural rules and acted upon them, but otherwise offers little reasoning concerning why Mackey's attempts to learn his costs and the correct due date are neither "valid" nor "honest."

b) The court reasons that if the rule requiring payment of opposing counsel's record costs were relaxed, litigants could not learn when no further appeals may be taken by the losing party in arbitration.

c) Because Mackey did not produce physical evidence of timely mailing, Mackey's appeal cannot be perfected and must be quashed.

d) The court reasons that the fact that the Prothonotary's office put Mackey on hold is irrelevant since Mackey could have easily continued trying to reach them within the statutory time limit.

e) Following Black and Brown, the court reasons that Mackey's attempts to learn what he owed and to make his payment on time are neither "valid" nor "honest" because Mackey tries to have it both ways: he argues that he could not read the pencil note, and also that he mistakenly juxtaposed the numbers. The court thus finds this inconsistent argument less than honest.

APPENDIX A (continued)

Cross-case level, determinate item

Pair 5 Item

Which one of the following statements best summarizes the legal issue raised by all three of the cases you have read?

a) Under what circumstances may losing parties in Pennsylvania arbitration decisions validly rely upon the record cost amounts provided to them by opposing counsel rather than the amount officially recorded?

b) Under what circumstances may a Pennsylvania court declare a party's failure to comply with procedural statutes for perfecting an appeal to be de minimis and disregarded?

c) *In what ways, if any, may a party fail to comply with the procedural statute requiring a losing party in arbitration to "pay all costs that have accrued on such suit or action . . . within 20 days after the entry of the award of the arbitrators on the docket" and yet proceed with an appeal?

d) When attempting to appeal an adverse arbitration decision to a regular appellate court in Pennsylvania, what portion of the total record cost owed to opposing counsel must a losing party pay?

e) Under what conditions may a court decide to void a statute because it has clearly put form over substance or has otherwise compromised justice by elevating mere technical detail?

APPENDIX A (continued)

Single-case, indeterminate item

Pair 10 Item

Assume for the moment that you decide to submit a motion to the Common Pleas Court to reconsider its decisions against your client, Mackey Plumbing Co. Which one of the following "unknowns" in the facts of the Mackey case would be most relevant to your motion?

a) Did Mackey have counsel, or was he acting pro se?

b) Does Mackey still have a copy of the dated fax he sent to the plaintiff?

c) Has Mackey Plumbing Co. ever previously lost an arbitration hearing and subsequently failed to pay record costs?

d) Does Mackey have some visual impairment?

e) *Is there a way of finding out whether the plaintiff's communication was deliberately versus accidentally misleading?

APPENDIX A (continued)

Cross-case, indeterminate item

Pair 12 Item

An advocate must routinely think through the way opposing counsel may analyze the law. Assuming that you have decided to appeal the Common Pleas Court decision (in Mackey) to Superior Court, which one of the following questions about opposing counsel's approach to the Mackey decision's relation to precedent is most relevant to you?

a) How might opposing counsel compare Mackey's situation with Yellow Cab's situation (in Meta)?

b) *How might opposing counsel compare Mackey's situation with the defendant's situation in Black and Brown?

c) How might opposing counsel make use of the dissent in the Meta decision?

d) To what extent will opposing counsel consider Mackey's compliance with other procedural requirements for appeal (e.g., such as payment of the arbitrators' fee) in reviewing the Common Pleas Court decision quashing his appeal?

e) Given that the Black and Brown court stressed that the appellant in that case had "express notice" that record costs be paid within the 20 day limit, how likely is it that opposing counsel will argue that Mackey similarly had "express notice" that he had to pay his record costs?

APPENDIX B

Sample TV2 Items

Single-case determinate item

Pair 1 Item:

Which one of the following statements most explicitly articulates the substantive reasoning used by the district court to support its decision to award summary judgment to the defendants in the *Bridell* case?

a) The court gave two reasons: first, because Bridell failed to present any evidence showing that Ribier engaged in discriminatory practices; and second, because Ribier's summary plan document was "written in a manner calculated to be understood by the average plan participant," as required by ERISA.

b) The court reasoned that Ribier Clothing Inc. should be allowed to reduce their medical benefits for two reasons: first, because of the steep, unpredictable rise in the cost of such benefits; and second, because "the paramount purpose of ERISA is to protect the solvency of employee benefit plans."

c) *The court advanced two reasons for awarding summary judgment: first, because Bridell was unable to convincingly demonstrate that an immutable, promised benefit had been revoked, and second, because Bridell failed to present any evidence showing that Ribier engaged in prohibited conduct or that the company interfered with any right to benefits under ERISA to which she was entitled.

d) The court awarded summary judgment to Ribier Clothing Inc. because Bridell failed to sufficiently establish the existence of an element essential to her case and on which she would need to bear the burden of proof at trial."

e) The court reasoned that summary judgment should be awarded to Ribier Clothing Inc. for two reasons: first, because "Bridell has not and could not reasonably contend that the SPD's earlier reference to 'lifetime medical benefits for eligible employees and their spouses' somehow renders [Ribier Clothing's] reservation of rights in the footnote disclaimer ambiguous;" and second, because ERISA does not require employers to provide any welfare benefits at all.

APPENDIX B (continued)

Cross-case, determinate item

Pair 5 Item

Which one of the following choices best summarizes the legal issue(s) addressed by *all three* of the cases you have just read?

a) *Under what conditions does a company's use of a reservation of rights clause in its benefits plan documents satisfy ERISA's competing goals of accurately disclosing the substance of employee benefits while also protecting a company's unilateral right to offer or not offer such welfare benefit plans?

b) Under what conditions does an employers' use of reservation of rights clauses in ERISA summary plan documents violate ERISA's fundamental goal of providing welfare benefit plans in a non-discriminatory manner?

c) Under what conditions may employers satisfy ERISA's goal of protecting employees' benefits while at the same time denying welfare benefits explicitly identified in both benefit plan documents and in specific, individual employee cases?

d) Under what conditions may an employer's prior history and behavior in increasing or reducing plan benefits to employees be used to clarify the meaning of disputed terminology in published welfare benefit plans?

e) Under what conditions may ERISA's goal of protecting employers' unilateral right to offer or not offer welfare plan benefits to employees be given more weight than ERISA's goal of protecting employees' reasonable expectations both for such benefits and accurate disclosure concerning them?

APPENDIX B (continued)

Single-case, indeterminate item

Pair 10 Item

Assume that you are considering ways in which you might help Bridell to appeal the decision of summary judgment against her. Which one of the "unknowns" in the facts of the *Bridell* case would provide the best grounds for a possible motion?

a) Was there any evidence that the "financial losses" suffered by Ribier had not resulted from extraordinary expenses related to their benefit plan, but were the result of high risk investments that failed to yield expected returns?

b) Are there any records of conversations between Ribier's senior managers and its insurers concerning the projected costs related to the lifetime coverage of the average AIDS patient?

c) *Are there any company records (letters, recorded phone calls, emails, etc.) that contain individual employee questions or confusions about how the SPD disclaimer affects "lifetime benefits," and can it be shown that some company personnel responded evasively or unclearly to these?

d) Did Bridell inform her employer, Ribier, of her AIDS diagnosis in a memo in which she also made reference to the terms of the existing plan and did a company representative acknowledge that memo in writing?

e) Did Ribier ever change the provisions of its medical plan either increasing or decreasing benefits?

APPENDIX B (continued)

Cross-case, indeterminate item

Pair 12 Item

Assume that your petition to appeal the *Bridell* case has been successful. As you prepare your case you will be thinking about how opposing counsel might look to other cases to effectively develop its case. In light of this task, which one of the following questions regarding opposing counsel's argument strategy would concern you the most?

a) To what extent will opposing counsel try to establish that just as medical benefits could be reduced for retirees as they were in *Alexander*, so can they be reduced for active employees like Bridell?

b) How might opposing counsel compare the disclaimer evidenced in Bridell's plan with the disclaimer found in Hamilton's plan?

c) To what extent might opposing counsel draw upon both *Hamilton* and *Alexander* to synthesize a rule concerning promises that inhere in benefit plan documents?

d) How might opposing counsel rely on the *Alexander* court's hypotheticals about possible federal legislation affecting the provision of private health care plans?

e) *To what extent will opposing counsel attempt to show that the case by case justification put forward in *Hamilton* rightfully extends to Bridell's case?

APPENDIX C

Item pairings

Single-case, determinate

Pair #1:
Understanding of court's reasoning in main case (Mackey/Bridell)

> TV1 Item:
> Which one of the following statements best summarizes the court's *reasoning* in *Mackey*?
>
> TV2 Item:
> Which one of the following statements most explicitly articulates the substantive reasoning used by the district court to support its decision to award summary judgment to the defendants in the *Bridell* case?

Pair #2:
Understanding of legal issues

> TV1 Item:
> Which one of the following questions best expresses the *legal issue* that the *Mackey* court sees itself as addressing?
>
> TV2 Item:
> Which one of the following statements best expresses the issue(s) which the District Court of New Jersey, on remand, will have to decide in view of the *Alexander* court's analysis and decision?

Pair #3:
Understanding of legal reasoning in supplemental cases

> TV1 Item:
> Which one of the following statements best summarizes the *reasoning* in the *Black and Brown* decision?
>
> TV2 Item:
> Which one of the following statements best summarizes the court's reasoning in *Alexander v. Primerica*?

APPENDIX C (continued)

Pair #4:
Understanding of more subtle (not the main) reasoning in supplemental cases

TV1 Item:
Which one of the following statements best summarizes the *reasoning* in the minority opinion (dissent) in *Meta*?

TV2 Item:
After the *Hamilton* court dealt with the threshold issue of whether or not Air Jamaica's Handbook constituted an ERISA plan, it went on to reason about a second issue. Which one of the following statements best summarizes the court's reasoning concerning this second issue?

Cross-case, determinate

Pair #5:
Summarization of all three cases

TV1 Item:
Which one of the following statements best summarizes the legal issue raised by *all three* of the cases you have read?

TV2 Item:
Which one of the following choices best summarizes the legal issue(s) addressed by *all three* of the cases you have just read?

Pair #6:
Similarities/differences in the facts

TV1 Item:
Assume for the moment that you decide to submit a motion to the Common Pleas Court to reconsider its decision against your client, Mackey Plumbing Co. Which one of the following statements best summarizes the similarities and differences between the facts in *Mackey* and the facts in *Meta*?

TV2 Item:
Assume for the moment that you decide to prepare an appeal to the Fifth Circuit to re-examine the district court decision against your client, Joan Bridell. Which one of the following summaries of the similarities between the facts in *Bridell* and the facts in *Alexander* would be most accurate *as well as* most useful for you on appeal?

APPENDIX C (continued)

Pair #7:
Similarities/differences in reasoning

TV1 Item:
Which one of the following statements best summarizes the similarities and differences between the courts' reasoning in *Meta* and the court's reasoning in *Black and Brown*?

TV2 Item:
Which one of the following statements best summarizes the differences in reasoning between the district court and circuit court decisions in *Hamilton*, so far as these differences are visible in the *Hamilton* opinion?

Pair #8:
Constitutionality/legality issues

TV1 Item:
As an advocate representing Mackey Plumbing Co., you may want to find out more about the constitutionality of the record cost statute. As far as this constitutionality issue is concerned, which one of the following statements seems most accurate, based upon the three cases you have read?

TV2 Item:
Which one of the following statements best summarizes the view of the courts in the three cases you read concerning the fundamental legality, under ERISA, of employers' use of disclaimers to reserve the right to change their welfare benefit plans?

Single-case, indeterminate

Pair #9:
How questions/ambiguities in supplemental case can help in appeal for client

TV1 Item:
Courts sometimes use phrasing that leaves room for interpretation by other courts. However, while an opinion may contain a number of statements that are ambiguous, some of these ambiguities may be more significant than others. Assuming you decide to appeal the *Mackey* decision to Superior Court, which one of the following ambiguities presented by the *Black and Brown* opinion is most relevant to you?

TV2 Item:
When you are researching among cases, you may find that they contained unanswered questions that, if logically extended to your case, could frame issues upon which you could build arguments for your client. Which one of the following questions remains unanswered in *Hamilton*, and, as such, might best provide you with an issue relevant to your appeal for Bridell?

Pair #10:
What are the unknowns in client's case that would be helpful for appeal

TV1 Item:
Assume for the moment that you decide to submit a motion to the Common Pleas Court to reconsider its decisions against your client, Mackey Plumbing Co. Which one of the following "unknowns" in the facts of the *Mackey* case would be most relevant to your motion?

TV2 Item:
Assume that you are considering ways in which you might help Bridell to appeal the decision of summary judgment against her. Which one of the "unknowns" in the facts of the *Bridell* case would provide the best grounds for a possible motion?

Pair #11:
Information that would best help client in a new case

TV1 Item:
All opinions are interpretations, and interpretations invite questioning. Assuming that you decide to appeal the *Mackey* decision to Superior Court, which one of the following questions presented by the *Mackey* opinion is most relevant to this task?

TV2 Item:
Although you did not represent Bridell at the District Court level, you have agreed to take Bridell's case to the Fifth Circuit. In your research you find that Section 510 of ERISA states: "It shall be unlawful for any person to discharge, fine, suspend, expel, discipline, or discriminate against a participant or beneficiary for exercising any right to which [s]he is entitled under the provisions of an employee benefit plan…" 29 U.S.C. § 1140. In light of this provision, which one of the following questions, if found in the affirmative, would be most likely to thwart Ribier's position in a new case?

