The Effects of Different Scoring Methodologies on Item and Test Characteristics of Technology-Enhanced Items

By

Cameron M. Clyne

University of Kansas

Submitted to the graduate degree program in Educational Psychology, and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson: Dr. William Skorupski

_____

Dr. Neal Kingston

_____

Dr. Bruce Frey

_____

Dr. Vicki Peyton

_____

Dr. Wayne Sailor

Date Defended: August 27, 2015

The Dissertation Committee for Cameron Clyne

certifies that this is the approved version of the following dissertation:

THE EFFECTS OF DIFFERENT SCORING METHODOLOGIES ON ITEM AND TEST CHARACTERISTICS OF TECHNOLOGY-ENHANCED ITEMS

_____

Chairperson: Dr. William Skorupski

Date approved: August 27, 2015

Abstract

Technology-enhanced (TE) item types have recently gained attention from educational test developers as a way to test constructs with higher fidelity. However, most research has focused on developing new TE item types, and less on researching best practices for scoring these new item types. The purpose of this study was to analyze the effect of adjusting scoring strategies of TE items on item and test characteristics. Descriptive statistics as well as tests of statistical significance were reported when appropriate. Additionally, figures representing the differences in test information and fit across forms were created to help show consistency in scoring effects. Results were consistent with prior research into differences between dichotomous and polytomous scoring strategies. Results indicate that the two best strategies for scoring TE items are partial-credit scoring and testlet response theory. The worst approach to scoring TE items is to score them as correct-only. Results of this study add to the research literature, as well as provides a practical guide to test developers when deciding which scoring strategy to use with new TE item development.

Acknowledgements

I would like to express my gratitude to Dr. William Skorupski for his time, resources, and instruction towards the completion of my dissertation. I have learned a remarkable amount from you in the past four years, and without your guidance this project would not have been as successful.

Additionally, I would like to thank the other members of my committee, Drs. Kingston, Frey, Peyton, and Sailor. Your support and input throughout this process has been invaluable to my education and my future in educational psychology.

Thank you to my family. Throughout my years of education, you have taken the time to support me in all my educational endeavors. Without my parents as role models, I would never have strived to reach the peak of educational achievement.

Finally, to my wife Colleen Clyne, thank you for your unconditional love and support. Without you by my side, this would never have been possible. All those long days of study, weekends of writing, and times I had to be absent were made possible by your encouragement. You are the reason I push myself to excel, and I thank you for being there every step of the way. This accomplishment is as much yours as it is mine.

Table of Contents

List of Equations

List of Tables

List of Figures

Page left intentionally blank

Chapter One: Introduction

In the 21st century, standardized testing has become a ubiquitous facet of the American education system. The history of standardized testing stretches back to the Han Dynasty, when testing was used to help select civil servants (Black, 1997; Madaus & O'Dwyer, 1999). Throughout modern history, assessments have been used to help select, diagnose, and gather information. Today, we use standardized testing for everything from college admissions to job applications.

E. L. Thorndike, a professor at Columbia University, started conducting research on more objective forms of testing. Thorndike believed that testing should be used by our society to identify and segregate the intellectual students from the general population (Gallagher, 2003). In 1913, Thorndike was quoted as saying:

> Educational Agencies are a great system of means not only of making men good and intelligent and efficient but also of picking out and labeling those who for any reason are good and intelligent and efficient…They help society by providing it not with better men but with the knowledge of which men are good. (Gallagher, 2003).

Thorndike's suggestions convinced schools in Pennsylvania, New Jersey, New York, Massachusetts, Michigan, Kansas, and California to begin using standardized measurement tools.

Not only have the uses of standardized testing expanded throughout history, but also our knowledge of the science of testing. The standardization of tests has led to an adjustment in the item types utilized in these assessments. As early as 1919, Chapman and Toops wrote about the benefits of multiple-choice items for bricklayers. Since that time, the use of multiple-choice items has exploded throughout the testing world. In the 1990's, multiple-choice assessments

began to receive criticism for the lack of improvement in student learning and educational outcomes (Osterlind, 1997). With the advent and wide-spread use of computers, assessment practices have begun to change (Zenisky & Sireci, 2002). Computer advances have influenced test construction, administration, scoring, and score reports. Additionally, a new item type has started working itself into the zeitgeist of the testing culture. With the prevalence of computers, test developers have been able to move away from the traditional paper-and-pencil assessment towards computer-based methods for administration (Bartram & Bayliss, 1984). As of 2009, Quellmalz and Pellegrino (2009) reported that more than 27 states had begun to pilot or provide operational assessments online. These first steps towards using computer-based assessments have laid the groundwork for the use of today's technology-enhanced (TE) item types.

Technology-enhanced item types were defined by Parshall, Spray, Kalohn, and Davey (2002) as "items that depart from the traditional, discrete, text-based, multiple-choice format." Technology-enhanced item types have gained attention from test developers as a new way to potentially improve measurement efficiency through reduced guessing or by more directly measuring the construct of interest (Parshall et al., 2002). Additionally, Sireci and Zenisky (2006) believe that these new item types are a way to improve the fidelity of assessments, and that they might be able to measure constructs not measurable using multiple-choice item types. Test fidelity refers to the degree to which a test simulates real world contextualized knowledge and skills (Lievens & Patterson, 2011). For example, flight simulations have high fidelity compared to multiple-choice items. Other commonly used terms used in place of "fidelity" are "performance" and "authentic." These new item types, along with the emergence of performance assessments, have started a movement to do away with the multiple-choice assessment (Osterlind, 1997). This movement has already influenced the perception the field and the public

have on multiple-choice assessments, and thus have begun the path away from multiple-choice assessments towards presumed higher fidelity assessments (Madaus & Dwyer, 1999).

The downside to the swift acceptance of new item formats is the inattention to evidence of validity and the effects that these new formats have on testing (Osterlind, 1997).

Due to the relatively new nature of TE items, only a small body of literature currently exists. In 2001, Huff and Sireci commented that too much focus has been spent on the development of different kinds of TE items and not enough time on the validation of the item types we currently have. Indeed, a literature search comes up with very few articles describing the validation or the comparison of psychometric properties of these newer TE item types with the traditional multiple-choice item type. Though very little research has been conducted, many test development companies are proceeding with the development of assessments with TE item types included. Not only has the validity not been fully researched, TE item types also tend to be expensive to develop, an important factor in deciding whether they are worth adding to an assessment.

Currently, there exists a large hole in the understanding of the psychometric properties of TE items. One of the most important areas for focused research is in the scoring of these item types. The full range of scoring options for TE items must be researched to determine best practice. This dissertation aims to contribute to the literature on the best methods for scoring TE items when using both Classical Test Theory (CTT) and Item Response Theory (IRT). Specifically, what are the effects of different scoring methodologies on the reliability of TE items?

Chapter Two: Literature Review

Review of Technology-Enhanced (TE) Item Types

With the boom in technology, educational assessment has begun to explore new ways to test various constructs. The recent move toward the Common Core has amplified the need for higher fidelity assessments that move beyond the now traditional multiple-choice assessments (Riddile, 2012). With the need to move away from strictly multiple-choice assessments, there is an increasing need to reexamine test development to incorporate these new item types (Zenisky & Sireci, 2002). Technology-enhanced item types utilize a new item format. An item format encompasses everything an examinee must complete to respond to an item (Sireci et al., 2006). Technology-enhanced items often require test takers to supply, develop, perform, or create content (Osterlind, 1997). Due to these new item types, test developers will have to reexamine how they handle data entry, reporting, analysis, and scoring of items.

Technology-enhanced item types may improve upon the standard multiple-choice item by overcoming their classic shortfalls, such as easily obtained correct answers through guessing, certain domains being difficult to assess, and low construct fidelity (Sireci & Zenisky, 2006). With recent advances in technology, test developers are creating new item types that allow students to interact with the actual item when responding (Example items can be found in Appendix A). For example, a student may be asked to sort geometric shapes by classification by dragging the items into categories on screen. This utilization of technology allows for more direct interaction with the desired construct and can help avoid some of the problems mentioned above with traditional multiple-choice assessments.

One of the main difficulties with researching these TE item types is that they consist of different methods for utilizing technology. Various researchers in this field have created taxonomic models for how to classify the various ways in which these TE items can be created (Sireci & Zenisky, 2006; Bennett, Ward, Rock, & LaHart, 1990; Scalise & Gifford, 2006). This paper will use the Scalise & Gifford (2006) taxonomy to discuss item types. This taxonomy fits well with the item types utilized in this study, and is the most modern of the taxonomies. In addition to Scalise & Gifford, other taxonomies have been created to help categorize new item types. These taxonomies will be described below.

Classification and Description of Technology-Enhanced Item Types

Item type taxonomy. This paper focuses on TE item types as a whole, treating each different type as a subtype of a larger group. Other researchers in this field have focused on what makes each subtype different, effectively creating different taxonomies. These taxonomies provide an easy way to discuss technology-enhanced item types.

Classically, there were two primary ways to divide item types: selected response and constructed response (Sireci & Zenisky, 2006). As more item types are created, this becomes too simplistic. A more complex taxonomy was developed by Bennett et al. (1990) and classifies items based on the extent of openness allowed in the item response. Bennett et al. (1990) created seven categories along this openness continuum in which items could be placed. These seven categories are multiple-choice, selection/identification, reordering/rearrangement, substitution/correction, completion, construction, and presentation/performance. Parshall, Davey, and Pashley (2000) built a taxonomy for technology-enhanced item types that utilized five dimensions. These five dimensions are item format, response action, media inclusion, level of interactivity, and scoring algorithm. Scalise & Gifford (2006) later used the categories proposed

by Bennett et al. (1990) as a piece of their item taxonomy. Scalise & Gifford (2006) placed these same categories on constraint of the item response. The constraint of the item response is determined by how much construction of the answer choice a test taker utilizes to solve a problem. On one end of the spectrum is a fully selected item type (e.g., multiple-choice), and on the other end of the spectrum are fully constructed item types (e.g., portfolio). For example, if a student has to write an essay response to a test question, this is considered a far less constrained item than if we give a student four options from which to choose. Within each category, items can be further classified by complexity. For example, within the multiple-choice category, the least complex item type is the true/false item; the most complex item within this category, is a multiple-choice item with media distractors. Scalise & Gifford (2006) call this taxonomy the Intermediate Constraint Taxonomy for E-Learning Questions and Tasks. This taxonomy is represented in Figure 1. Moving from left to right and top to bottom, item responses become less constrained and more complex. The bottom right of this figure shows the diagnosis and teaching item type, which is an example of the least constrained but most complex item type featured under the presentation/portfolio category.

Most Constrained → *Least* Constrained

Less Complex ↓ *More* Complex	*Fully Selected*	*Intermediate Constraint Item Types*				*Fully Constructed*
	1. Multiple Choice	**2. Selection/ Identification**	**3. Reordering/ Rearrangement**	**4. Substitution/ Correction**	**5. Completion**	**6. Construction**	**7. Presentation/ Portfolio**
	1A. *True/False* (Haladyna, 1994c, p.54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed* (Parshall et al, 2002, p. 87)	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
	1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
	1C. *Conventional or Standard Multiple Choice* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
	1D. *Multiple Choice with New Media Distractors* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) & *Automated Editing* (Breland et al, 2001, pp.1-64)	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

Figure 1. Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks (Scalise & Gifford, 2006).

The key to using this taxonomy for TE item types is the middle aspect of this scale. The categories that fall in the middle (selection/identification, reordering/arrangement, substitution/correction, and completion) are where the majority of the technology-enhanced item types used in this paper would be classified. Utilizing this taxonomy, 28 item types can be classified.