APPENDIX C (continued)

Cross-case, indeterminate

Pair #12:
Opposing counsel's strategy in client's case

TV1 Item:
An advocate must routinely think through the way opposing counsel may analyze the law. Assuming that you have decided to appeal the Common Pleas Court decision (in *Mackey*) to Superior Court, which one of the following questions about opposing counsel's approach to the *Mackey* decision's relation to precedent is most relevant to you?

TV2 Item:
Assume that your petition to appeal the *Bridell* case has been successful. As you prepare your case you will be thinking about how opposing counsel might look to other cases to effectively develop its case. In light of this task, which one of the following questions regarding opposing counsel's argument strategy would concern you the most?

Pair #13:
Comparison of client's case to all other cases regarding the most relevant/important lines of questioning

TV1 Item:
Because the *Mackey* case was decided in a lower level court (Court of Common Pleas), the judge in that case would need to examine decisions in higher appellate courts to learn the applicable law. In Pennsylvania this court would be Superior Court. Given this circumstance, which one of the following questions about the relationship between the *Mackey* decision and the other two Superior Court decisions (*Meta* and *Black and Brown*) is most important for you to think about?

TV2 Item:
All court opinions involve the construction of arguments; in turn, arguments invite questions. In light of this, which one of the following lines of questioning related to Bridell's previous defeat in district court is best supported by the cases you have read?

APPENDIX C (continued)

Pair 14:
Comparison of client's case to one other case regarding the most relevant questioning

TV1 Item:
Most case analysis takes place in a specific context, with a specific problem in view. Assuming you decide to appeal the Mackey decision to Superior Court, which one of the following questions about the *Mackey* court's reliance on the *Black and Brown* decision is most relevant?

TV2 Item:
The *Alexander* court warns any court or lawyer who reads its opinion that "slight similarity" is insufficient to the purpose of comparing facts from one case to the facts in another case. In addition, the facts a lawyer selects for comparison must be significant within the context of the case at hand. Given these two considerations, which one among the following questions might present the strongest challenge to the *Bridell* court's analogical reasoning ?

APPENDIX D

Example of Item Justification

TV1 Pair 10 Item Justification

Best answer: (e)
Justification: This unknown may be relevant even if current law explicitly requires defendants to confirm the amount of record costs and the date they are due by contacting the Prothonotary's office. If evidence can be produced that the plaintiff deliberately sent erroneous information to the defendant, then an advocate for Mackey Plumbing Co. might argue that there was an intent to deceive on the plaintiff's part. On a motion to reconsider the case, the court might for this reason have some sympathy for Mackey's error in underpaying the amount and missing the correct due date.

Bad answer: (a)
Justification: This answer is bad because even if Mackey Plumbing Co. was acting *pro se* the court would not likely find that the defendant's non-compliance could be excused for that reason alone. The *Mackey* court stated, after all, that the "defendant should have known the rules and acted upon them instead," and many *pro se* defendants have likely met the requirement before with no difficulty.

Bad answer: (b)
Justification: This unknown is less relevant than (e) because there does not appear to be any dispute as to whether this letter was sent or concerning what it said. Indeed, apparently it *was* sent, since the plaintiff responded to Mackey Plumbing Co. with a pencil note.

Bad answer: (c)
Justification: This unknown is of little relevance because (even) if Mackey Plumbing Co. had previously failed to meet the record cost requirement following an adverse arbitration decision, this fact would only suggest that Mackey had learned little from the experience. Certainly this fact would not induce any sympathy on the part of the court, likely the very opposite.

Bad answer: (d)
Justification: This unknown may at first glance seem to be of some relevance to an advocate for Mr. Mackey, because if some sort of visual disability could be established, it might explain why Mr. Mackey transposed the record cost figure in the peculiar manner that he did. Yet there is nothing in the *Mackey* court's account of the facts to suggest that Mr. Mackey had or claimed any sort of disability as an excuse. And, even if he did have some such disability, the excuse would be bootless because the company would still be able to learn the amount owed and the due date by telephoning the Prothonotary.

APPENDIX E

Rating Sheet for Item Pairs

Pair #_____

1. Do both items have at least one student who has responded in the manner intended by the item writers?

    - 3 = Both items clearly display at least one student whose reasoning matches the intended justification.
    - 2 = Evidence matching the intended justification was not found for one of the two items
    - 1 = Evidence matching the intended justification was not found for both items

2. What are the strategies/reasoning utilized by the ideal students to respond to TV1 item? To the TV2 item?

    - 3 = The strategies/reasoning are equivalent or fairly equivalent between the two items
    - 2 = There are some differences in the strategies/reasoning between the two items
    - 1 = Significant differences exist in the strategies/reasoning between the two items

3. For each distractor, is there evidence that at least one student's reasoning matches the intent of the item writers?

TV1 – Option A_____	TV2 – Option A_____
TV1 – Option B_____	TV2 – Option B_____
TV1 – Option C_____	TV2 – Option C_____
TV1 – Option D_____	TV2 – Option D_____
TV1 – Option E_____	TV2 – Option E_____

4. What are the types of errors made by students who incorrectly respond to the item?

TV1 Construct-relevant Errors:	TV2 Construct-relevant Errors:
TV1 Construct-irrelevant Errors:	TV2 Construct-irrelevant Errors:

APPENDIX F

Dimensions for construct-equivalence between item pairs

Dimension	Criteria
Equivalent No changes (or only minor changes) suggested for either item	1. Each item has at least one student whose reasoning matches the intended justification for correctly responding. 2. The general strategies and reasoning utilized to answer the item are similar. 3. Most distractors in each item have at least one student whose reasoning matches the intended justification on why they are incorrect.
Equivalent One item should be revised	1. Each item has at least one student whose reasoning matches the intended justification for correctly responding. 2. The general strategies and reasoning utilized to answer the item are similar. 3. Most distractors in each item have at least one student whose reasoning matches the intended justification on why they are incorrect. 4. Evidence points to necessary changes in the stem, key, and/or distractors for only one item.
Equivalent Both items should be revised	1. Each item has at least one student whose reasoning matches the intended justification for correctly answering. 2. The general strategies and reasoning utilized to answer the item are similar. 3. Most distractors in each item have at least one student whose reasoning matches the intended justification on why they are incorrect. 4. Evidence points to necessary changes in the stem, key, and/or distractors for both items.
Not equivalent	1. There is a lack of reasoning for one or both of the items matching the intended justification for the key. 2. The necessary strategies and reasoning to answer each item are very different. 3. Most of the distractors are not functioning as intended.

APPENDIX G

Summary table for think-aloud analyses

Item Pair	Analysis of ideal response	Analysis of strategy use	Analysis of distractors	Conclusion on item
Pair #1	Both items clearly display at least one student whose reasoning matches the intended justification.	Some differences exist in strategies/ reasoning between the two items.	Evidence supports all distractors functioning as intended.	Equivalent (minor revisions in both items)
Pair #2	Both items clearly display at least one student whose reasoning matches the intended justification.	Significant differences exist in the strategies/ reasoning between the two items.	Option E in TV1 is not considered by students. Option A in TV2 does not have evidence to support functioning.	Not equivalent
Pair #3	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Evidence supports all distractors functioning as intended.	Equivalent
Pair #4	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Evidence supports all distractors functioning as intended.	Equivalent (minor revisions in TV1 item)

APPENDIX G (continued)

Pair #5	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Option B for TV1 does not have clear evidence to support functioning	Equivalent (revision in TV1 item)
Pair #6	Both items clearly display at least one student whose reasoning matches the intended justification	Significant differences exist in the strategies/ reasoning between the two items.	Evidence supports all distractors functioning as intended.	Not equivalent (revision in TV1 item)
Pair #7	Evidence matching the intended justification was not found for one of the two items	Strategies and reasoning are equivalent between the two items.	Option A and E in TV1 do not have clear evidence to support functioning	Equivalent (revision for TV1 item; minor revision for TV2 item)
Pair #8	Both items clearly display at least one student whose reasoning matches the intended justification	Significant differences exist in the strategies/ reasoning between the two items.	Option A in TV2 needs revision	Not equivalent (revision for TV1 item; minor revision for TV2 item)
Pair #9	Both items clearly display at least one student whose reasoning matches the intended justification (slightly less clear for TV2 item)	Strategies and reasoning are equivalent between the two items.	Most options in TV1 item are eliminated due to test-based reasons. Option A and E in TV2 do not have clear evidence to support functioning	Equivalent (revision for TV1 item)

APPENDIX G (continued)

Pair #10	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Evidence supports all distractors functioning as intended.	Equivalent
Pair #11	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Option B and C do not have clear evidence to support functioning Minor revisions for option D in TV1; Minor revisions for option C in TV2	Equivalent (minor revisions for both items)
Pair #12	Evidence matching the intended justification was not found for one of the two items	Strategies and reasoning are equivalent between the two items.	Minor revisions for option D in TV1	Equivalent (minor revisions for TV1 item; revisions for TV2 item)
Pair #13	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Evidence supports all distractors functioning as intended.	Equivalent (minor revisions for both items)
Pair #14	Both items clearly display at least one student whose reasoning matches the intended justification	Strategies and reasoning are equivalent between the two items.	Unclear evidence for TV1 Option C	Equivalent (minor revisions for TV1 item)