Technology-enhanced item types. The above taxonomies have attempted to categorize various item types to provide ease of identification. The Scalise & Gifford taxonomy can help

identify various technology-enhanced item types. These items can range from highlighting text, clicking on graphics, or dragging, moving, or reordering objects (Parshall et al., 2000). The technology-enhanced item types described in this section are similar to those that will be used in this research paper.

Scalise & Gifford (2006) described an item type that would fall under the reordering/rearrangement category with mid to low complexity. This item type is called a categorizing item type, and asks test takers to categorize an object into an appropriate parent class. For example, test takers may be asked to categorize a mathematical equation into a family of similar equations (e.g., linear or Quadratic). In this example, the responses would be different mathematical equations, and the bucket would be a box labeled as linear or Quadratic. Similar items may ask test takers to categorize items into multiple categories: for instance, a set of numbers to be placed into either a rational or irrational bucket. To increase difficulty and complexity, some options may be listed that do not fit into either category.

Another item that falls into the reordering/rearranging category is similar to the previous item, except that after the test taker categorizes the item, he/she then has to rank order the items within the category. For example, a test taker may have to choose from a list of words, some of which are pieces of the biological taxonomy, and then order them from broadest to narrowest. This additional step may help tap into deeper cognitive complexity.

Davey, Godwin, & Mittelholtz (1997) described an editing item type where test takers review a passage for various errors. Once an error is identified, the test taker can click on the sentence, and is given a set of alternatives. Once an alternative is selected, that new text is inserted into the original passage so that the test taker can decide if that was the best option. This

item type would fall in the selection/identification category of the Scalise & Gifford taxonomy. A similar item was utilized in this study, and will be further described in the methods section.

Another common TE item is the figural-response item type. This item type can take various forms, but all require the test taker to manipulate a graphical element of an item. This may occur by adjusting a graph, or selecting a hot spot on a graphic, or by dragging labels to designated spots on a graphic (Martinez, 1991; Wan & Henly, 2012; Parshall et al., 2000). Within this study, figural-response item types required test takers to drag words to label various aspects of a graphic. For example, an item might ask a student to label the various aspects of a barn. In this version of figural response, the taxonomy categorization would be selection/identification.

Parshall et al. (2000) described another item that appears in this study. This item would be categorized as reordering/rearranging. Test takers are asked to order a set of items according to a specified rule. For example, test takers may be asked to order the planets of our solar system from closest to the sun to farthest.

An additional item type utilized in this study is a matching item type (Scalise & Gifford, 2006). Test takers are asked to match a word on the left of the screen with the possible response on the right side of the screen. For example, students might be asked to match a word to its part of speech.

Another item that appears in this study is called a select-text item. These items are most similar to a substitution/correction item type (Scalise & Gifford, 2006). Test takers are asked to scan through a passage and select the text that matches the question asked. For example, test takers would be asked to read a passage, and select the sentence that identifies the main plot of the story.

Finally, Scalise & Gifford (2006) described a single numerical entry constructed-response item type, which is the most open ended technology-enhanced item type that can still be computer scored without complicated algorithms. Test takers are asked a question, and are given a single box in which they can type a response. For example, test takers may be given a simple mathematical word problem, and asked to type an answer to that question in the box below.

In addition to the item types listed in the research above, this study also utilized one other item type not found in the research. This is a Punnett square item type. A Punnett square helps to determine the probability of certain genetic outcomes of offspring. In this item type, test takers are asked a question that contains information about the mating pairs, and then are asked to complete the Punnett square. This technology-enhanced item type is completely unique to the assessments used in this study. If this item were to be classified by the Scalise & Gifford (2006) Taxonomy, it would most likely be in the completion category, in the most complex category of matrix completion.

Potential Benefits of Technology-Enhanced Item Types

The potential benefits of utilizing technology-enhanced item types have only begun to be investigated. These potential benefits, and the effects on psychometric properties need to be further explored before the various item types can be fully added to our lexicon (Zenisky & Sireci, 2002).

One potential benefit of technology-enhanced item types is the possibility of being able to create items to match any type of construct needed, while keeping a high amount of fidelity (Zenisky & Sireci, 2002). These various types of items provide an almost limitless ability to measure a large variety of constructs. This ability to match the construct of interest is also known as construct representation. According to Sireci and Zenisky (2006), construct representation is

the "ability of a test to fully represent all the knowledge, skills, and abilities inherent in the construct measured." It is believed that items that are able to increase fidelity of the behavior of interest will also have higher construct representation. An increase in construct representation is a great potential benefit of these item types, but there is also potential for construct-irrelevant variance (CIV). Sireci and Zenisky (2006) define CIV as variance caused by "attributes unintentionally measured by a test that affect test scores (e.g., English proficiency affecting math test performance)." It is feared that items with additional technology and interactions may begin to test not only the construct of interest, but also the test taker's technological savvy.

Another often cited benefit of technology-enhanced item types is test takers' preference for these items. As cited previously, multiple-choice items often have been maligned, so alternative item types are increasingly desired by test takers. Bennett and Sebrechts (1997) asked test takers about their perceptions of sorting tasks and standard multiple-choice questions. They found that test takers generally preferred the sorting items, and felt the sorting items were a better assessment of their ability.

Additionally, Parshall et al. (2000) cited the decreased chance of guessing a correct answer as an added benefit of utilizing technology-enhanced item types. For example, assuming no knowledge of the correct answer, for a four-option multiple-choice item, a test taker has a 25% chance of answering the item correctly when randomly guessing. If there were a four-option reordering/rearranging item, the chance of guessing correctly (ordering all four options correctly) is greatly reduced at only 4.1%. While promising, more research is needed to test the psychometric properties of these various technology-enhanced item types. Currently, most validity research for TE items focuses on comparing new item types to multiple-choice item types (see next section), and testing the degree to which new item types provide item information

and discrimination. The next section will describe the research that has been conducted to identify evidence of validity while using these item types.

Comparison of Technology-Enhanced Item Types and Multiple-Choice Item Types

The majority of research into technology-enhanced item types focuses on information, information efficiency, item discrimination, and construct equivalence as compared to multiple-choice items. The authors of these studies have attempted to determine if these new item types are better (i.e., worth the cost) than traditional multiple-choice items.

The expanded use of TE item types has provided new opportunities to study the psychometric properties of these items. Wan and Henly (2012) compared multiple-choice items with figural-response, short constructed-response, and extended constructed-response items on a state science assessment. They found that a figural-response item produced more information than a multiple-choice item for high school grades. This increase in information was fairly small in size. The short constructed-response items also produced more information in high school grades, but similar information in lower grades compared to multiple-choice items. Finally, extended response items provided the most information across all ability levels in all grades. It appears that each of these item types provided different levels of information depending on the grade level. This might be due to the item complexity and the age of the test taker.

Jodoin (2003) compared multiple-choice items and technology-enhanced items built for the Microsoft Certification Systems Engineer (MCSE) certification program. The multiple-choice items either consisted of basic four-option multiple-choice, or multiple-select multiple-choice items. The technology-enhanced items were either drag-and-connect (similar to matching

item types described previously) and build-a-tree (similar to categorizing) item types. The researchers utilized a 3-PL model for dichotomous items and a graded response model for all polytomous items. They found that the average information for the innovative item types exceeded the information provided by the standard multiple-choice items across all ability levels. Jodoin (2003) also determined that the time to complete the technology-enhanced item types was greater than the multiple-choice item types.

To compare the information provided by each item type while accounting for the time to respond to the question, researchers can use information efficiency. Information efficiency was defined by Wan and Henly (2012) as the "mean weighted item information divided by the average time spent on an item within an item type." When they took the efficiency of each item into account, they found that the figural-response items had similar information efficiency to multiple-choice items. However, a limitation of this study was the use of pilot data to calculate the information efficiency. Jodoin (2003) on the other hand, found that multiple-choice items provided more expected information per unit time than technology-enhanced item types. Specifically, the multiple-choice items provided less information overall, but they provided nearly double the information per unit of time. This finding reinforces Bennett et al. (1990) who suggested using multiple-choice items due to their ability to be answered quickly, thus allowing more content coverage and representation of the domain of interest, all while increasing the reliability of the assessment.

Wan and Henly (2012) compared the discrimination of each of these item types, and found that figural-response and multiple-choice items provided very similar discrimination ability. Additionally, both the figural-response and multiple-choice items were more discriminating than the constructed-response items. In contrast, Martinez (1991) examined

figural-response and multiple-choice items that were built to be parallel. Items were then correlated with the total score of the same format. For example, multiple-choice items were correlated with the total score of the multiple-choice items, and the figural-response items were correlated with the total score of the figural response items. Martinez (1991) found that figural-response items have superior discrimination in comparison to multiple-choice items. Specifically, the author found that this increase in discrimination was moderated by whether the figural-response items were easier or harder than the multiple-choice items. As figural-response items got easier, the differences in the item discriminations decreased. Martinez's conclusion was that in comparison to parallel multiple-choice items, figural-response items were comparable or better.

Additional research into technology-enhanced item types centered on the construct equivalence of these items to standard multiple-choice items. Wan and Henly (2012) found that a single-factor model best fit the data and provided reasonable fit statistics. Though the authors did suggest that analysis of the items utilized as technology-enhanced could have easily been rewritten as multiple-choice items without changing the construct measured. They further suggested that this is not uncommon and that the appeal of these new item types may be overrated. Finally, Traub (1993) compared constructed-response items and multiple-choice items for reading comprehension and concluded that the different format did not have any distinct effect on the construct they were designed to measure. Additionally, the authors suggested that if any differences do exist, they would likely be very small.

In summary, the research into the psychometrics of these items is currently fairly limited. It appears that some technology-enhanced items may provide more information than standard multiple-choice items, but that they also take more time for students to complete. In the future,

the time required may decrease as students become more familiar with these new item types. Additionally, it appears that TE items most likely assess constructs similar to those assessed by standard multiple-choice items. Obviously, more research is needed, and different and more varied technology-enhanced item types will need to be compared before any definitive answer can be provided. It does appear that TE items are popular with students, and that further research will need to determine if that translates to better statistics and test characteristics.

Basics of Test Theory

There are two main testing theories that drive the majority of test development. The first theory commonly utilized in test development is Classical Test Theory (CTT). Classical Test Theory is founded on a single simple principle: X=T+E (Hambleton & Jones, 1993). In this equation, the X stands for the observed score. This is the score received from the administration of an assessment, and is an unweighted sum of all the item scores. The T stands for the true score. The true score is the test takers actual ability level on the construct of interest. This is defined as the expected value of X over a theoretically infinite number of repeated test administrations. Unfortunately, the true score is an unknowable quantity; which is the reason for the utilization of assessments. The final piece of the equation is E, which stands for Error. Error is any unsystematic variability that prevents the test from measuring the true score. This equation is simple, but provides the backbone for the CTT approach to test development.

Item Analysis in CTT

In classical test theory, there are two main measurements of item analysis: *p*-values and item-total correlations. A *p*-value is an estimate of the difficulty of an item. In the simplest version, a *p*-value for a dichotomous item is simply the proportion of test takers who responded

to that particular item correctly. For example, if 100 test takers took an item, and 65 of those test takers responded correctly. The *p*-value would be the proportion correct, 65/100, or .65. When discussing items that are scored polytomously, the *p*-value is no longer simply the proportion of test takers who responded to that question correctly; it is the proportion of possible points (e.g., the item could have 2, 3, 4, etc. possible points) earned on an item. For example, if a polytomous item has a *p*-value of .65, this indicates that the proportion of possible scores was .65.