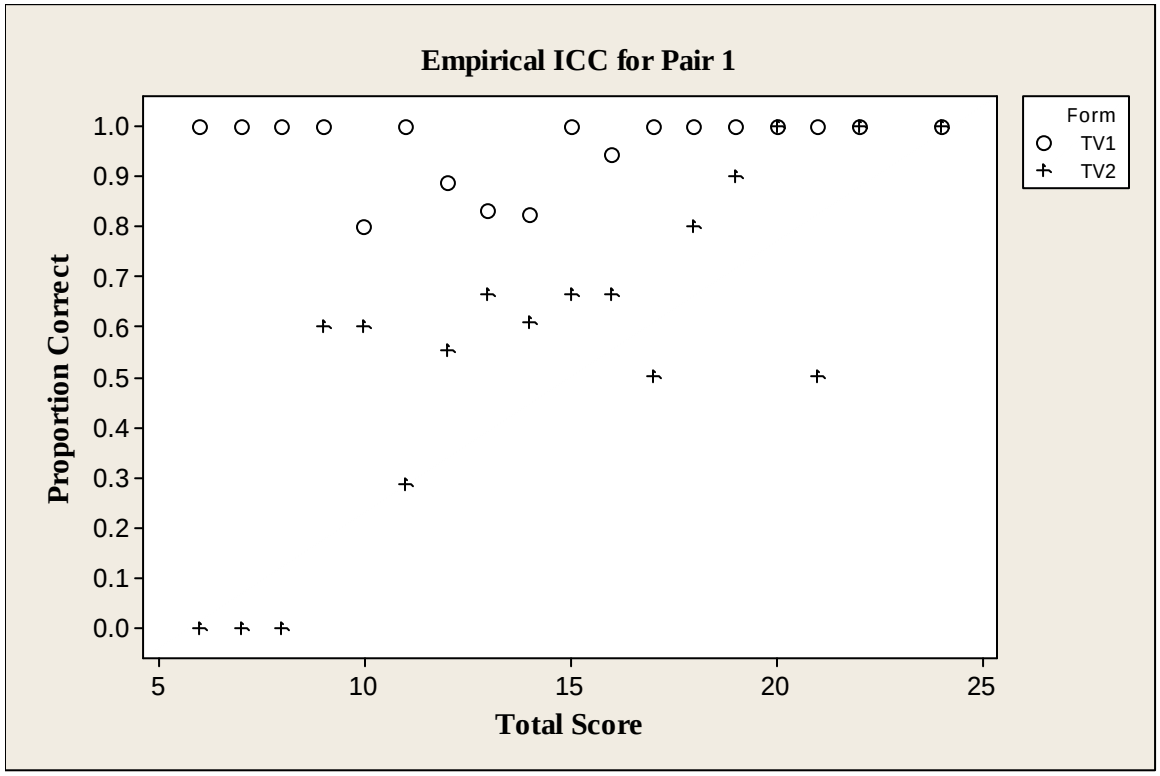
APPENDIX H

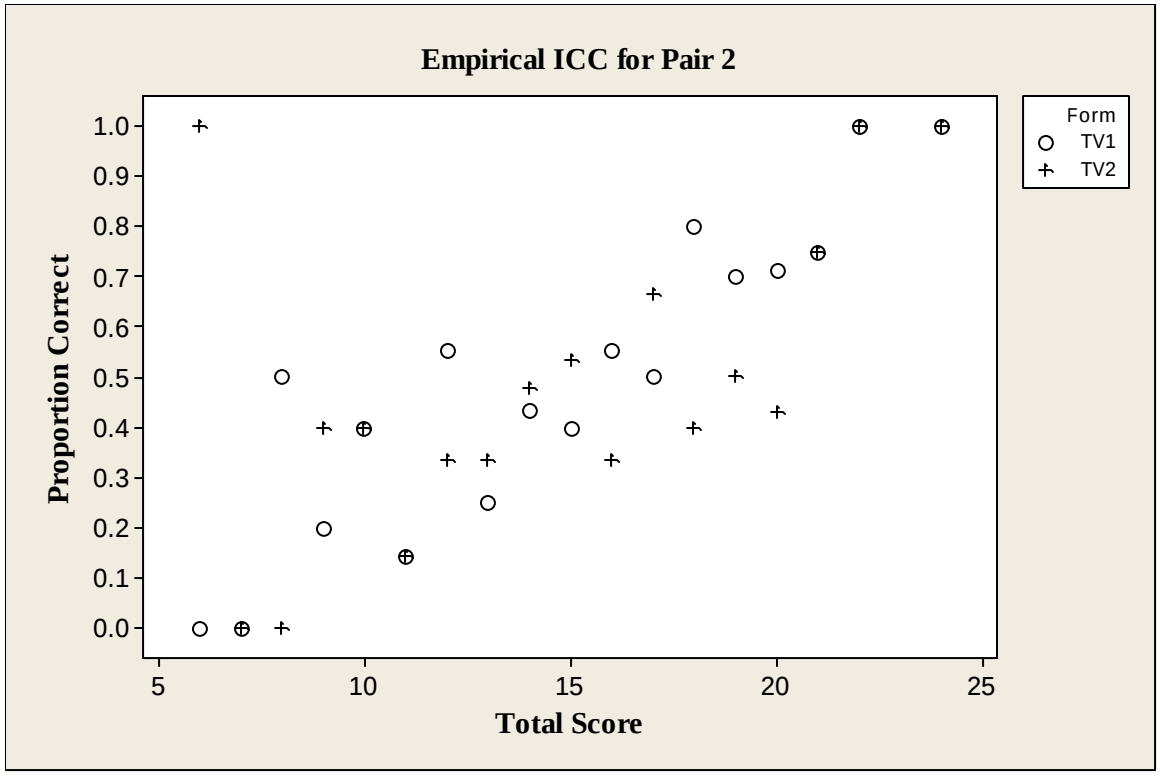
Classical Item Analysis

	TV1 Item		TV2 Item	
Pair Number	**Item difficulty (p-value)**	**Item discrimination (point biserial)**	**Item difficulty (p-value)**	**Item discrimination (point biserial)**
1	0.94	0.14	0.65	0.12
2	0.50	0.23	0.44	0.00
3	0.95	0.01	0.92	0.02
4	0.45	0.07	0.64	0.16
5	0.67	0.17	0.66	0.30
6	0.33	0.28	0.45	0.26
7	0.43	0.08	0.78	0.10
8	0.76	0.10	0.43	0.25
9	0.51	0.09	0.46	0.19
10	0.49	0.08	0.45	0.30
11	0.76	0.06	0.61	0.05
12	0.32	0.08	0.24	-0.03
13	0.46	0.09	0.42	0.28
14	0.48	0.16	0.25	0.23

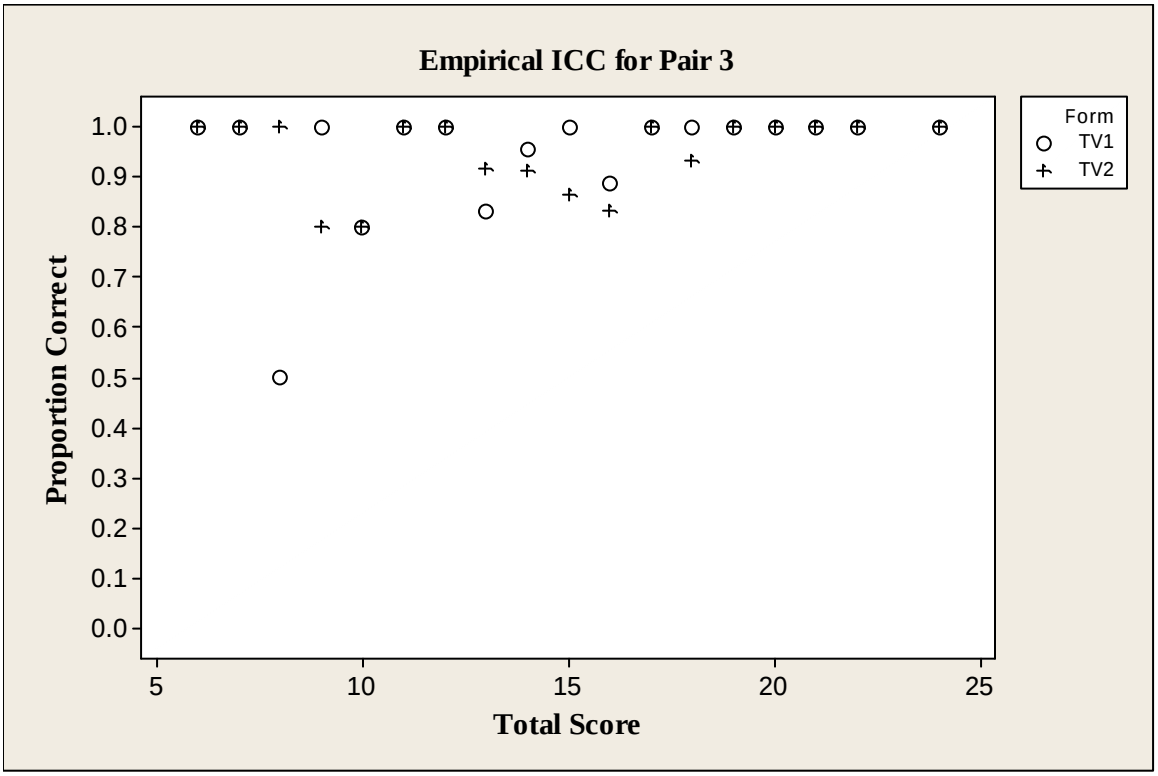
APPENDIX I

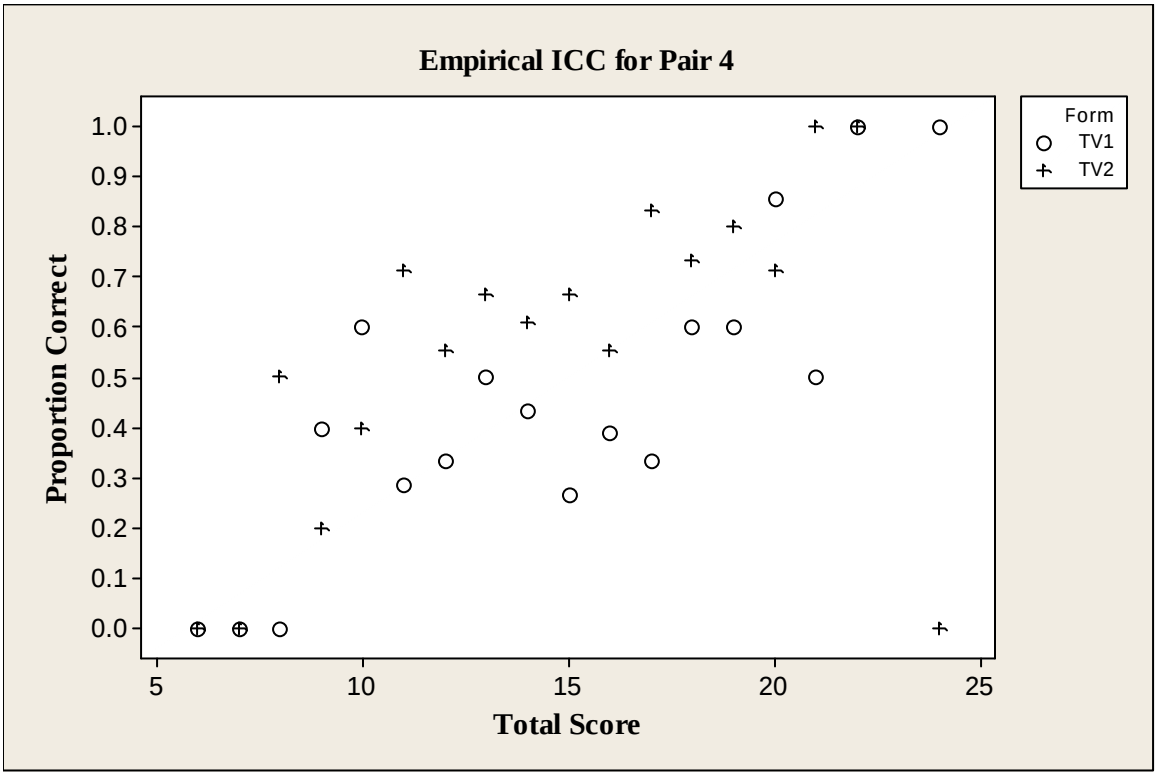
Empirical Item Characteristic Curves of Item Pairs



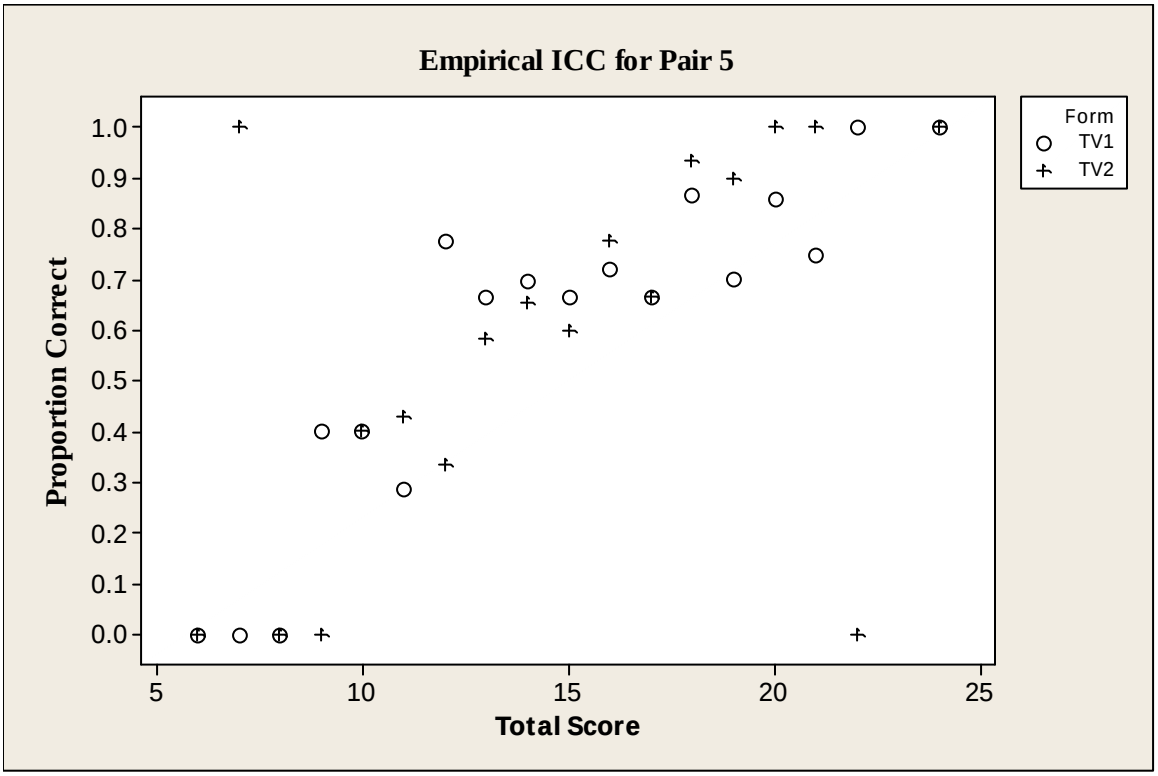


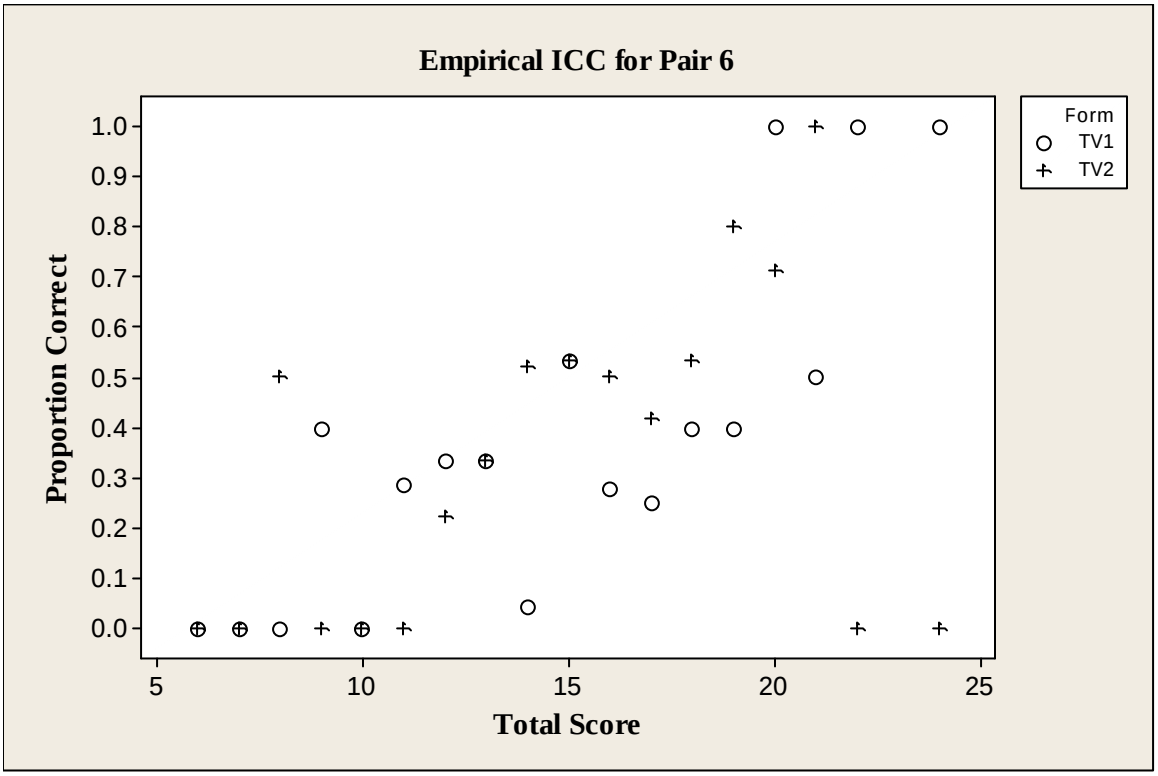
APPENDIX I (continued)