The other important item statistic is an item-total correlation or discrimination index. This value indicates how well the item differentiates between "good" test takers and "bad" test takers. The discrimination index is easy to calculate; it is simply the correlation between the scores on a particular item and the total scores. The index is a simple correlation, which means it can range from -1 to 1, with 0 indicating no relationship. The closer the discrimination index is to 1, the better the item is at distinguishing among test takers.

Reliability in CTT

Reliability is defined by Lord and Novick (1968) as "the squared correlation ρ_{XT} between observed score (X) and true score (T)." More broadly, reliability is the ratio between true score variance and observed score variance. For this study, coefficient alpha will be utilized as the main estimate of test reliability. The formula for Cronbach's Alpha (1951) can be found in Equation 1.

$$\boldsymbol{\rho_{CC'} \geq \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_c^2}\right]}$$

Equation 1. Cronbach's Alpha (1951)

Item Response Theory (IRT)

The second theory of test development is Item Response Theory (IRT). Item Response Theory is a testing theory that links a latent trait to a set of item responses. Using mathematical assumptions, IRT enables a test developer to create a scale for an assessment that is invariant to both the examinees who take the assessment, as well as the specific items on the assessment. The basic equation for a 1-Parameter Logistic (1-PL) IRT model is shown in Equation 2. When plotted, this equation provides us with an item characteristic curve (ICC). An item characteristic curve is a mathematical model of the relationship between ability and item performance. A visual representation of this is an ogive curve that can be plotted using Equation 2.

$$P(u=1|\theta_i)=\frac{e^{(\theta_i-b_j)}}{1+e^{(\theta_i-b_j)}}$$

Equation 2: 1-PL

In Equation 2, θ_i represents the ability of the test taker on a given latent trait. The value of θ_i is continuous and scale indeterminate; typically θ_i is standardized with a mean of 0 and a standard deviation of 1. The second piece of this equation is the b_j parameter. In IRT, the b_j represents the difficulty of an item. Specifically, this is the point where a test taker has a 50% chance of answering a question correctly conditional on the latent trait. Equation 2 is the simplest version of IRT; Equation 3 (2-PL) and Equation 4 (3-PL) expand upon the original equation by adding additional parameters.

$$P(u=1|\theta_i)=\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}$$

Equation 3: 2-PL

The 2-PL model (Equation 3) adds the a_j parameter to the original equation. This parameter allows for the estimation of a discrimination index. Specifically, this allows the slope of the ICC to take on different values across items. The greater the value of the a_j parameter, the steeper the slope of the ICC. As the steepness of the slope increases, the more discriminating that item is at a specific level of θ_i.

$$P(u=1|\theta_i) = c_j + (1-c_j)\frac{e^{a_j(\theta_i - b_j)}}{1+e^{a_j(\theta_i - b_j)}}$$

Equation 4: 3-PL

The 3-PL model (Equation 4) adds the c_j parameter to the 2-PL equation. The c_j parameter is considered a pseudo guessing parameter, and can take on a value from 0 to 1. This parameter adjusts the ICC to account for a test takers' ability to guess an answer correctly.

The previous models are utilized on dichotomous assessments. As different item types are integrated into the assessment, more complex IRT models are used. Polytomous IRT models allow for items to have more score points than just 0 or 1. For the purposes of this dissertation, only two of the many polytomous models will be described. The first model is the Generalized Partial Credit Model (GPCM). This model allows for the use of items with multiple score points, as well as includes the discrimination parameter. The Generalized Partial Credit Model is shown in Equation 5. The a_j, b_{jk}, and θ_i parameters are defined the same as with the previous models, except now there is a b_{jk} parameter for every category boundary.

$$P(u_{ij} = x \mid \theta_i) = \frac{e^{\sum_{k=0}^{x} a_j(\theta_i - b_{jk})}}{\sum_{h=0}^{m_j} e^{\sum_{k=0}^{h} a_j(\theta_i - b_{jk})}}$$

Equation 5: Generalized Partial Credit Model (GPCM)

Information in IRT

In item response theory, reliability is discussed in terms of information and precision. Information can be described at the item and test level. Item information describes an item's ability to distinguish examinees with higher latent trait levels from examinees with lower latent trait levels. Typically, item information is displayed as an item information function (IIF). This function graphically shows where an item best discriminates examinees with higher latent trait levels from lower latent trait levels. If item information from all test items is combined, the resulting information will be test information. Test information informs the test developer about the certainty of the ability estimation at any level of θ_i (Sireci, Thissen, & Wainer, 1991; Wang, Bradlow, & Wainer, 2002). Specifically, the more information a test provides, the higher the reliability of that test (Wan & Henly, 2012). This can also be graphically displayed as a test information function (TIF).

Item Scoring

Basics of scoring various item types. Item scoring is at the heart of all test development. Decisions about scoring must be made at both the item and test level when developing assessments. Item-level scoring focuses on evaluating the work of the test taker and the outcomes are numerical values (e.g., 0, 1). Luecht (2001) defined item-level scoring as "the process of codifying the response(s) and other relevant information that the examinee provides

into a numerical quantity that reliably and validly represents the examinee's performance." Test-level scoring or evidence accumulation, focuses on how to analyze the item-level scoring (Wainer et al., 2006). An example of this is using an IRT model to determine a student's latent trait estimate.

Within the item level, there are multiple factors to the majority of scoring schemes (Luecht, 2001). The first is the examinee response. The examinee response could be as simple as the selection of response option A, or as complex as identifying the coordinates of a marker placed on an xy coordinate grid. That response is then compared to the answer expression, which is the idealized answer to that item. For example, if the key of a multiple choice item is B, and the student responded with A, then they would receive a 0. This answer expression can be further broken down into three distinct parts. The first is the response that is evaluated. In the example above, this is the student's response of A. The second part is the answer key, which in this example is B. Finally, the student's response and the key are compared to provide a numerical value. In this case, the student would receive a 0 for this item. This basic process can be used with item types ranging from multiple choice, to technology enhanced (e.g., categorizing). In terms of multiple-choice items, scoring typically uses a Boolean "IF f(x,y) Then assign(value)." (Luecht, 2001). This simple approach to scoring becomes more complicated as we add more response possibilities, as well as partial credit.

Current scoring practices. Traditionally, assessments have scored items dichotomously, or simply right/wrong. This scoring scheme has led to criticism due to lack of information about why the student answered correctly or incorrectly (Rogers & Ndalichako, 2000). For example, students might have answered the dichotomous item correctly because of in-depth knowledge about the subject, or because they have partial knowledge and were able to narrow down the

responses and guess. Conversely, they might have a strong grasp of the content, but due to a misleading response option, answered the item incorrectly. Regardless of the reasons for incorrect responses, students' knowledge is not fully being measured by this scoring scheme (Grunert, Raker, Murphy, & Holme, 2013). However, even with the criticism of dichotomous scoring listed above, there are benefits to the simplicity of this scoring scheme, foremost being the ease of explanation to students and parents, and the researched validity of the scheme (Rogers & Ndalichako, 2000). The following section will describe alternatives to dichotomous scoring. Due to a lack of in-depth research into technology-enhanced items, the majority of the research will be taken from items such as multiple-select multiple-choice (e.g., Pick-N) item types.

Scoring Methodology

Simply put, there are two main methods for scoring items. The first, and often most used, is dichotomous scoring. As mentioned previously, this method scores items as right or wrong. The second methodology is polytomous scoring. Items scored polytomously allow for more variation in the points awarded. This type of scoring allows for a finer grain evaluation of the students' knowledge. As described later, polytomous scoring can take many different forms. Since innovative item types are new, and there is limited research on how to score these items, most of this section will focus on how to score an item type that is similar to some of the TE items. This item type will be referred to as a multiple-select multiple-choice (MSMC). In the research this is sometimes called a Pick-N item, type X, or multiple true/false.

An MSMC item asks students to "choose all that apply" (Bauer, Holzer, Kopp, & Fischer, 2010). Specifically, students are given many answer choices (normally more than four), and asked to pick all the correct responses (greater than one, but less than the total number of

options). With these types of items, scoring them dichotomously may reduce the amount of information provided by the item (Albanese & Sabers, 1988). With MSMC items, if a student identifies three out of four correct answers and the item is scored dichotomously, the student is treated the same as a student who did not identify any of the correct responses. With most content, there may be valuable information available, as there are students between those who know everything, and those who know nothing. Additionally, Ripkey, Case, & Swanson (1996) found that when these types of items were treated as dichotomous, the items became very difficult. This occurred if even one response option was more difficult than the others. This difficulty reduced the variability in the scores, thus reducing the amount of information provided by the test overall.

There are multiple ways in which these types of items can be scored polytomously. Bauer et al. (2010) attempted to score MSMC items using three different methods (two of which were polytomous). The first method, which was called partial-credit scoring 1 (PS50), assigned 1 point if the student selected all correct answers, 0 points if they selected no correct answers, and .5 point if the student selected at least 50% of the correct responses. This method is similar to what will later be called threshold scoring. A variation of this threshold method was proposed by Albanese & Sabers (1988). They suggested giving half credit to a respondent who answered above chance levels.

The second polytomous method described by Bauer et al. (2010) was called partial-credit scoring 2 (PS1/*m*). This method, which will later be called partial-credit scoring, assigns 1/*m* points for each correct response. In this method, *m* stands for the total number of correct responses. For example, if there is an item with four correct responses, then each correct response would receive .25 point. If a student answered three out of the four responses correctly,

he/she would receive .75 out of 1 for that item. This full partial-credit scoring was also described by Albanese and Sabers (1988) and Ripkey et al. (1996).

Albanese and Sabers (1988) also created a scoring method that falls between the threshold and the partial-scoring method. In this hybrid version, not all responses get partial credit. They suggested giving partial credit to scores once they exceed the level of chance. For example, if an item has eight total responses where four of the responses are correct, the chance of responding to any single correct answer randomly is .50. So in this case, a student would have to respond to at least two of the correct responses before receiving partial credit. Once that threshold has been met, a student would receive credit. In this example, a student would receive 0 points for fewer than two correct answers, .5 point for two correct answers, .75 point for three correct answers, and 1 point for all four correct answers.

As mentioned previously, there is very little research focusing on innovative item types, and the best way to score them. Scalise & Gifford (2006) suggested that reordering and rearrangement items can be scored as either dichotomous or polytomous. Furthermore, they suggested that in these types of items, certain order placements may be valued (or weighted) as more important than others. For example, if the item involves reordering parts of a formal letter, not placing the salutation and closing in the right places may be more important (or have a higher point value) than correctly identifying where the parts of the letter body should be placed.

Testlet scoring. Another possible way to score TE item types is to treat them as testlets. A testlet is a group of items that share a single theme or stimulus (Wang et al., 2002; Wainer et al., 2006). In the case of a TE item, each possible response would be treated as an individual question, with the overarching item being the shared stimulus. For example, if a matching item asks students to drag words to match their definitions, then each word would be considered an

individual item within a testlet. Treating the items in this way would allow the use of testlet response theory.

Testlet response theory allows for items to be conditionally dependent. A major assumption in IRT is that a response to one item is completely independent from all other items conditional on the latent trait being measured. When items are conditionally dependent, a response to one item might affect the response to another item, even after accounting for the latent trait measured by the test. Due to this, the IRT assumption of conditional independence may not be met when using testlet items. This conditional dependence could occur because of a student's misinterpretation of the item, understanding of that particular topic, item fatigue, etc. (Wang et al., 2002). Failing to account for conditional dependence within a testlet has been found to yield 10-15% overestimation of reliability (Sireci et al., 1991).

The Testlet Response Theory (TRT) model is similar to previously presented IRT models with the addition of the γ_{id}. The γ_{id} parameter describes the interaction between a test taker and an item within a testlet. This additional parameter will model the potential for conditional dependence that could occur with the use of testlets. The 3-PL version can be found in Equation 6.

$$P(u=1|\theta_i) = c_j + (1-c_j)\frac{e^{a_j(\theta_i - b_j + \gamma_{id(j)})}}{1+e^{a_j(\theta_i - b_j + \gamma_{id(j)})}}$$

Equation 6: Testlet Response Theory Model (TRT)

Basically, testlet response theory allows for an additional interaction term for test takers answering specific items within a testlet (Wang et al., 2002). Wang et al., (2002) demonstrated that estimated latent traits and item parameters are biased when testlet dependencies are ignored.

Additionally, they found that the amount of dependency that exists varies across testlets. Lee, Kolen, Frisbie, and Ankenmann (2001) determined that treating items with a shared stimulus as individual items did violate the assumption of local item independence and unidimensionality. These assumptions were satisfied if the unit of measurement was the testlet instead of each item.

Research on Multiple-Select Multiple-Choice Scoring

Much of the current research on polytomous scoring methods has focused on the multiple-select multiple-choice item type. This item type is sometimes called Type X, Pick-N, or multiple true/false. This type of item does not meet the criteria of a technology-enhanced item type, but some of the aspects of these items do have some similarities to TE items. Mainly, each response that is evaluated can either be treated as a piece of the whole item, or as a singular item with a shared stimulus. The main difference between these item types is the different actions required of students who are taking the assessments. For this reason, the next section will describe research that focused on the MSMC item type.

As discussed previously, items can basically be scored dichotomously or polytomously. Research on MSMC item types has shown that scoring them dichotomously creates rather difficult items (Ripkey et al., 1996; Bauer et al., 2010). Specifically, Bauer et al. (2010) found that when adjusting the scoring of MSMC items between correct only or awarding partial credit, 20% of the items became too difficult when correct-only scoring was applied, while only 0.03% of the items were too difficult when partial-credit scoring was utilized. Additionally, the authors found that dichotomously scored tests were more difficult by 4-5 points compared to polytomously scored tests.

Albanese and Sabers (1988) also conducted research on MSMC items. Overall, they found that scoring these item types as correct only resulted in low reliability. Theoretically this

makes sense as more difficult items have far less variance, which in turn effects the reliability of the assessment. Bauer et al. (2010) also found that dichotomous scoring of MSMC items provided the lowest estimates of reliability. An alternative is scoring the item utilizing a threshold methodology which provided the highest estimates of Cronbach's Alpha. The partial-credit methodology was right behind the threshold method. However, using the Spearman-Brown prophecy formula, there was very little difference between partial-credit and threshold scoring when projected out to 100 items. Additionally, Albanese and Sabers (1998) compared the internal consistency (coefficient alpha) of MSMC items scored with partial credit to the same MSMC items scored as individual items. They found that even though they were linear transformations, the alpha levels of the items scored as individual items was higher than those scored polytomously as a single item. In addition, they found that random guessing may not be an issue; therefore scoring methods that award points for any correct responses (partial credit) tended to be slightly better than methods that do not give credit until chance levels have been reached (threshold).

In addition to difficulty and reliability, item-total correlations or item discrimination are important indexes to consider when determining the benefits of a scoring method. Similar to the decrease in reliability due to difficult items, the difficulty of correct-only scoring also resulted in a lower item discrimination (Ripkey et al., 1996). Specifically, Ripkey et al. (1996) found that, depending on the type of scoring used, MSMC items became either the most or least discriminating. When items were scored as polytomously their item discrimination increased. In contrast, Bauer et al. (2010) found very small differences in item discrimination between dichotomous and two different polytomous scoring methods. The polytomous scoring methods were slightly higher than the dichotomous method, and resulted in fewer item-total correlation

values below .2. Additionally, dichotomous scoring resulted in the only negative item-total correlation. Overall though, the differences mentioned above were very small.

Comparison of Scoring Methodology

The previous section focused on non-technology-enhanced item types, specifically items known as multiple-select multiple-choice. This next section will discuss research on dichotomous versus polytomous scoring for innovative item types, and other items that lend themselves to this type of scoring.

According to Jodoin (2003), evidence is building that indicates allowing technology-enhanced item types to be scored polytomously increases the amount of information, which is provided by dichotomous scoring methods. Additionally, innovative item types do tend to take longer for students to complete than standard multiple-choice items, so it is still questionable if the added information is worth additional testing time. Vispoel and Kim (2014) also cited mounting evidence that scoring items polytomously versus dichotomously improves evidence of reliability and validity. In their study, they utilized the Balanced Inventory of Desirable Responding (BIDR) and scored some of the Likert-type items either polytomously or dichotomously. Items that were given high ratings (i.e., 6 or 7) were given a score of 1, all other items received a score of 0. A 2-PL IRT model was used for the dichotomous items and the partial-credit model (PCM) and Graded Response Model (GRM) were fit to the polytomous items. They found that internal consistency was higher when polytomous scoring was used instead of dichotomous scoring. Additionally, they found that of the two polytomous models for scoring the assessment, the PCM was the less desirable of the two. Additionally, they found that their results matched prior findings that scoring items polytomously results in higher internal consistency, test-retest, and convergent validity coefficients in relation to dichotomous scoring

methods. Finally, they suggested that using GRM was the best practice over all other IRT methods utilized in their study.

Donoghue (1994), utilized data from a field-test of the National Assessment of Educational Progress (NAEP) reading assessment. The author found that when scoring constructed-response items polytomously the amount of information provided was between 2.1 and 3.1 times more information than when scored dichotomously. Though the amount of information from polytomous scoring of constructed-response items was impressive, Donoghue found that multiple-choice items provided more information per minute than the polytomously scored constructed-response items. Donoghue (1994) did suggest that there is support for treating each response option as a separate item, thus making it worth 0 to *k*-1 score points.

Davey et al., (1997) utilized a simulation study to determine the effects of treating these item types as conditionally independent. The authors found that the more conditional dependence added to the response data, the poorer the performance of response-pattern-based scoring. Grunert et al. (2013) adjusted an already developed multiple-choice assessment to allow for partial credit on items where the incorrect responses potentially provided additional information. They found that allowing polytomous scoring resulted in higher mean percentage scores for the test as a whole. Additionally, they found that low-performing students' scores did not gain much from allowing partial credit. They theorized that this was due to lower performing students making bigger mistakes that, regardless of the scoring type, resulted in receiving no credit for the items. Along those lines, high-scoring test takers also gained little from adjusting the scoring type due to the majority of those test takers already receiving full credit on the item. The middle performers gained the most from the adjustment in scoring, though approximately 26% of

percentile rankings remained the same, and the increasing and decreasing of scores balanced each other out.

Jiao, Liu, Haynie, Woo, and Gorham (2012) evaluated the effects of scoring method on a single type of TE item in a computer adaptive testing (CAT) environment. They also followed up their data analysis with a simulation study based on the parameters estimated from the real data. The TE items used in this study were fill-in-the-blank, ordered-response, multiple-response, and hot-spot items. A potential issue with this study was the relatively small number of items that had their scores adjusted in comparison to the total item pool utilized for the CAT. Specifically, there were 92 total items adjusted out of a total of 16,870. Due to the small number of items adjusted, the real data analysis did not reveal any differences in ability estimation or classification decisions. The simulation study used the ability estimates from the polytomous scoring calibration and treated them as true parameters. The authors then generated item response data and re-scored the items as dichotomous. They found that there was less bias when using polytomous scoring than when using dichotomous scoring at the upper end of the ability scale. Additionally, classification accuracy was slightly higher for polytomous scoring, though the number of examinees affected was only 0.7%. Considering that the number of TE item types was small in comparison to the total number of items, the increase in measurement precision was still prevalent. Finally, Rogers and Ndalichako (2000) compared finite-state scoring with number-right, one, two, and three-parameter item-response scoring methods on a state test of reading comprehension. They found that there was a strong agreement between these four methods of scoring.

In summary, the evidence appears to be leaning towards the added information and enhanced reliability of polytomous scoring on TE item types. The majority of these studies all

focused on utilizing IRT methods for scoring and scaling. Additionally, most of the studies that did focus on TE items only had one or two different kinds of TE items to test. More research needs to be conducted to test the effects in both CTT and IRT of adjusting the scoring method of these technology-enhanced item types.

Chapter Three: Methods

The purpose of this research is to systematically test the effects on test characteristics of utilizing different scoring strategies for TE item types. The TE item types have been gathered from the Career Pathways Assessment System (cPass). Two assessments were released operationally to students in Kansas and Mississippi. The two assessments were delivered between October 2013 and June 2014 and covered different content areas. The first assessment is the General Career and Technical Education (CTE) assessment. This assessment covers basic academic foundations, as well as 21st-century skills (e.g., leadership and communication). The second assessment is the Comprehensive Agriculture assessment, which covers agribusiness, animal systems, plant systems, food products and processing, and natural resources/environmental science. Students in the career and technical education pathway system major in an area of interest as early as their freshman year in high school. These majors are known as pathways, and each pathway covers a large swath of content related to a specific field. For instance, those who are interested in field crops, might take classes in the plant sciences area. Students who utilized the General CTE assessment are juniors or seniors in high school, and can major in any of the CTE pathways. Students who take the Comprehensive Agriculture assessment are also juniors and seniors, but must have taken a variety of classes in the agriculture clusters (e.g., agribusiness, animal systems, plant systems, etc.).

The cPass system has students take a General assessment (described above), which covers content areas common among all the different pathways. Then each student takes their pathway-specific assessment. In this case, the pathway is Comprehensive Agriculture. Students who take this assessment have had multiple courses covering many different areas in agriculture. These assessments consist of 100 items each and contain multiple-choice items, as

well as various types of TE item types. This research will look at both the General CTE, as well as the Comprehensive Agriculture assessments. Both assessments are utilized to help with the generalizability of the research conclusions, as each assessment covers completely different topics. Additionally, each test contains a unique TE item type not seen on the other assessment. The following sections will describe each assessment, including the types of items utilized and the demographics of the students who completed the assessments.

General Career and Technical Education Assessment

The General CTE assessment was designed for students to take as a precursor to their pathway-specific assessments. Development consisted of building two operational forms, each with 100 items developed with input from secondary, post-secondary, and industry content experts. Each form was built to exact specifications indicated by the test blueprint (Test specifications can be found in Appendix B). Each of the two forms were built with 17 TE item types. The rest of the items are either multiple-choice or situational judgment tasks. For the use of our research, non-TE item scoring will not be altered throughout analysis. Additionally, numerical entry (constructed-response items) do not lend themselves to polytomous scoring. Therefore, if there is a numerical entry item, the scoring will not be altered throughout analysis. Each of the two forms share 12 common items sampled from the test blueprint. Students were randomly assigned to one of the two forms by the Kansas Interactive Testing Engine (KITE™). This randomization occurs when the student logs in to the system for the first time. Additionally, this log-in helps create consistent groups of students who can more easily be compared.

Each of the two forms contain a combination of unique and core TE items. Table 1 specifies the type of TE items that appear on each form. Please note that the editing item type is

unique to the General CTE assessment, and does not appear on the Comprehensive Agriculture assessment.