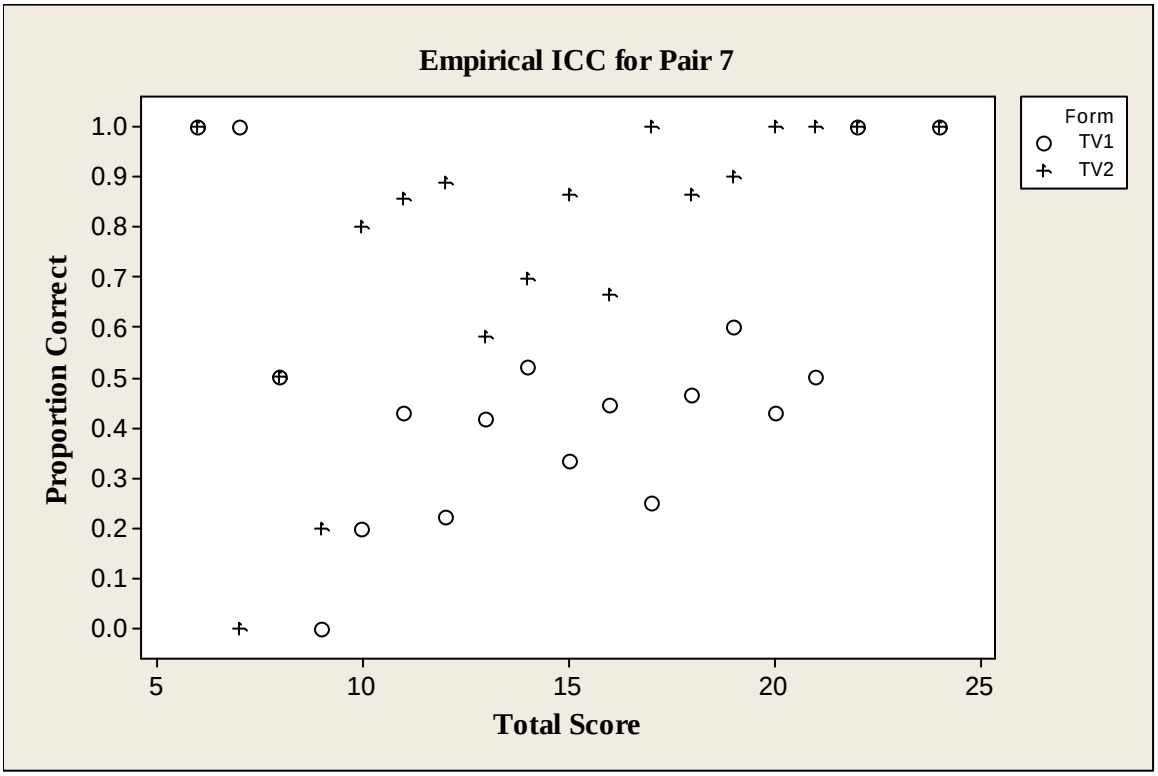


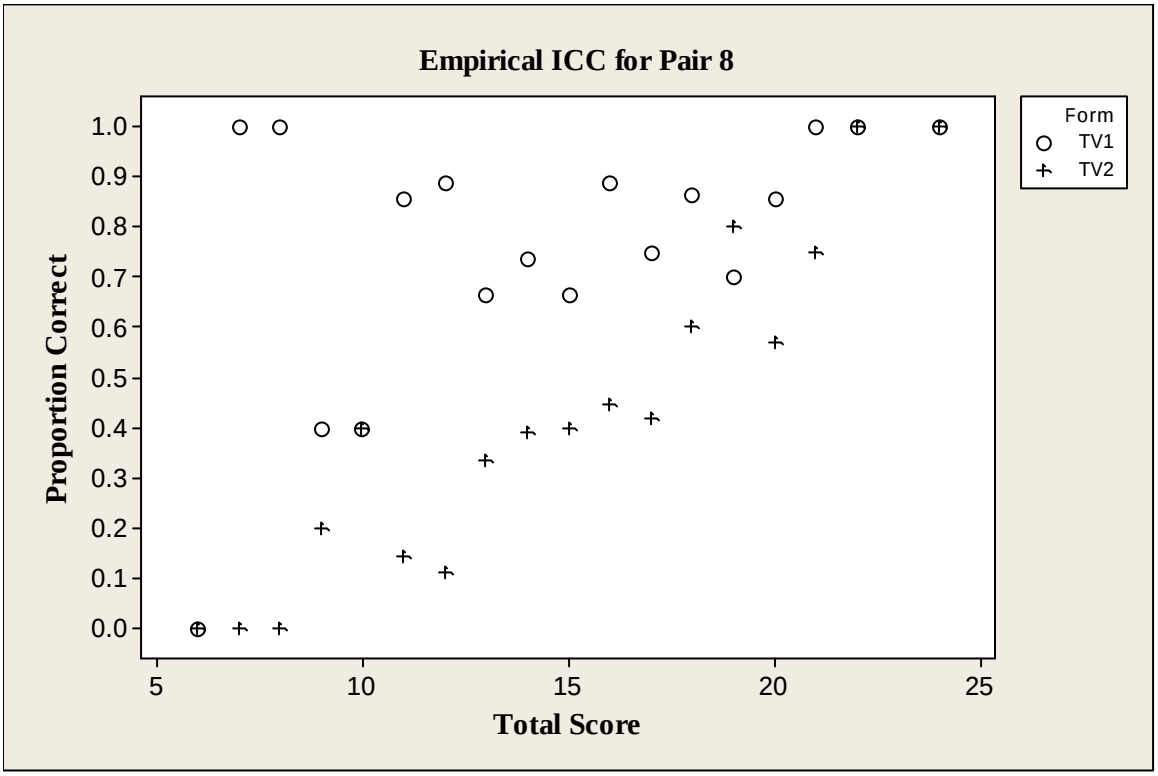
APPENDIX I (continued)



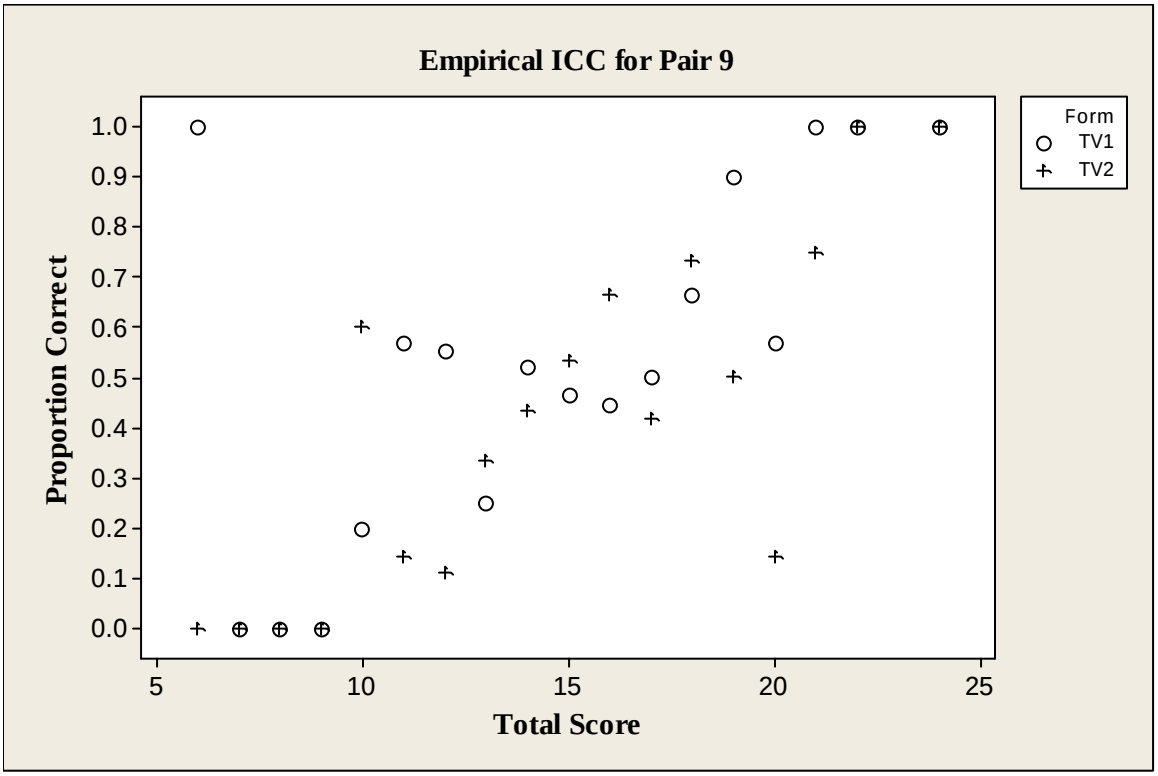


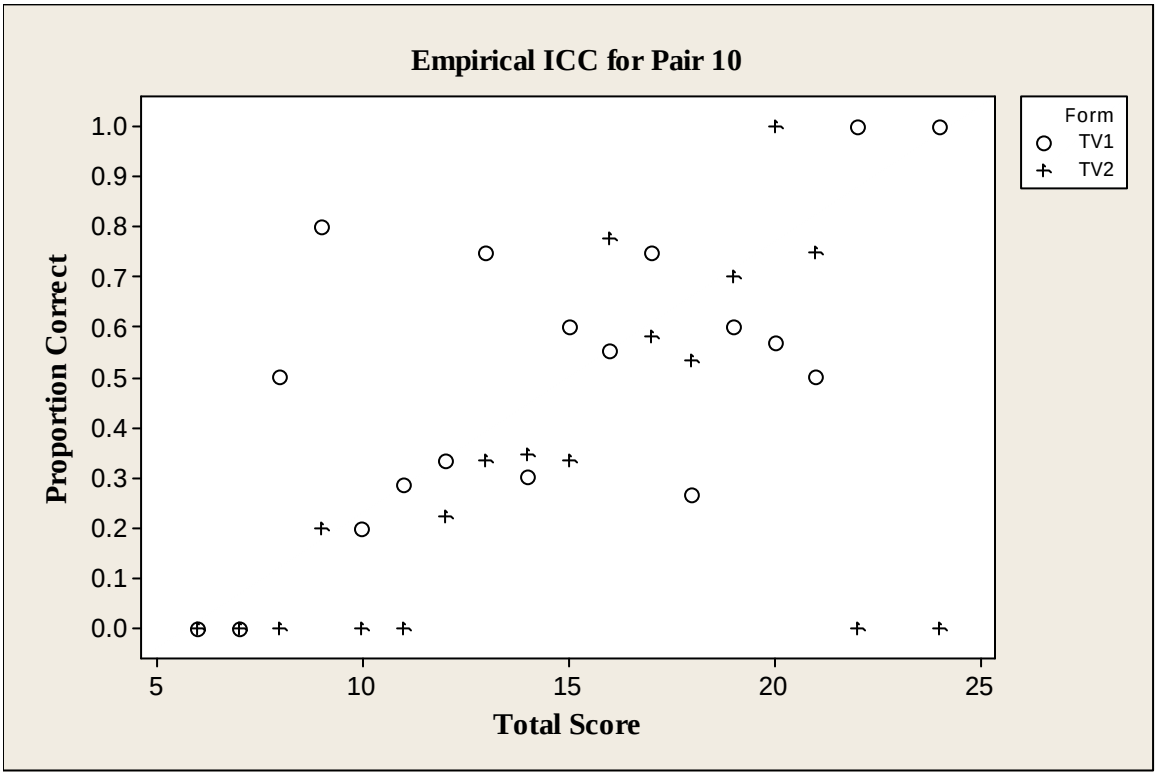
APPENDIX I (continued)