Table 1. *General CTE Technology-Enhanced Item Types by Form*

Form	Constructed Response	Editing	Matching	Categorizing	Reordering/ Rearranging	Substitution/ Correction
A	1	2	5	5	2	2
B	0	4	5	2	1	5

Comprehensive Agriculture Assessment

The Comprehensive Agriculture assessment was designed for students to take at the end of their secondary course work within the Agriculture Pathway. This assessment was specifically built for students who have a broad interest in agriculture. The Comprehensive Agriculture assessment covers agribusiness, animal systems, food products and processing, natural resources/environmental science, and plant systems. Students who take this assessment will typically be in their junior and senior years of high school. The first year of operational assessments utilized three forms, each with 100 items developed by secondary, post-secondary, and industry content experts. Each form was built to exact specifications indicated by the test blueprint (Test specifications can be found in Appendix B). Each of the three forms were built with exactly 20 technology-enhanced items. The rest of the items were multiple-choice items. For the use of our research, non-TE item scoring will not be altered throughout analysis. Additionally, numerical entry (constructed-response items) do not lend themselves to polytomous scoring. Therefore, if there is a numerical entry item, the scoring will not be altered throughout analysis. Each of the three forms share 30 common items that appear on all three

forms and are sampled from the test blueprint. Students were randomly assigned to one of the three forms (A, B, or C) by the Kansas Interactive Testing Engine (KITE). This randomization occurs when the student logs in to the system for the first time. Additionally, this log-in helps create consistent groups of students who can more easily be compared. Each of the three forms contain a combination of unique and core TE items. Table 2 specifies the type of TE items that appear on each form. Please note that the Punnett square item type is unique to the Comprehensive Agriculture assessment, and does not appear on the General CTE assessment.

Table 2. *Comprehensive Agriculture Technology-Enhanced Item Types by Form*

Form	Constructed Response	Figural Response	Matching	Categorizing	Reordering/ Rearranging	Punnett Square
A	0	2	4	8	5	1
B	1	5	4	8	2	0
C	0	4	5	8	3	0

General CTE Participants

Participants for the General CTE assessment were junior and senior students from Kansas and Mississippi. A total of 1,028 students were assessed using the General CTE assessment, and were randomly assigned to two forms built to be parallel. In order to ensure that students used for the analysis made a legitimate attempt to take the assessment, as well as to remove any student who may have had technological issues with the assessment system, any student who did not complete at least 10% of the items was removed. Any test taker who responded to at least 10% of the items is considered to have made a successful attempt, and thus all non-responses were

treated as incorrect. This left a total of 859 test takers total. Form A had a total of 406 total students. Due to federal regulations, demographic information is not required for students taking these assessments; of the students who provided gender information (n = 271), 62.73% were male and 37.26% were female. Students who provided data (n = 256) were mainly from Kansas (74.60%), with the remaining students being from Mississippi (25.39%). The students who provided information (n = 271) were mainly Caucasian (93.35%), with the next largest group being American Indians or Alaskan Native (4.05%). The remaining students were Black or African American (1.84%) and Asian (0.73%). Form B had 453 students who responded to at least 10% of the items on the assessment. Out of the 453 total students for Form B, 302 provided information on gender. For Form B, 57.61% were male and 42.38% were female. Students who provided data (n = 302) were mainly from Kansas (75.82%), with the remaining students being from Mississippi (24.17%).The students who provided information (n = 301) were mainly Caucasian (93.36%) with the next largest group being American Indians or Alaskan Native (2.33%). The remaining students were Black or African American (1.99%), Native Hawaiian or Other Pacific Islander (1.33%), and Asian (1.00%).

Comprehensive Agriculture Participants

Participants for the Comprehensive Agriculture assessment were junior and senior students from Kansas and Mississippi. A total of 455 students were assessed using the Comprehensive Agriculture assessment, and were randomly assigned to three forms built to be parallel. In order to ensure that students used for the analysis made a legitimate attempt to take the assessment, as well as to remove any student who may have had technological issues with the assessment system, any student who did not respond to at least 10% of the items on the assessment was removed. Any test taker who responded to at least 10% of the items is

considered to have made a successful attempt, and thus all non-responses were treated as incorrect. This left a total of 386 test takers. Form A had 125 students who responded to at least 10% of the questions. Due to federal regulations, demographic information is not required for students taking these assessments; of the students who provided gender information (n = 85), 61.2% were male and 38.8% were female. Students who provided data (n = 84) were mainly from Kansas (67.9%), with the remaining students being from Mississippi (32.1%). The students who provided information (n=85) were mainly Caucasian (96.5%), with the remaining students being American Indians or Alaskan Native (3.5%). Form B had 127 students who responded to at least 10% of the questions. Out of the 127 total students for Form B, 73 provided information on gender. For Form B, 52.1% were male students and 47.9% were female. Students who provided data (n = 96) were mainly from Kansas (61.5%), with the remaining students being from Mississippi (38.5%).The students who provided information (n = 73) were mainly Caucasian (94.5%), with the next two groups being American Indians or Alaskan Native (2.73%) and Asians (2.73%). Form C had 134 students who responded to at least 10% of the questions. Out of the 134 total students for Form C, 81 provided information on gender. For Form C, 61.7% were male and 38.3% were female. Students who provided data (n = 116) were mainly from Kansas (56.9%), with the remaining students being from Mississippi (43.1%). The students who provided information (n = 81) were mainly Caucasian (93.8%), with the next two groups being American Indians or Alaskan Native (4.9%) and Black or African Americans (1.23%).

Scoring Methodology

This research utilized multiple scoring methods. Each of these methods (with the exception of testlet response theory) were applied to both CTT and IRT methods. The section below will describe the multiple methods used for scoring these items. Each TE item was scored

as correct only, threshold, threshold partial, partial credit, subtractive, and utilizing testlet response theory. The first four methods will be described in terms of CTT. These were converted for use in IRT polytomous models. For example, with partial-credit scoring utilizing IRT, instead of each correct response receiving a fraction of the total score, each additional correct response would place a student in a higher score category. The same can be done with threshold and subtractive scoring methods. The only method that will only be utilized with IRT is the testlet response theory scoring method.

The first method for scoring TE items will be a dichotomous method. This method will be referred to as correct-only scoring. When scoring TE items as correct only, test takers will have to respond correctly to all aspects of the item in order to receive one possible point. If a test taker misses any component of the item, then no points will be awarded. This type of scoring will be considered the base type, and it is not expected to be the best way to improve validity evidence. Items scored as correct only tend to be more difficult, which can reduce item variance and covariance, thus reducing reliability. The IRT version of this method would treat this item as a dichotomous item.

The second method will be polytomously scored. Polytomously scored items allow test takers to receive multiple score points other than 0 or 1. The second scoring method will be referred to as partial credit. Partial-credit scoring will give test takers credit for each response they have answered correctly. The point value for each correct response is determined by dividing 1 (total item score) by the number of possible correct responses (m). For example, a categorizing TE item with six different words that must be sorted, for each response that is correctly sorted the test taker would receive .166 point. All correct points are then summed. For example, if four responses were correctly sorted, the test taker would receive .664 point (.166*4).

This methodology was described by Bauer et al. (2010). The IRT version of this method would treat this item as a six-category polytomous item.

The third scoring method is called subtractive scoring. Subtractive scoring is a variation of the partial-credit scoring method. In subtractive scoring, students receive partial credit (1/*m*) for each correct response they provide. For each incorrect response, they lose partial credit (1/*m*). As per the previous example, if a student were to correctly respond to four out of the six responses, the student would receive .332 point (.166*4 - .166*2). In this type of scoring, a lower bound of zero would prevent any negative point values. The IRT version of this method would also treat this item as a six-category polytomous item. However, in this approach the estimated threshold parameters between categories are likely to be larger than those derived from partial-credit scoring, as higher-category scores are more difficult to receive.

This fourth method will be called threshold scoring. Test takers will have a threshold that will need to be surpassed in order to be given a score greater than 0. For example, if an item has six words that must be sorted into two buckets, the value of the threshold would be set at 50% of the total score. In this example, a test taker would have to correctly sort at least three words into the proper buckets in order to receive a score greater than 0. If that threshold were reached without answering all items correctly, half a score point would be assigned. As with correct-only scoring, if the question is answered entirely correctly, a student would receive 1 score point. This methodology will allow a test taker to receive one of three scores for an item, 0, .5, or 1. This is similar to the methodology utilized by Albanese and Sabers (1988) for multiple-select multiple-choice items. The IRT version of this method would treat this item as a three-category polytomous item.

The fifth scoring method, referred to as threshold partial, is a variation on the threshold method. As with threshold scoring, test takers have a threshold that will need to be surpassed in order to be given a score greater than 0. Unlike the threshold method where there are only three possible scores, the test takers get 1/*m* point after the threshold is met. As with the threshold scoring example, test takers would need to correctly sort at least three words into the proper buckets in order to receive a score greater than 0. If in the same example, a test taker correctly sorts four out of the possible six correct responses, a score of .66 would be awarded. As with correct-only scoring, if the question is answered entirely correctly, a student would receive 1 score point. Threshold-partial scoring is similar to one of the methodologies utilized by Albanese & Sabers (1988) for multiple-select multiple-choice items. Using the IRT version of this method, the item would be treated as a five-category polytomous item.

The final scoring method utilizes testlet response theory, and is strictly for IRT. In this method, each item is treated as a testlet. Each score point will be an individual dichotomous item that is treated as sharing a common stimulus. In the current example, instead of having one item with 6 possible points, there would be six testlet items that are treated as conditionally dependent on the stimulus (that is, item response will be conditionally independent only after accounting for the latent trait and the person-by-testlet interaction parameter).

Statistical Software

The software programs utilized in this study included SPSS (IBM: Version 20.0), R (R Core Team, 2015), and R2Openbugs (Sturtz, Ligges, & Gelman, 2005). SPSS was utilized for all within-subjects Analysis of Variance (ANOVA), and R and R2Openbugs were utilized for IRT and TRT calibration and parameter estimation.

Data Cleaning

Data were gathered from the administration of both the General CTE and Comprehensive Agriculture cPass assessments. The raw response strings from each of the TE items were analyzed and converted to either a 0 or 1. For example, if a student correctly dropped a word next to its appropriate description, that student would receive a 1 for that response. These conversions were utilized to determine the final score for that student, on each individual TE item, for each scoring methodology. These score conversions were produced using Microsoft Excel 2013, and the scores were double checked against available machine scoring. For example, during operational use, multiple-bucket items were scored as partial credit. For these item types, the partial-credit score converted from the raw response were compared to the machine score for partial credit. In each case, these scores matched perfectly. Additionally, each student's response string was analyzed for missing items. If a student did not respond to at least 10% of the items, the student was determined to have not made a successful attempt and was deleted from further analysis. Students who responded to at least 10% of the questions were believed to have successfully attempted the test. In this case, all non-responses were adjusted to an incorrect response of 0. Additionally, when conducting the IRT analysis, certain items when scored correct-only or subtractive did not have any correct responses. These items were deleted from that form. Finally, to make sure IRT models were estimated correctly, items that had standardized residuals greater than 20 (unless the item was a TE item) were eliminated, and the IRT model was re-calibrated.