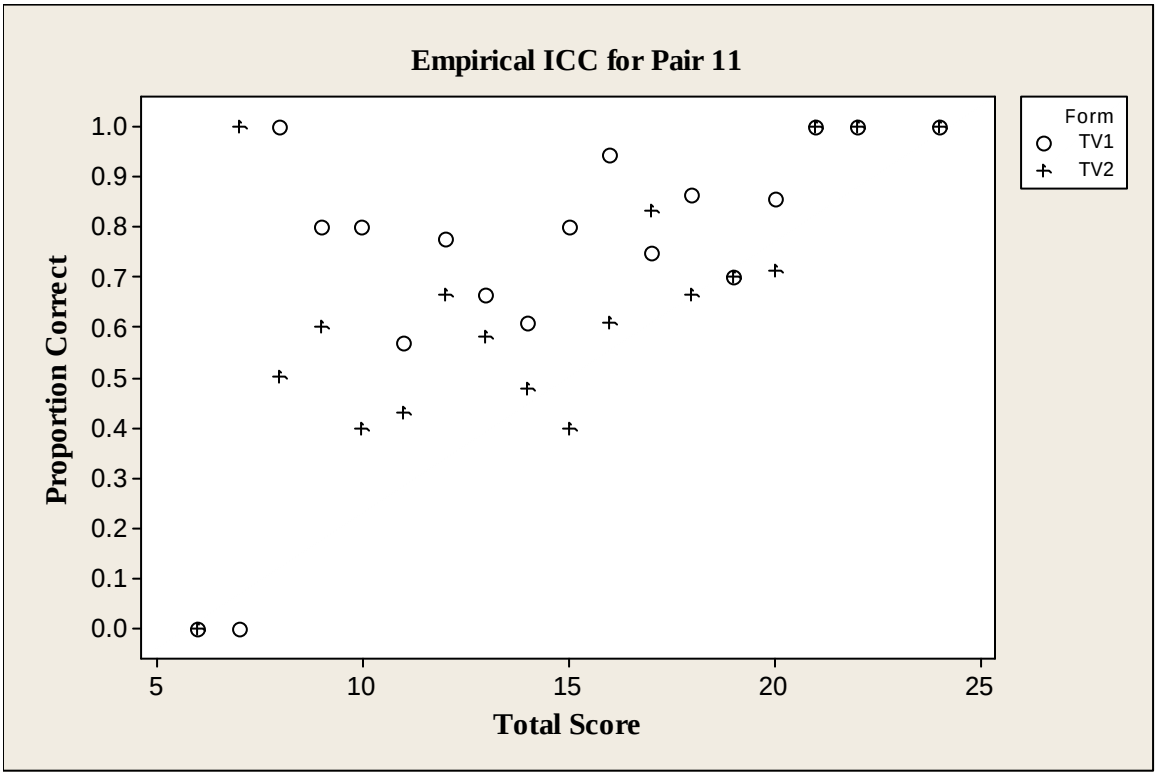


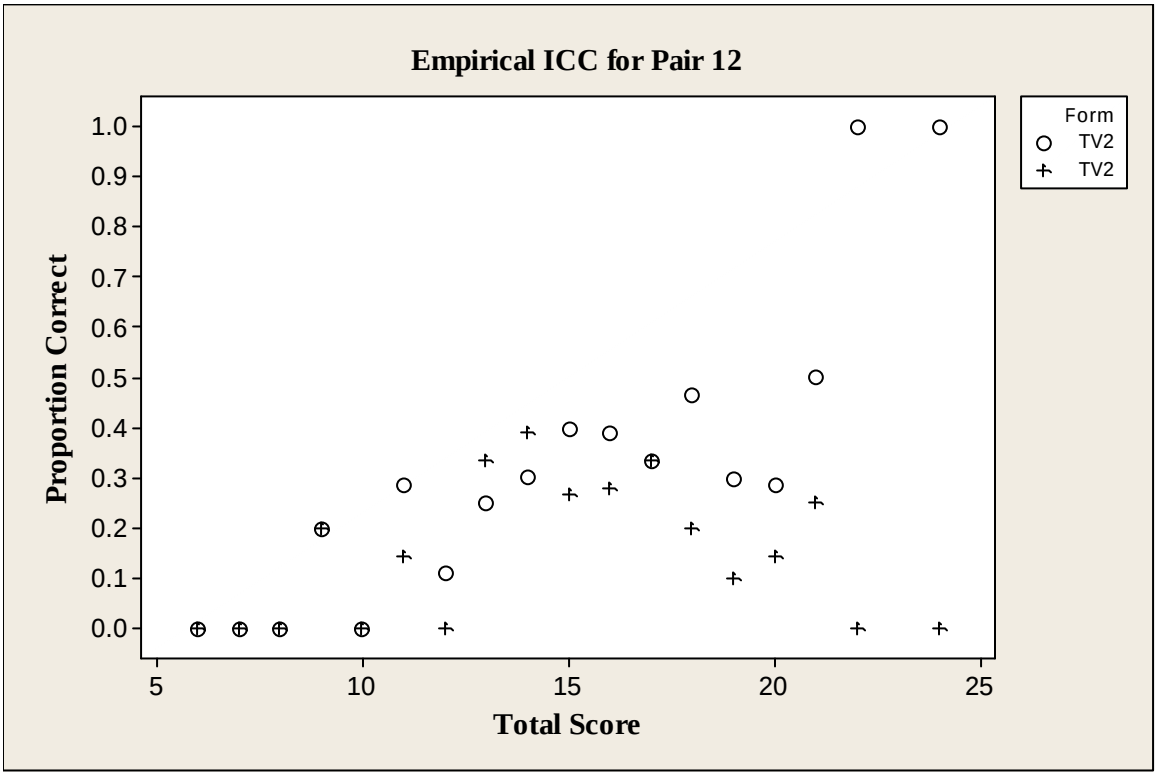
APPENDIX I (continued)



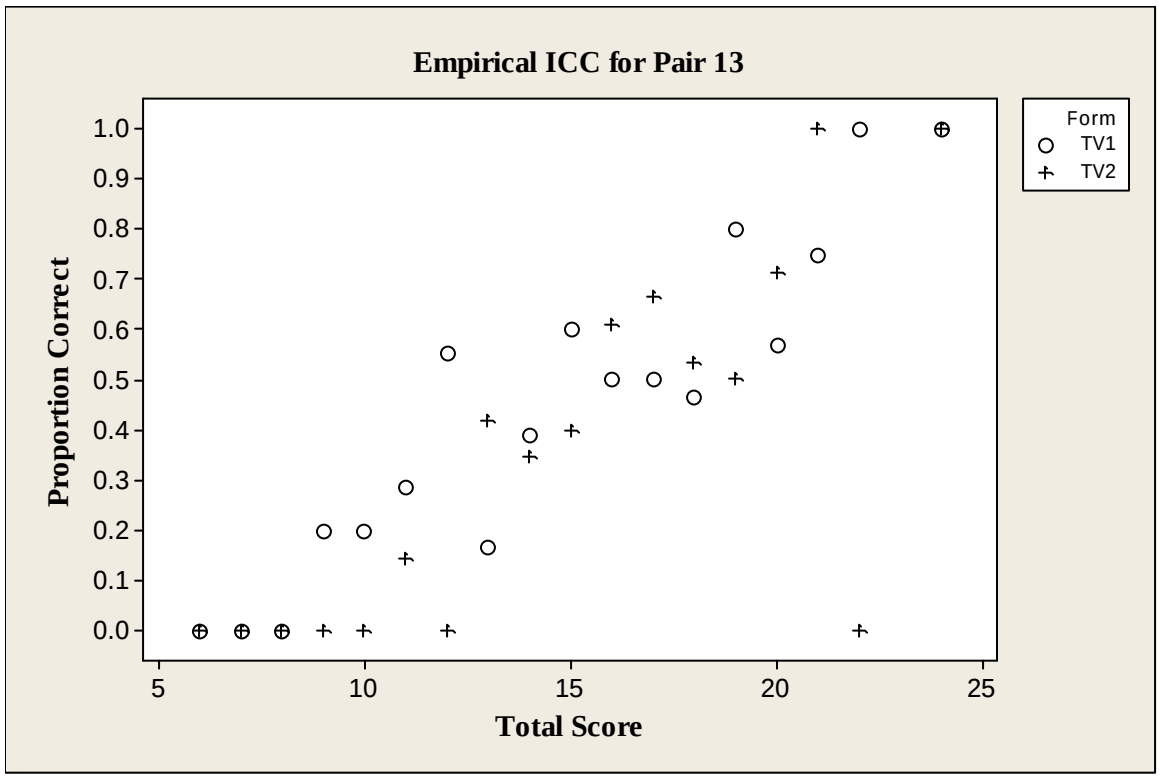


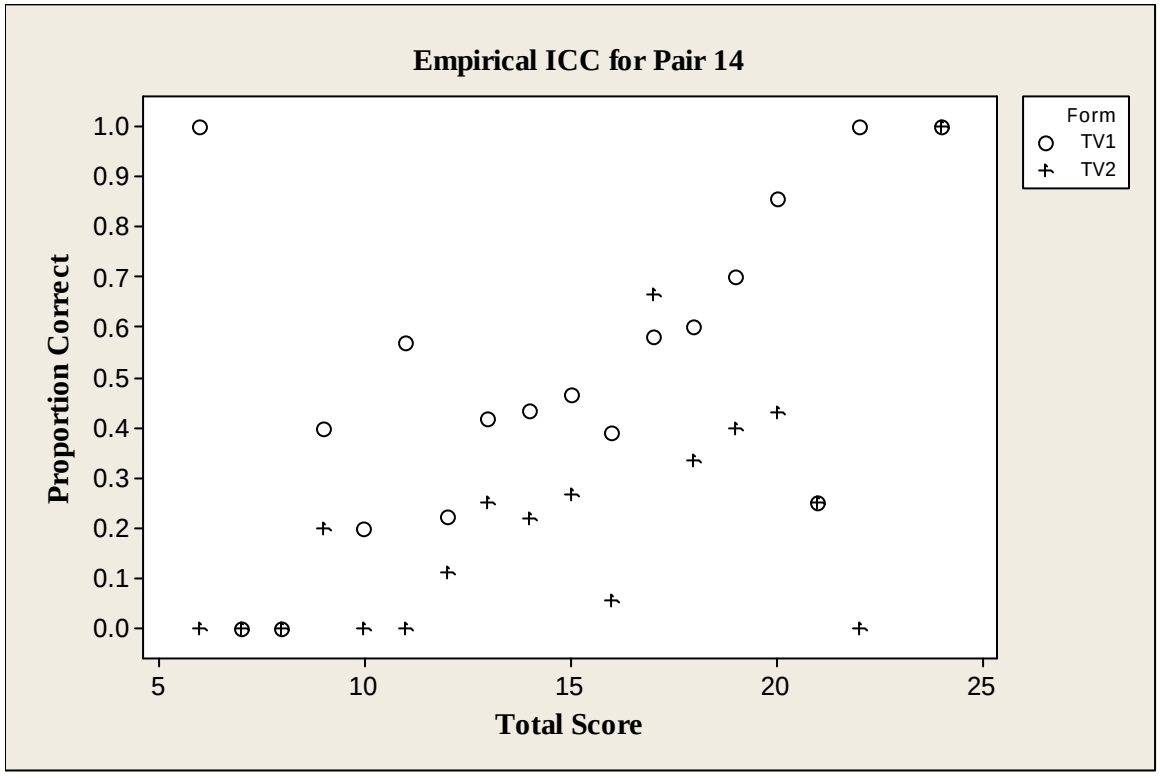
APPENDIX I (continued)





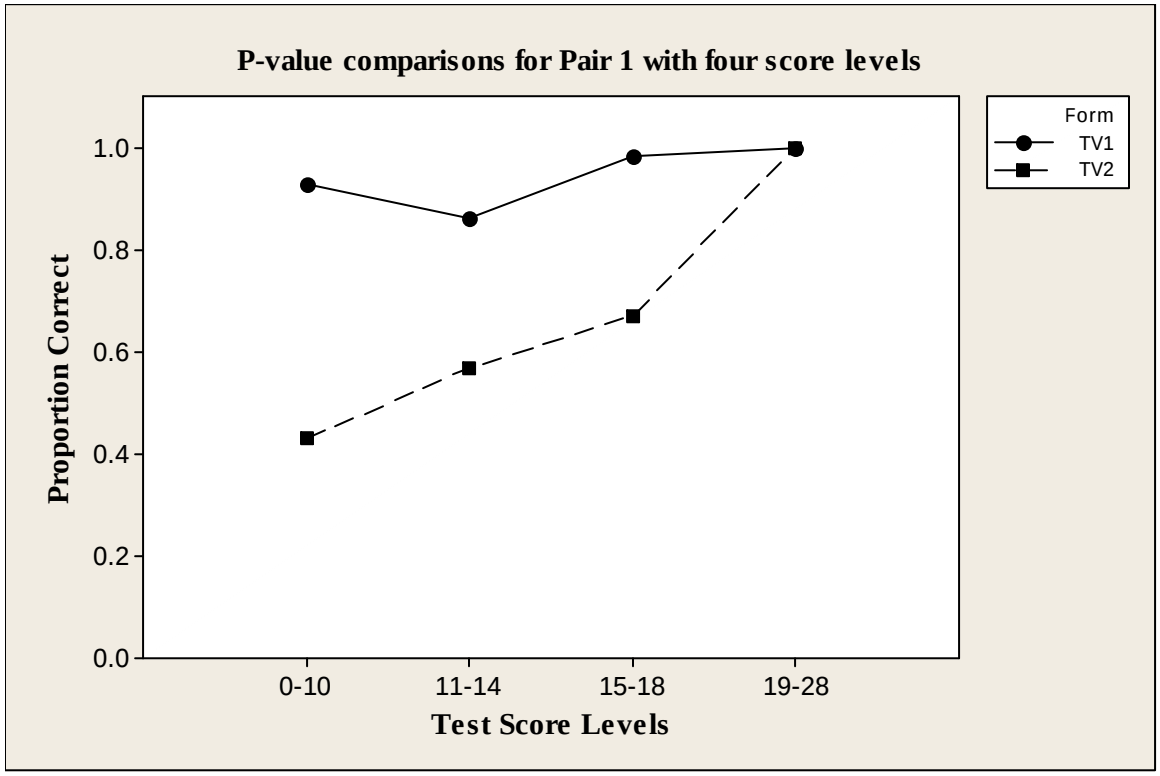
APPENDIX I (continued)

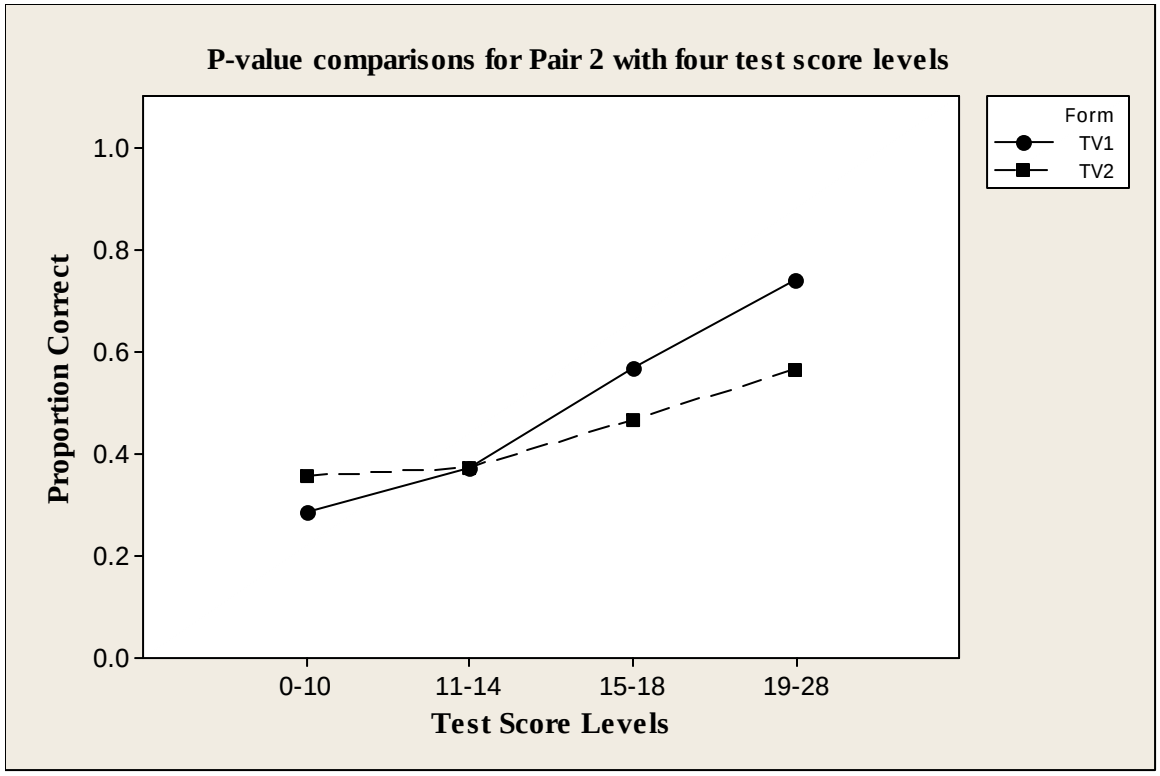


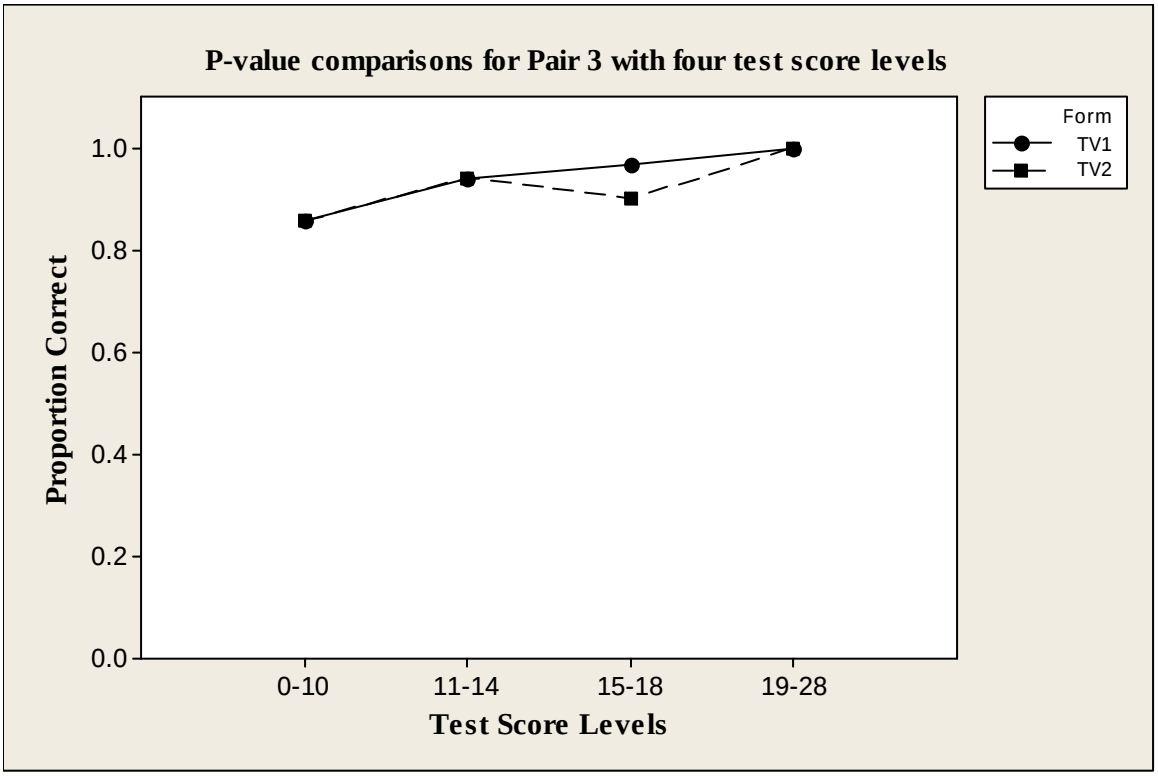


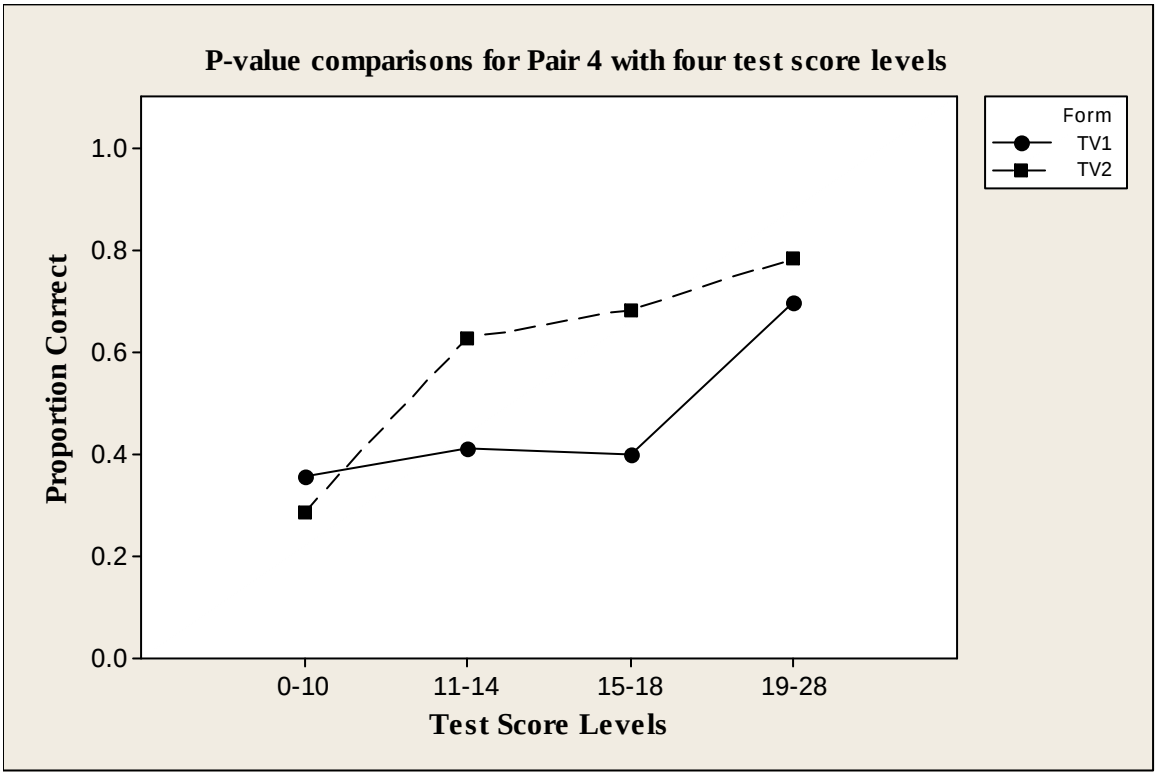
APPENDIX J

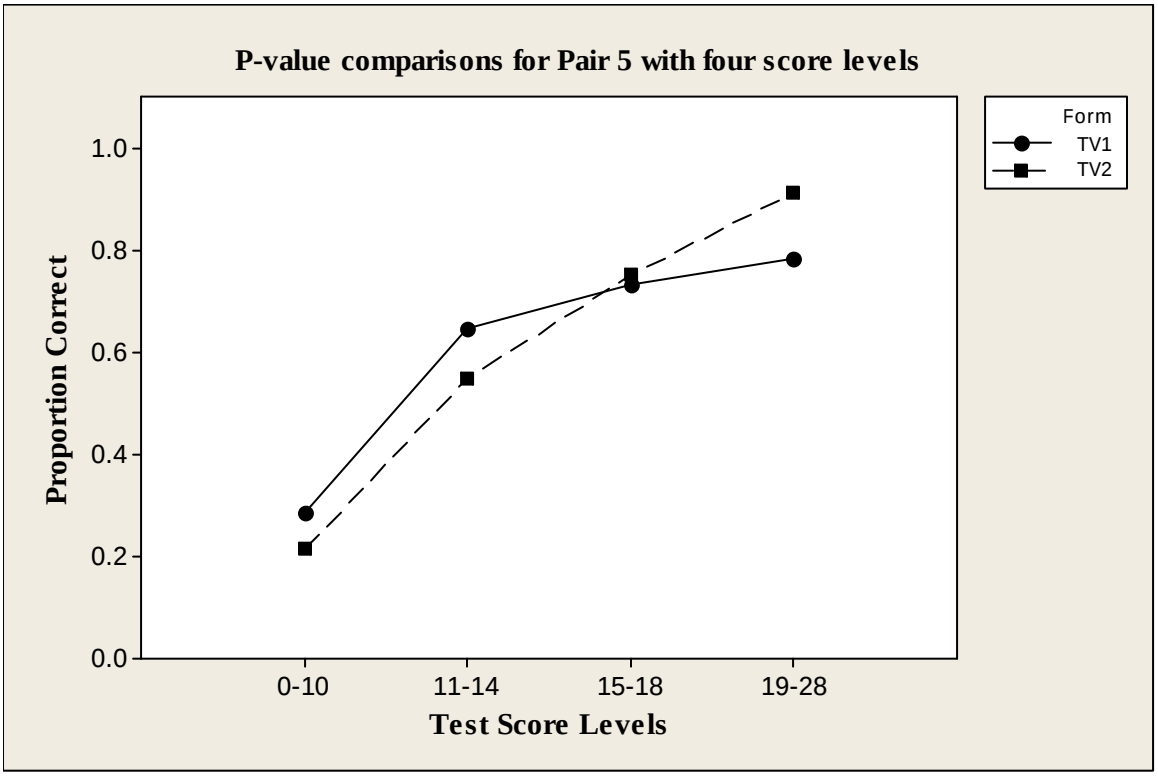
Conditional *p*-value plots with four score levels