Item Calibration

All item response theory calibrations utilized the MIRT package in R. Depending on the scoring methodology, either a 2-PL, GRM, or 2-PL and GRM estimation was utilized. Due to sample size restrictions a 3-PL model was not suitable to the data. For calibration of the testlet response theory only utilizing TE items, R2Open bugs was used to estimate the parameters of the model. Using a Bayesian approach, priors for the *a* parameter were set using a log normal distribution, while *b*, and Ɵ parameters were set to a normal distribution with a mean of 0 and a standard deviation of 1. Utilizing two chains, 75,000 iterations for each chain were produced, with a burn in of 70,000. Convergence of the two chains in most cases signified proper estimation of the model.

Research Questions

Research Question 1: How does scoring method of TE item types affect basic psychometric properties (i.e., *p*-value, *r*, IRT based stats)? The first research question seeks to determine how adjusting the scoring scheme effects basic item-level statistics. For each data set (General CTE and Comprehensive Agriculture), raw response data were scored using the five different scoring methods. Each different scoring method should result in different scores for each student. Item level analysis was conducted on each of the five scoring schemes.

1a: **How does scoring method of TE item types affect basic psychometric properties when all items are utilized in analysis?** Specifically, what affect does adjusting the scoring methodology have in a mixed format assessment? Utilizing CTT, *p*-values and item-total correlations were calculated. All statistics were calculated using the full 100-item assessment. The only statistic that isn't affected by the number of items on the assessment is the *p*-value. The

p-value estimate is the same for both research questions 1a and 1b. Both averages and standard deviations of each of the CTT item statistics for each of the scoring schemes were calculated. Additionally, for *p*-values and item-total correlations a repeated measures analysis of variance (ANOVA) was conducted on each statistic to determine the significance of any differences between scoring methodologies. A Bonferroni correction was then utilized when comparing main effects of each scoring methodology.

IRT parameters were estimated for all 100-item forms using a combination of 2-PL and Graded Response (GRM). Each form was estimated independently to help determine the effects of score adjustment at the test form level. Item parameters (*a*, *b*) means and standard deviations were calculated and compared. For items with multiple *b* parameters (polytomous items), these *b* parameters were averaged before being compared to help with comparability. Additionally, for the *a* and *b* parameters a repeated measures analysis of variance (ANOVA) was conducted on each parameter to determine the significance of any differences between scoring methodologies. A Bonferroni correction was then utilized when comparing main effects of each scoring methodology.

1b: **How does scoring method of TE item types affect basic psychometric properties when only TE items are utilized in analysis?** Specifically, what affect does adjusting the scoring methodology have in an assessment with only TE items? Utilizing CTT, *p*-values and item-total correlations were calculated. All statistics were calculated using only the TE items on each form of the assessment. The only statistic that isn't affected by the number of items on the assessment is the *p*-value. The *p*-value estimate is the same for research questions 1a and 1b. Both means and standard deviations of each of the CTT item statistics for each of the scoring schemes were calculated. Additionally, for *p*-values and item-total correlations a repeated

measures analysis of variance (ANOVA) was conducted on each statistic to determine the significance of any differences between scoring methodologies. A Bonferroni correction was then utilized when comparing main effects of each scoring methodology.

IRT parameters were estimated utilizing only the TE items on each form. Depending on the form either 2-PL, GRM, or a combination of 2-PL and GRM were utilized. Each TE form was estimated independently to help determine the effects of score adjustment when the test only consists of TE items. Item parameters (*a*, *b*) means and standard deviations were calculated and compared. For items with multiple *b* parameters (polytomous items), these *b* parameters were averaged to help with comparability. The *a* parameters produced by the testlet response theory were averaged to calculate a single *a* parameter to be compared. This was similar to the procedure utilized for the *b* parameters of polytomous items. Additionally, for the *a* and *b* parameters a repeated measures analysis of variance (ANOVA) was conducted on each statistic to determine the significance of any differences between scoring methodologies. A Bonferroni correction was then utilized when comparing main effects of each scoring methodology.

Research Question 2: How does the scoring method of TE items affect reliability of scores from each test form? The second research question seeks to determine the relationship between changing scoring schemes of TE items and the reliability of the test forms. As with the last research question, only TE item scores will be adjusted.

2a: **How does the scoring method of TE items affect reliability of scores from each test form when all items are utilized?** Specifically, what effect does adjusting the scores of the TE items have on the reliability of the whole form. Although all non-TE items scoring remained unaltered, the covariances should change, thus altering the form's reliability. For CTT, reliability was calculated using Cronbach's Alpha for each form. Specifically, two forms for General CTE,

and three forms for Comprehensive Agriculture were analyzed for a total of five different forms. Cronbach's Alpha is reported for each of the test forms for each scoring methodology. Additionally, the standard error of measurement (*SEM*) is calculated for each form of each assessment. This allows for a better understanding of the effect of changing scoring methods on the variability and precision of the assessments.

For IRT, test information functions (TIFs) were compared to determine which methods provide the most information to the test developer. Comparing test information functions can help determine which scoring method provides the most information at any given level of Θ_i. Specifically, the higher the peak of the distribution, the more information about the test takers ability can be determined.

2b: How does the scoring method of TE items affect reliability of scores from each test form when only TE items are utilized? Specifically, what effect does adjusting the scores of the TE items have on the reliability of an assessment created with only TE items? For CTT, reliability was calculated using Cronbach's Alpha for each TE form. Two forms for General CTE, and three forms for Comprehensive Agriculture were analyzed for a total of five different TE forms. Cronbach's Alpha is reported for each of the test forms for each scoring methodology. Additionally, the standard error of measurement (*SEM*) is calculated for each form of each assessment. This allows for a better understanding of the effect of changing scoring methods on the variability and precision of the assessments.

For IRT, test information functions (TIFs) were compared to determine which methods provide the most information to the test developer. Comparing test information functions can help determine which scoring method provides the most information at any given level of Θ_i. Specifically, the higher the peak of the distribution, the more information about the test takers'

abilities can be determined.

Research Question 3: When using IRT, how does adjusting the scoring method affect the model fit? Is there a type of scoring method that allows for better model fit of the data? To answer this research question, the data was scored using all the different methods and estimated using combinations of the 2-PL and GRM. The resulting model's standardized residuals were compared to determine the best model fit.

3a: When using IRT, how does adjusting the scoring method affect the model fit when calibrating with all items? All forms were calibrated using a combination of 2-PL and GRM. Density graphs of standardized residuals were created to compare how well the models fit the data for each of the five scoring methods. These side-by-side comparisons can help determine which scoring methodology provides the best model fit.

3b: When using IRT, how does adjusting the scoring method affect the model fit when calibrating only TE items? All TE forms were calibrated using a combination of 2-PL, GRM, or 2-PL and GRM. Density graphs of standardized residuals were created to compare how well the models fit the data for each of the five scoring methods. These side by side comparisons can help determine which scoring methodology provides the best model fit.

Chapter Four: Results

The purpose of this study is to explore the effect of scoring methods on the reliability of an assessment. This chapter is organized by sections based on the main research questions. When appropriate, descriptive statistics are reported first, followed by an omnibus test, tests of main effects, and/or figures of the differences between scoring methods. Not all research questions lend themselves to tests of statistical significance. In these cases only figures were used to display the results of the analysis.

Research Question 1: How does scoring method of TE item types affect basic psychometric properties (i.e., *p*-value, *r*, IRT based stats)? Each TE item was scored using one of the six scoring methodologies. For the CTT statistics, only the correct-only, partial-credit, subtractive, threshold, and threshold-partial scoring were utilized.

Item Difficulty

***p*-values.** Across all forms and all assessments, items scored utilizing correct-only scoring tended to be more difficult than the other scoring strategies. *P*-values were calculated for each TE item on each form of the General CTE and Comprehensive Agriculture assessments for each scoring methodology. Table 3 shows the mean *p*-values for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments. The correct-only scoring methodology had the lowest mean *p*-value for both forms (General Form A: $M = .33$, $SD = .18$; General Form B: $M = .29$, $SD = .21$; Comprehensive Agriculture Form A: $M = .36$, $SD = .21$; Comprehensive Agriculture Form B: $M = .21$, $SD = .22$; Comprehensive Agriculture Form C: $M = .30$, $SD = .20$). Conversely, the highest mean *p*-value utilized the partial-credit methodology for both forms (General Form A: $M = .56$, $SD = .12$; General Form B: $M = .51$, $SD = .16$;

Comprehensive Agriculture Form A: *M* = .65, *SD* = .12; Comprehensive Agriculture Form B: *M* = .48, *SD* = .17; Comprehensive Agriculture Form C: *M* = .56, *SD* = .11).

Table 3. *Mean* p-*Values*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	.33	.18	16	.29	.21	19	.36	.21	19	.21	.22	20	.30	.20
Partial Credit	15	.56	.12	16	.51	.16	19	.65	.12	19	.48	.17	20	.56	.11
Subtractive	15	.43	.17	16	.38	.21	19	.47	.19	19	.31	.22	20	.39	.16
Threshold	15	.47	.13	16	.44	.18	19	.54	.15	19	.36	.20	20	.46	.15
Threshold Partial	15	.51	.14	16	.47	.19	19	.60	.15	19	.41	.21	20	.50	.14

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the *p*-values of the different scoring methodologies for all forms on both assessments. The Mauchly's Test of Sphericity was significant for all within-subjects ANOVAs conducted. All five omnibus tests indicate a significant difference ($p < .01$) between their mean *p*-value when utilizing different scoring strategies. Additionally, all five of the omnibus tests have large partial eta squares, with Comprehensive Agriculture Form A having the largest (partial eta squared = .775).

Table 4. *p-Value Repeated Measures ANOVA*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	.451	4	.113	43.649	.000	.757
	Error	.145	56	.003			
General Form B	Scoring	.466	4	.117	47.626	.000	.760
	Error	.147	60	.002			
Comp. Agriculture Form A	Scoring	.981	4	.245	65.307	.000	.775
	Error	.285	76	.004			
Comp. Agriculture Form B	Scoring	.765	4	.191	53.529	.000	.748
	Error	.257	72	.004			
Comp. Agriculture Form C	Scoring	.803	4	.201	59.395	.000	.758
	Error	.257	76	.003			

Each form showed a significant difference between the *p*-values utilizing a different scoring strategy. To further clarify which scoring strategies were responsible for the overall difference, a test of the main effects for each form was conducted. To reduce the chance of a type I error occurring due to multiple comparisons, a Bonferroni adjustment was utilized.

Results from the test of main effects for General Form A can be found in Table 5. The largest mean difference (.226) of *p*-values can be found between correct-only and partial-credit scoring ($p < .01$). The only non-significant difference ($p = .284$) was found between threshold and subtractive scoring. Though technically significant, the mean difference between threshold and threshold-partial scoring (.040) was very small. All other item pairings were significant at $\alpha = .05$.

Table 5. *General CTE Form A p-Value Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.226*	.026	.000	-.312	-.140
	Subtractive	-.098*	.026	.023	-.186	-.010
	Threshold	-.143*	.017	.000	-.199	-.087
	Threshold Partial	-.183*	.023	.000	-.261	-.105
Partial Credit	Correct Only	.226*	.026	.000	.140	.312
	Subtractive	.128*	.018	.000	.069	.186
	Threshold	.083*	.016	.001	.030	.137
	Threshold Partial	.043*	.010	.007	.010	.076
Subtractive	Correct Only	.098*	.026	.023	.010	.186
	Partial Credit	-.128*	.018	.000	-.186	-.069
	Threshold	-.045	.018	.284	-.106	.016
	Threshold Partial	-.085*	.011	.000	-.122	-.048
Threshold	Correct Only	.143*	.017	.000	.087	.199
	Partial Credit	-.083*	.016	.001	-.137	-.030
	Subtractive	.045	.018	.284	-.016	.106
	Threshold Partial	-.040*	.011	.030	-.077	-.003
Threshold Partial	Correct Only	.183*	.023	.000	.105	.261
	Partial Credit	-.043*	.010	.007	-.076	-.010
	Subtractive	.085*	.011	.000	.048	.122
	Threshold	.040*	.011	.030	.003	.077

Results from the test of main effects for General Form B can be found in Table 6. The largest mean difference (.221) in *p*-values was also between correct-only and partial-credit scoring ($p < .01$). The difference between subtractive and threshold scoring was just significant ($p = .048$). The mean difference in *p*-values between subtractive and threshold scoring was practically insignificant (.054). All comparisons between main effects were significant after a Bonferroni adjustment at $\alpha = .05$.

Table 6.*General CTE Form B p-Value Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.221*	.023	.000	-.298	-.145
	Subtractive	-.092*	.022	.009	-.165	-.018
	Threshold	-.146*	.021	.000	-.214	-.078
	Threshold Partial	-.177*	.024	.000	-.255	-.099
Partial Credit	Correct Only	.221*	.023	.000	.145	.298
	Subtractive	.130*	.015	.000	.081	.179
	Threshold	.076*	.014	.001	.030	.121
	Threshold Partial	.045*	.010	.005	.011	.078
Subtractive	Correct Only	.092*	.022	.009	.018	.165
	Partial Credit	-.130*	.015	.000	-.179	-.081
	Threshold	-.054*	.016	.048	-.108	.000
	Threshold Partial	-.085*	.014	.000	-.130	-.040
Threshold	Correct Only	.146*	.021	.000	.078	.214
	Partial Credit	-.076*	.014	.001	-.121	-.030
	Subtractive	.054*	.016	.048	.000	.108
	Threshold Partial	-.031*	.007	.006	-.055	-.007
Threshold Partial	Correct Only	.177*	.024	.000	.099	.255
	Partial Credit	-.045*	.010	.005	-.078	-.011
	Subtractive	.085*	.014	.000	.040	.130
	Threshold	.031*	.007	.006	.007	.055

Results from the test of main effects for Comprehensive Agriculture Form A can be found in Table 7. The largest mean difference (.286) in *p*-values was also between correct-only and partial-credit scoring ($p < .01$). All comparisons between main effects were significant after a Bonferroni adjustment ($p < .01$).

Table 7. *Comprehensive Agriculture Form A p-Value Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.286*	.031	.000	-.386	-.186
	Subtractive	-.110*	.022	.001	-.179	-.041
	Threshold	-.177*	.020	.000	-.239	-.114
	Threshold Partial	-.229*	.029	.000	-.321	-.137
Partial Credit	Correct Only	.286*	.031	.000	.186	.386
	Subtractive	.176*	.018	.000	.119	.233
	Threshold	.109*	.013	.000	.067	.152
	Threshold Partial	.057*	.008	.000	.031	.082
Subtractive	Correct Only	.110*	.022	.001	.041	.179
	Partial Credit	-.176*	.018	.000	-.233	-.119
	Threshold	-.067*	.014	.001	-.111	-.022
	Threshold Partial	-.119*	.013	.000	-.162	-.077
Threshold	Correct Only	.177*	.020	.000	.114	.239
	Partial Credit	-.109*	.013	.000	-.152	-.067
	Subtractive	.067*	.014	.001	.022	.111
	Threshold Partial	-.052*	.011	.001	-.086	-.019
Threshold Partial	Correct Only	.229*	.029	.000	.137	.321
	Partial Credit	-.057*	.008	.000	-.082	-.031
	Subtractive	.119*	.013	.000	.077	.162
	Threshold	.052*	.011	.001	.019	.086

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 8. The largest mean difference (.266) in *p*-values was between correct-only and partial-credit scoring ($p < .01$). All comparisons between main effects were significant after a Bonferroni adjustment ($p < .05$).

Table 8. *Comprehensive Agriculture Form B p-Value Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.266*	.029	.000	-.361	-.172
	Subtractive	-.096*	.022	.004	-.166	-.026
	Threshold	-.148*	.021	.000	-.214	-.081
	Threshold Partial	-.192*	.029	.000	-.284	-.099
Partial Credit	Correct Only	.266*	.029	.000	.172	.361
	Subtractive	.171*	.017	.000	.116	.225
	Threshold	.119*	.015	.000	.071	.166
	Threshold Partial	.075*	.013	.000	.034	.115
Subtractive	Correct Only	.096*	.022	.004	.026	.166
	Partial Credit	-.171*	.017	.000	-.225	-.116
	Threshold	-.052*	.014	.016	-.097	-.007
	Threshold Partial	-.096*	.013	.000	-.138	-.054
Threshold	Correct Only	.148*	.021	.000	.081	.214
	Partial Credit	-.119*	.015	.000	-.166	-.071
	Subtractive	.052*	.014	.016	.007	.097
	Threshold Partial	-.044*	.010	.003	-.076	-.012
Threshold Partial	Correct Only	.192*	.029	.000	.099	.284
	Partial Credit	-.075*	.013	.000	-.115	-.034
	Subtractive	.096*	.013	.000	.054	.138
	Threshold	.044*	.010	.003	.012	.076

Results from the test of main effects for Comprehensive Agriculture Form C can be found in Table 9. The largest mean difference (.260) in *p*-values was between correct-only and partial-credit scoring ($p < .01$). All comparisons between main effects were significant after a Bonferroni adjustment ($p < .01$).

Table 9. *Comprehensive Agriculture Form C* p-*Value Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.260*	.031	.000	-.359	-.161
	Subtractive	-.086*	.021	.006	-.152	-.020
	Threshold	-.155*	.019	.000	-.215	-.096
	Threshold Partial	-.195*	.026	.000	-.279	-.111
Partial Credit	Correct Only	.260*	.031	.000	.161	.359
	Subtractive	.174*	.016	.000	.124	.224
	Threshold	.104*	.015	.000	.056	.153
	Threshold Partial	.065*	.009	.000	.037	.093
Subtractive	Correct Only	.086*	.021	.006	.020	.152
	Partial Credit	-.174*	.016	.000	-.224	-.124
	Threshold	-.069*	.013	.000	-.110	-.029
	Threshold Partial	-.109*	.011	.000	-.145	-.073
Threshold	Correct Only	.155*	.019	.000	.096	.215
	Partial Credit	-.104*	.015	.000	-.153	-.056
	Subtractive	.069*	.013	.000	.029	.110
	Threshold Partial	-.039*	.009	.004	-.068	-.010
Threshold Partial	Correct Only	.195*	.026	.000	.111	.279
	Partial Credit	-.065*	.009	.000	-.093	-.037
	Subtractive	.109*	.011	.000	.073	.145
	Threshold	.039*	.009	.004	.010	.068

Figure 2 shows the change in mean *p*-values across the five different scoring methodologies for each of the test forms. A consistent pattern emerges from the figure. Regardless of the test or form, scoring an item correct only will provide the hardest items. Conversely, partial-credit scoring consistently provides the highest *p*-values. The other scoring methodologies are also consistent, with subtractive being more difficult than threshold and threshold partial.

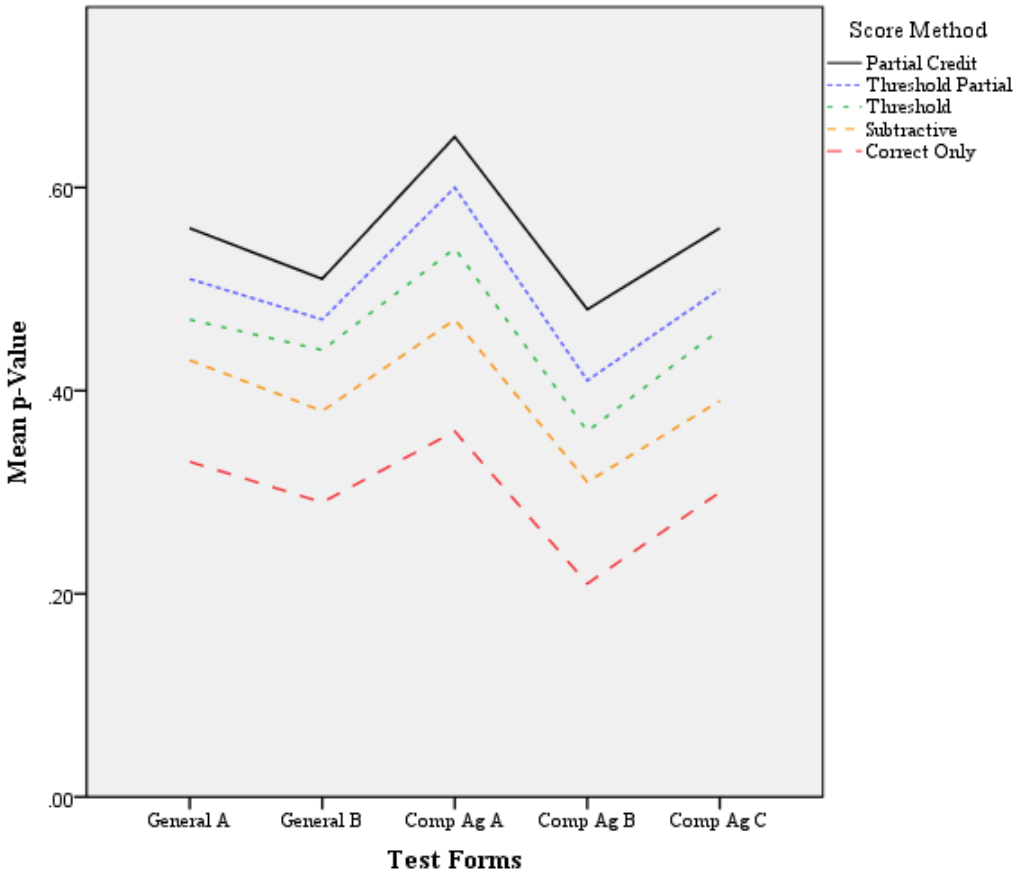


Figure 2. Mean p-values.

***b* parameters.** Mixed IRT models were estimated across all forms and all assessments. A combination of a 2-PL and GRM were utilized to estimate the *b* parameter for each item. Difficulty parameters from items with multiple categories (polytomous items) are averaged to get a single difficulty rating for analysis. Correct-only scoring had the highest *b* parameters (General Form A: *M* = 1.02, *SD* = 1.20; General Form B: *M* = 1.15, *SD* = 1.30; Comprehensive Agriculture Form A: *M* = 2.02, *SD* = 2.55; Comprehensive Agriculture Form B: *M* = 2.50, *SD* = 2.75; Comprehensive Agriculture Form C: *M* = 1.57, *SD* = 2.42). The lowest *b* parameters were from partial-credit scoring (General Form A: *M* = -0.25, *SD* = 0.56; General Form B: *M* = -0.13, *SD* = 0.65; Comprehensive Agriculture Form A: *M* = -0.16, *SD* = 0.67; Comprehensive

Agriculture Form B: *M* = 0.05, *SD* = 0.72; Comprehensive Agriculture Form C: *M* = -0.52, *SD* = 0.63). Table 10 shows the mean *b* parameters for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments when all items within each form were utilized for calibration.