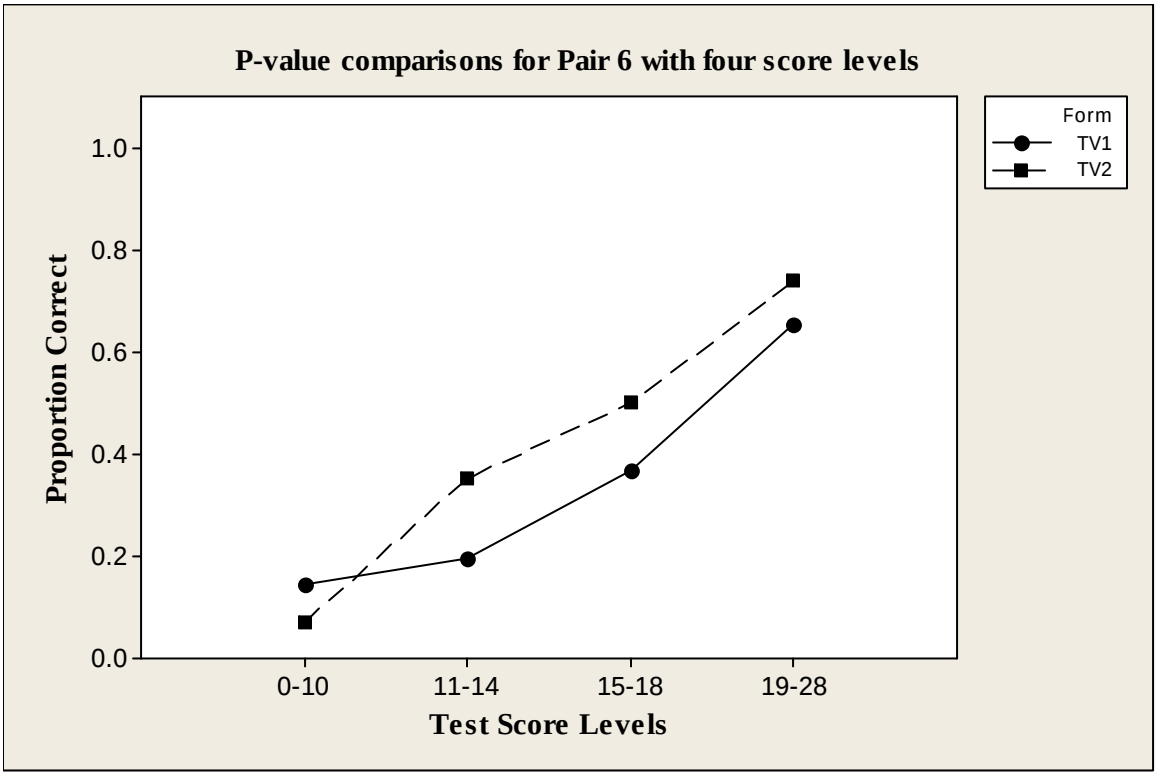


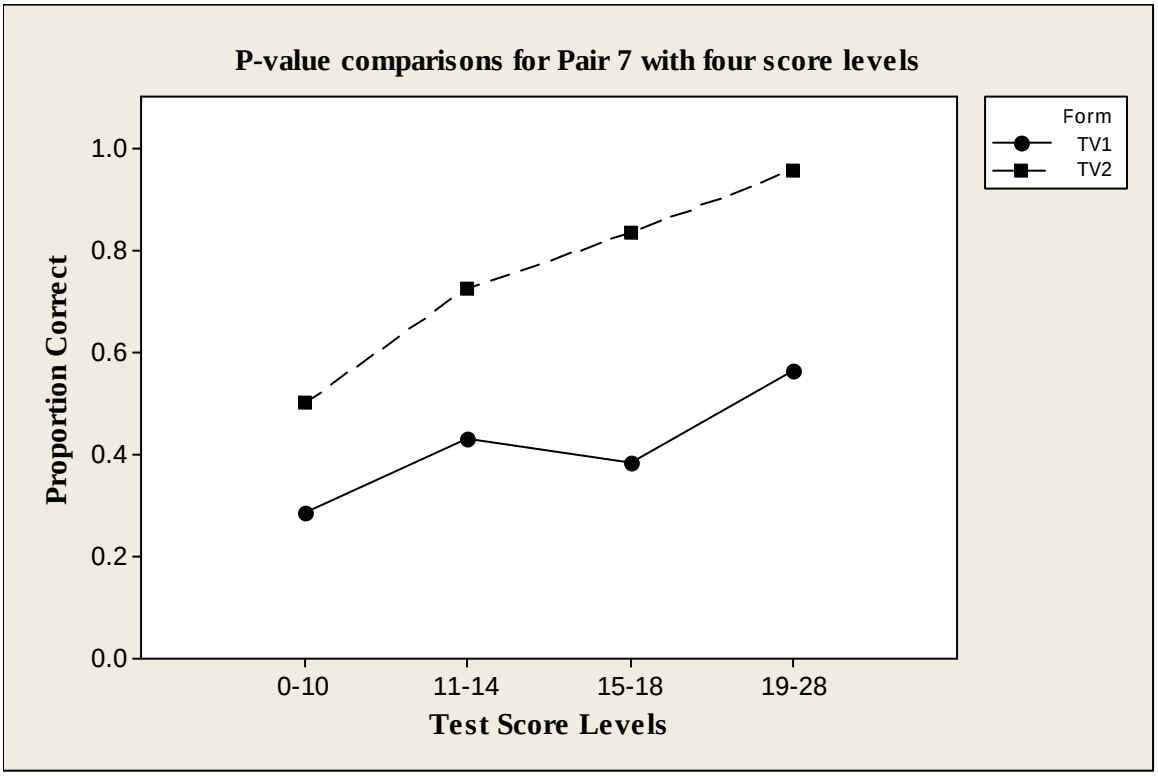


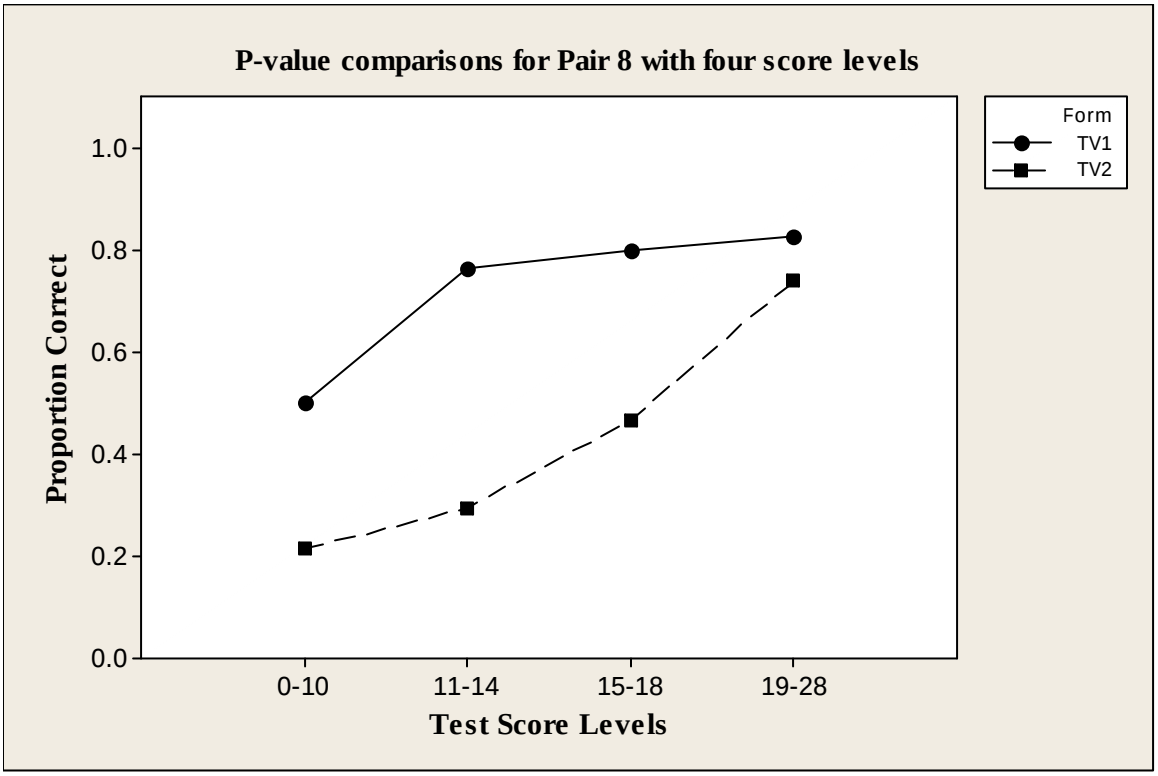

P-value comparisons for Pair 3 with four test score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels


P-value comparisons for Pair 4 with four test score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels

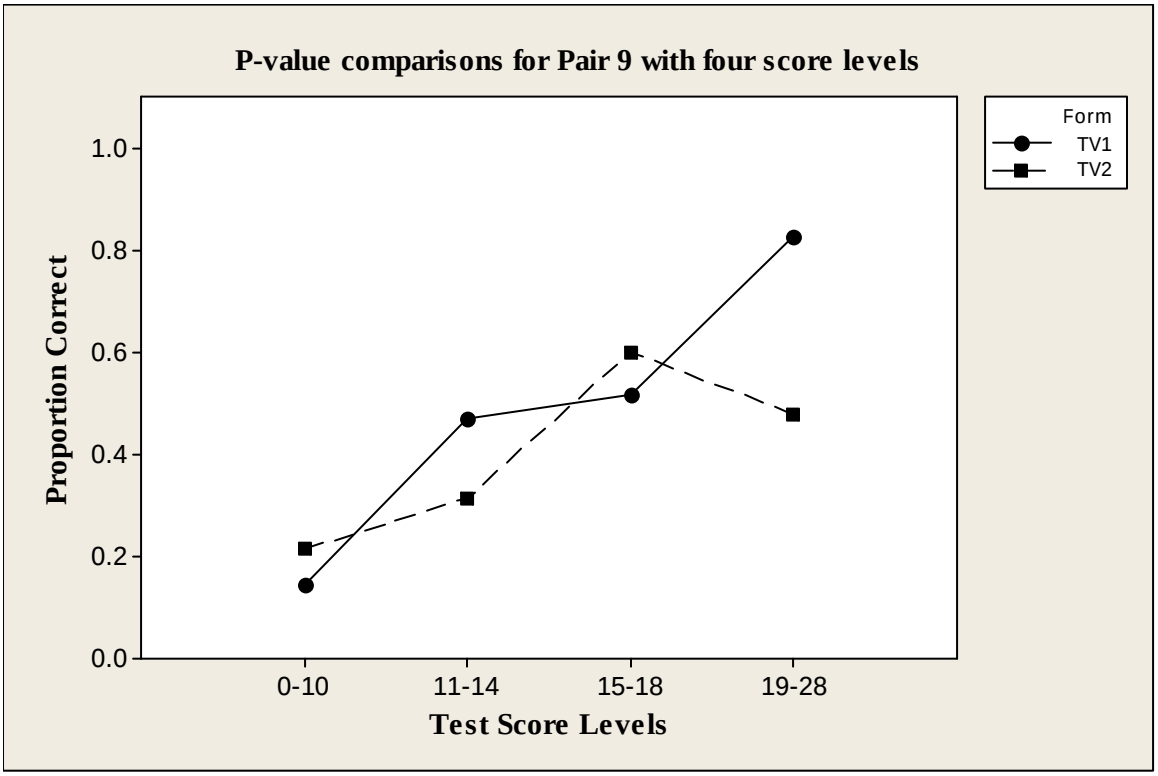

P-value comparisons for Pair 5 with four score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels

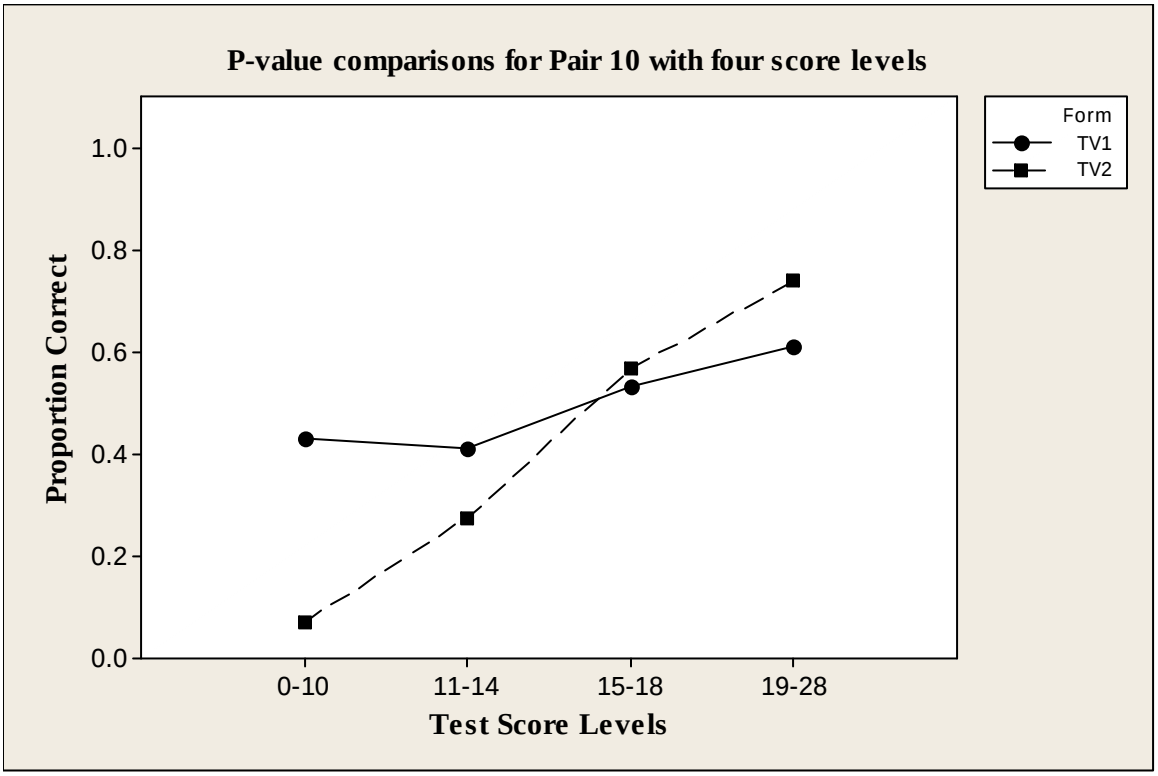

P-value comparisons for Pair 6 with four score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels


P-value comparisons for Pair 7 with four score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels

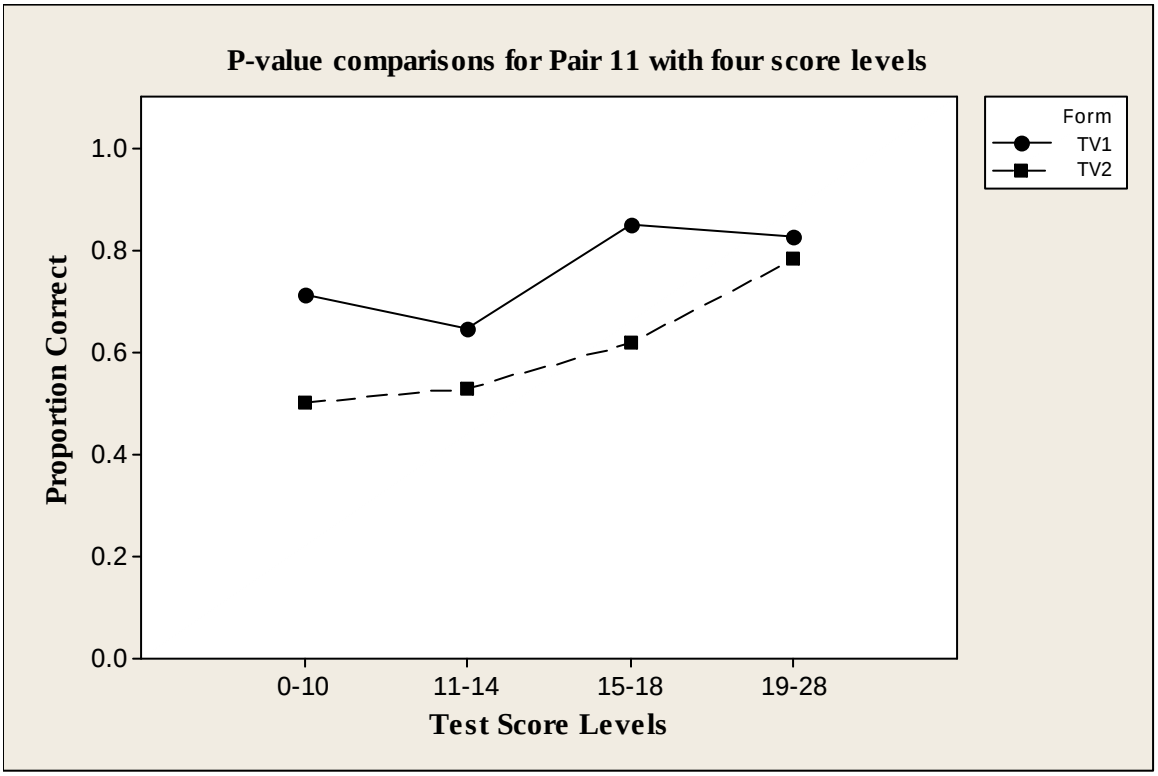

P-value comparisons for Pair 8 with four score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels

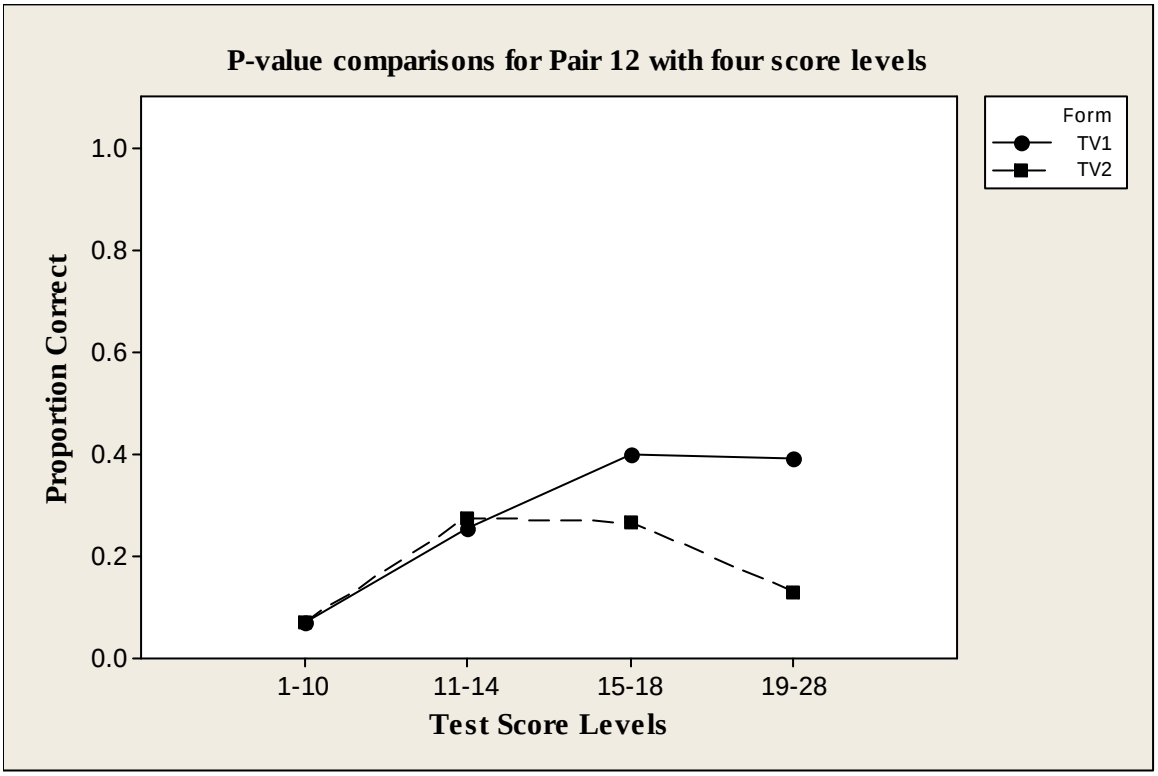
APPENDIX J (continued)

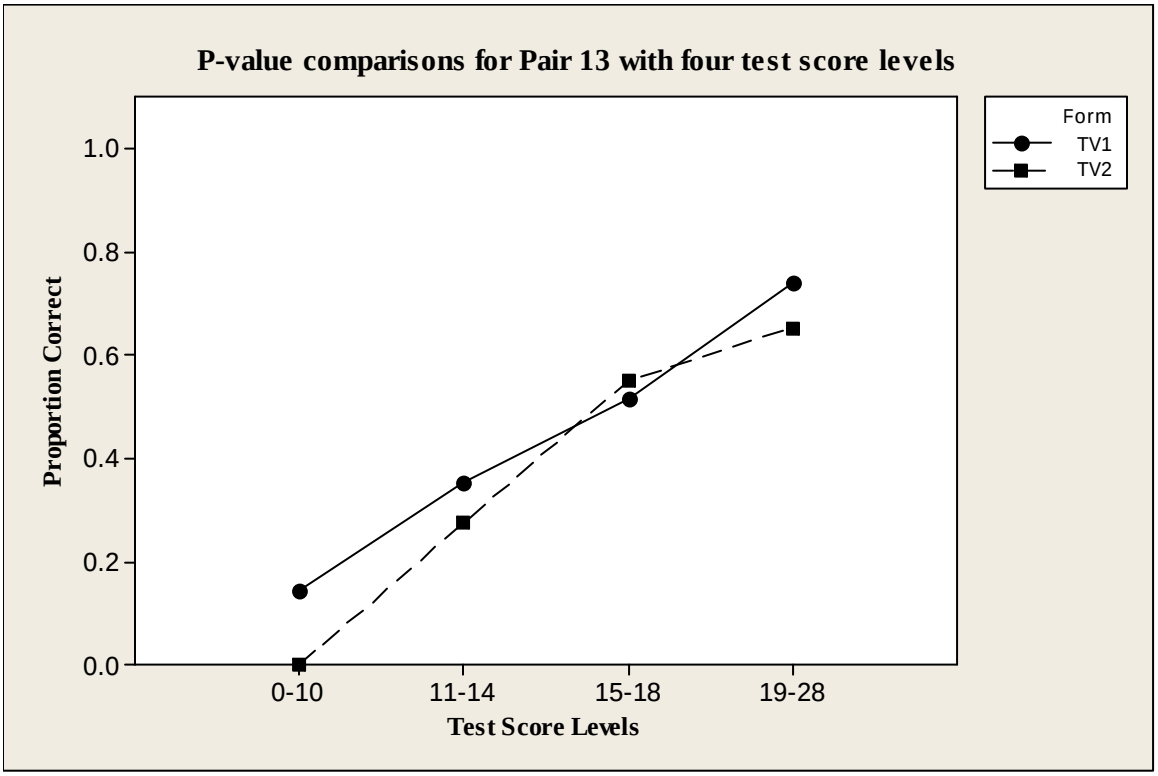


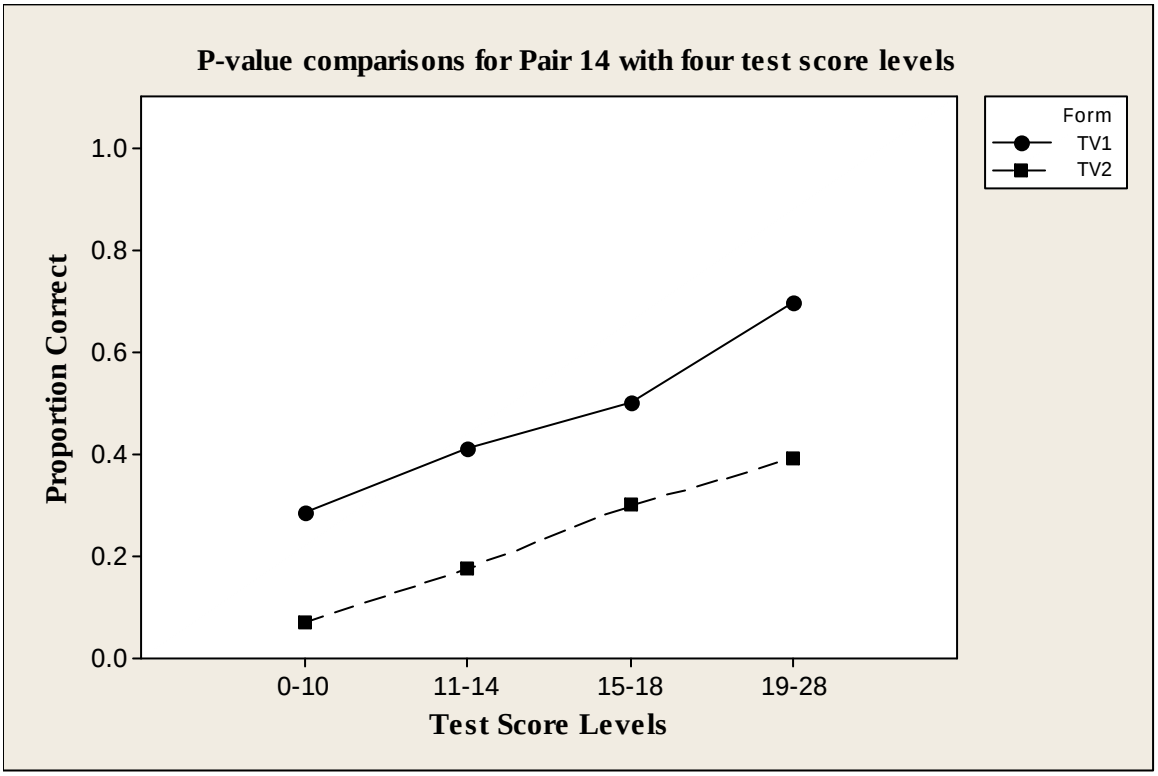


APPENDIX J (continued)






P-value comparisons for Pair 13 with four test score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels


P-value comparisons for Pair 14 with four test score levels
Form
TV1
TV2
Proportion Correct
1.0
0.8
0.6
0.4
0.2
0.0
0-10
11-14
15-18
19-28
Test Score Levels

APPENDIX K

Summary table for comparison of methods

Pair	Think-aloud	Expert Judgment	Statistical DIF Methods
1	Equivalent	Equivalent	**Nonequivalent**
2	**Nonequivalent**	Equivalent	Equivalent
3	Equivalent	Equivalent	Equivalent
4	Equivalent	Equivalent	Equivalent
5	Equivalent	Equivalent	Equivalent
6	**Nonequivalent**	Equivalent	Equivalent
7	Equivalent	Equivalent	**Nonequivalent**
8	**Nonequivalent**	Equivalent	**Nonequivalent**
9	Equivalent	Equivalent	Equivalent
10	Equivalent	Equivalent	Equivalent
11	Equivalent	**Nonequivalent**	Equivalent
12	Equivalent	Equivalent	Equivalent
13	Equivalent	**Nonequivalent**	Equivalent
14	Equivalent	**Nonequivalent**	**Nonequivalent**

Sarah E. Rzasa Zappe
Vita

EDUCATION

Masters of Science, Educational Psychology (2002)
Penn State University, University Park, PA

Bachelor of Arts, Psychology, Summe Cum Laude (1998)
University of Connecticut, Storrs, CT

WORK EXPERIENCE

Director of Assessment and Instructional Support for the College of Engineering
Penn State University
July 2007 to present

Educational Testing and Assessment Specialist/Research Assistant
Schreyer Institute for Teaching Excellence, Penn State University
July 2004 to June 2007

SELECTED PUBLICATIONS AND PRESENTATIONS

Zappe, S. E., Guertin, L. A., & Kim, H. (April, 2006). Just-In-Time Teaching: A web-based method of integrating classroom assessment and instruction. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Suen, H. K., & Rzasa, S. E. (2003). Psychometric foundations of behavioral assessment. In S. N. Haynes & E. M. Heiby (vol. eds). *Comprehensive Handbook of Psychological Assessment*, *Volume 3: Behavioral assessmen*t, 37-56, Hoboken, NJ: Wiley.

Rzasa, S. E. (April, 2003). Item analysis on a developmental rating scale using both statistical and qualitative methods. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

SERVICE AND AFFILIATIONS

American Educational Research Association
American Society for Engineering Education
National Council of Measurement in Education