Table 10. *Mean* b *Parameters, All Items*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	1.02	1.20	15	1.15	1.30	19	2.02	2.55	16	2.50	2.75	18	1.57	2.42
Partial Credit	15	-0.25	0.56	15	-0.13	0.65	19	-0.16	0.67	16	0.05	0.72	18	-0.52	0.63
Subtractive	15	-0.07	1.28	15	0.26	1.34	19	0.81	1.28	16	1.35	1.94	18	0.62	1.18
Threshold	15	0.23	0.69	15	-0.02	1.08	19	0.55	1.08	16	0.93	1.37	18	0.32	1.00
Threshold Partial	15	0.33	0.87	15	0.26	0.81	19	0.51	1.10	16	0.87	1.41	18	0.26	0.93

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the *b* parameters of the different scoring methodologies for all forms on both assessments. The Mauchly's Test of Sphericity was significant for all within-subjects ANOVAs conducted. All five omnibus tests with a Greenhouse-Geisser correction indicate a significant difference ($p <$ 01) between their mean *b* parameters when utilizing different scoring strategies (Table 11). Additionally, all five of the omnibus tests have large partial eta squares, with General Form B having the largest (partial eta squared = .669).

Table 11. b *Parameter Repeated Measures ANOVA, All Items*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	14.205	2.770	5.129	13.027	.000	.482
	Error	15.267	38.779	.394			
General Form B	Scoring	77.862	1.053	73.970	26.289	.000	.669
	Error	38.503	13.684	2.814			
Comp. Agriculture Form A	Scoring	48.492	1.169	41.476	18.345	.000	.505
	Error	47.580	21.045	2.261			
Comp. Agriculture Form B	Scoring	51.171	1.458	35.109	15.836	.000	.514
	Error	48.471	21.863	2.217			
Comp. Agriculture Form C	Scoring	40.798	1.214	33.600	15.674	.000	.480
	Error	44.249	20.642	2.144			

Each form showed a significant difference between the different scoring strategies' *b* parameters when all items were utilized in calibration. To further clarify which scoring strategies were responsible for the overall difference, a test of the main effects for each form was conducted. To reduce the chance of a type I error occurring due to multiple comparisons, a Bonferroni adjustment was utilized.

Results from the test of main effects for General Form A can be found in Table 12. The largest *b* parameter mean difference (1.266) can be found between partial-credit and correct-only scoring ($p < .01$). Additionally, correct-only scoring and threshold scoring also had a significant difference ($p = .007$). The mean difference in *b* parameters between correct-only scoring and threshold was .787. Subtractive scoring was only statistically significant when compared to correct-only scoring ($p = .001$).

Table 12. *General CTE Form A* b *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	1.266	.188	.000	.641	1.892
	Subtractive	1.087	.208	.001	.397	1.778
	Threshold	.787	.181	.007	.185	1.389
	Threshold Partial	.687	.249	.154	-.141	1.516
Partial Credit	Correct Only	-1.266	.188	.000	-1.892	-.641
	Subtractive	-.179	.204	1.000	-.856	.498
	Threshold	-.479	.071	.000	-.715	-.243
	Threshold Partial	-.579	.173	.047	-1.153	-.005
Subtractive	Correct Only	-1.087	.208	.001	-1.778	-.397
	Partial Credit	.179	.204	1.000	-.498	.856
	Threshold	-.300	.203	1.000	-.977	.376
	Threshold Partial	-.400	.226	.980	-1.151	.350
Threshold	Correct Only	-.787	.181	.007	-1.389	-.185
	Partial Credit	.479	.071	.000	.243	.715
	Subtractive	.300	.203	1.000	-.376	.977
	Threshold Partial	-.100	.147	1.000	-.589	.389
Threshold Partial	Correct Only	-.687	.249	.154	-1.516	.141
	Partial Credit	.579	.173	.047	.005	1.153
	Subtractive	.400	.226	.980	-.350	1.151
	Threshold	.100	.147	1.000	-.389	.589

Results from the test of main effects for General Form B can be found in Table 13. The largest *b* parameter mean difference (1.280) can be found between partial-credit and correct-only scoring ($p < .01$). Additionally, correct-only and threshold scoring also had a significant difference ($p < .01$). The mean difference in *b* parameters between correct-only and threshold scoring was 1.171. Subtractive scoring was only statistically significant when compared to correct-only scoring ($p = .003$).

Table 13. *General CTE Form B* b *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	1.280	.212	.000	.575	1.985
	Subtractive	.889	.184	.003	.276	1.503
	Threshold	1.171	.129	.000	.742	1.600
	Threshold Partial	.893	.173	.001	.316	1.469
Partial Credit	Correct Only	-1.280	.212	.000	-1.985	-.575
	Subtractive	-.391	.195	.643	-1.038	.256
	Threshold	-.109	.139	1.000	-.570	.351
	Threshold Partial	-.387	.073	.001	-.629	-.145
Subtractive	Correct Only	-.889	.184	.003	-1.503	-.276
	Partial Credit	.391	.195	.643	-.256	1.038
	Threshold	.281	.167	1.000	-.273	.835
	Threshold Partial	.003	.181	1.000	-.597	.604
Threshold	Correct Only	-1.171	.129	.000	-1.600	-.742
	Partial Credit	.109	.139	1.000	-.351	.570
	Subtractive	-.281	.167	1.000	-.835	.273
	Threshold Partial	-.278	.077	.029	-.535	-.021
Threshold Partial	Correct Only	-.893	.173	.001	-1.469	-.316
	Partial Credit	.387	.073	.001	.145	.629
	Subtractive	-.003	.181	1.000	-.604	.597
	Threshold	.278	.077	.029	.021	.535

Results from the test of main effects for Comprehensive Agriculture Form A can be found in Table 14. The largest *b* parameter mean difference (2.183) can be found between partial-credit and correct-only scoring ($p = .002$). Additionally, correct-only and threshold-partial scoring also had a significant difference ($p = .007$). The mean difference in *b* parameters between correct-only and threshold-partial scoring was 1.516. Similarly, the difference between correct-

only and threshold scoring was statistically different (p = .010). The mean difference between *b* parameters was 1.474. Subtractive scoring was not statistically significant when compared to threshold scoring (p = .239).

Table 14. *Comprehensive Agriculture Form A* b *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	2.183	.461	.002	.710	3.656
	Subtractive	1.209	.372	.045	.019	2.399
	Threshold	1.474	.375	.010	.274	2.674
	Threshold Partial	1.516	.369	.007	.336	2.695
Partial Credit	Correct Only	-2.183	.461	.002	-3.656	-.710
	Subtractive	-.974	.153	.000	-1.464	-.484
	Threshold	-.709	.105	.000	-1.044	-.374
	Threshold Partial	-.668	.108	.000	-1.011	-.324
Subtractive	Correct Only	-1.209	.372	.045	-2.399	-.019
	Partial Credit	.974	.153	.000	.484	1.464
	Threshold	.265	.107	.239	-.078	.608
	Threshold Partial	.306	.096	.049	.001	.612
Threshold	Correct Only	-1.474	.375	.010	-2.674	-.274
	Partial Credit	.709	.105	.000	.374	1.044
	Subtractive	-.265	.107	.239	-.608	.078
	Threshold Partial	.042	.033	1.000	-.063	.147
Threshold Partial	Correct Only	-1.516	.369	.007	-2.695	-.336
	Partial Credit	.668	.108	.000	.324	1.011
	Subtractive	-.306	.096	.049	-.612	-.001
	Threshold	-.042	.033	1.000	-.147	.063

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 15. The largest *b* parameter mean difference (2.449) can be found between

partial-credit and correct-only scoring (p = .003). Additionally, correct-only and threshold-partial scoring also had a significant difference (p = .015). The mean difference in b parameters between correct-only and threshold-partial scoring was 1.630. Similarly, the difference between correct-only and threshold scoring was statistically different (p = .017). The mean difference between b parameters was 1.577. Subtractive scoring was not statistically significant when compared to threshold scoring (p = .757) or threshold-partial scoring (p = .376).

Table 15. *Comprehensive Agriculture Form B* b *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	2.449	.531	.003	.705	4.193
	Subtractive	1.156	.318	.025	.109	2.202
	Threshold	1.577	.415	.017	.213	2.941
	Threshold Partial	1.630	.423	.015	.242	3.019
Partial Credit	Correct Only	-2.449	.531	.003	-4.193	-.705
	Subtractive	-1.294	.335	.015	-2.393	-.194
	Threshold	-.872	.179	.002	-1.461	-.283
	Threshold Partial	-.819	.194	.007	-1.456	-.182
Subtractive	Correct Only	-1.156	.318	.025	-2.202	-.109
	Partial Credit	1.294	.335	.015	.194	2.393
	Threshold	.421	.221	.757	-.304	1.147
	Threshold Partial	.475	.208	.376	-.210	1.159
Threshold	Correct Only	-1.577	.415	.017	-2.941	-.213
	Partial Credit	.872	.179	.002	.283	1.461
	Subtractive	-.421	.221	.757	-1.147	.304
	Threshold Partial	.053	.045	1.000	-.095	.202
Threshold Partial	Correct Only	-1.630	.423	.015	-3.019	-.242
	Partial Credit	.819	.194	.007	.182	1.456
	Subtractive	-.475	.208	.376	-1.159	.210
	Threshold	-.053	.045	1.000	-.202	.095

Results from the test of main effects for Comprehensive Agriculture Form C can be found in Table 16. The largest *b* parameter mean difference (2.084) can be found between partial-credit and correct-only scoring ($p = .003$). Additionally, correct-only and threshold-partial scoring also had a significant difference ($p = .038$). The mean difference in *b* parameters between correct-only and threshold-partial scoring was 1.305. Subtractive scoring was only statistically

significant when compared to partial-credit scoring (p = .003). The mean difference in *b* parameters was 1.141.

Table 16. *Comprehensive Agriculture Form C* b *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	2.084	.462	.003	.596	3.573
	Subtractive	.943	.324	.098	-.102	1.988
	Threshold	1.250	.390	.052	-.006	2.506
	Threshold Partial	1.305	.389	.038	.050	2.560
Partial Credit	Correct Only	-2.084	.462	.003	-3.573	-.596
	Subtractive	-1.141	.184	.000	-1.733	-.549
	Threshold	-.834	.112	.000	-1.196	-.473
	Threshold Partial	-.780	.097	.000	-1.093	-.466
Subtractive	Correct Only	-.943	.324	.098	-1.988	.102
	Partial Credit	1.141	.184	.000	.549	1.733
	Threshold	.307	.156	.655	-.195	.809
	Threshold Partial	.361	.139	.188	-.087	.810
Threshold	Correct Only	-1.250	.390	.052	-2.506	.006
	Partial Credit	.834	.112	.000	.473	1.196
	Subtractive	-.307	.156	.655	-.809	.195
	Threshold Partial	.055	.037	1.000	-.065	.174
Threshold Partial	Correct Only	-1.305	.389	.038	-2.560	-.050
	Partial Credit	.780	.097	.000	.466	1.093
	Subtractive	-.361	.139	.188	-.810	.087
	Threshold	-.055	.037	1.000	-.174	.065

Figure 3 shows the change in mean *b* parameters across the five different scoring methods for each of the test forms utilizing all items. Across all scoring methodologies and forms, except threshold, threshold partial, and subtractive scoring for the General CTE forms,

there is a consistent pattern. Threshold, threshold partial, and subtractive scoring for the General forms are not consistent. For Form A, subtractive-scored items were easier and threshold-scored items were harder. Form B was opposite, where subtractive-scored items were harder and threshold-scored items were easier. All other scoring methods and forms were consistent in mean item difficulty. Partial-credit scoring provided the lowest *b* parameters across all items and all forms, whereas correct-only scoring was consistently the hardest scoring methodology.

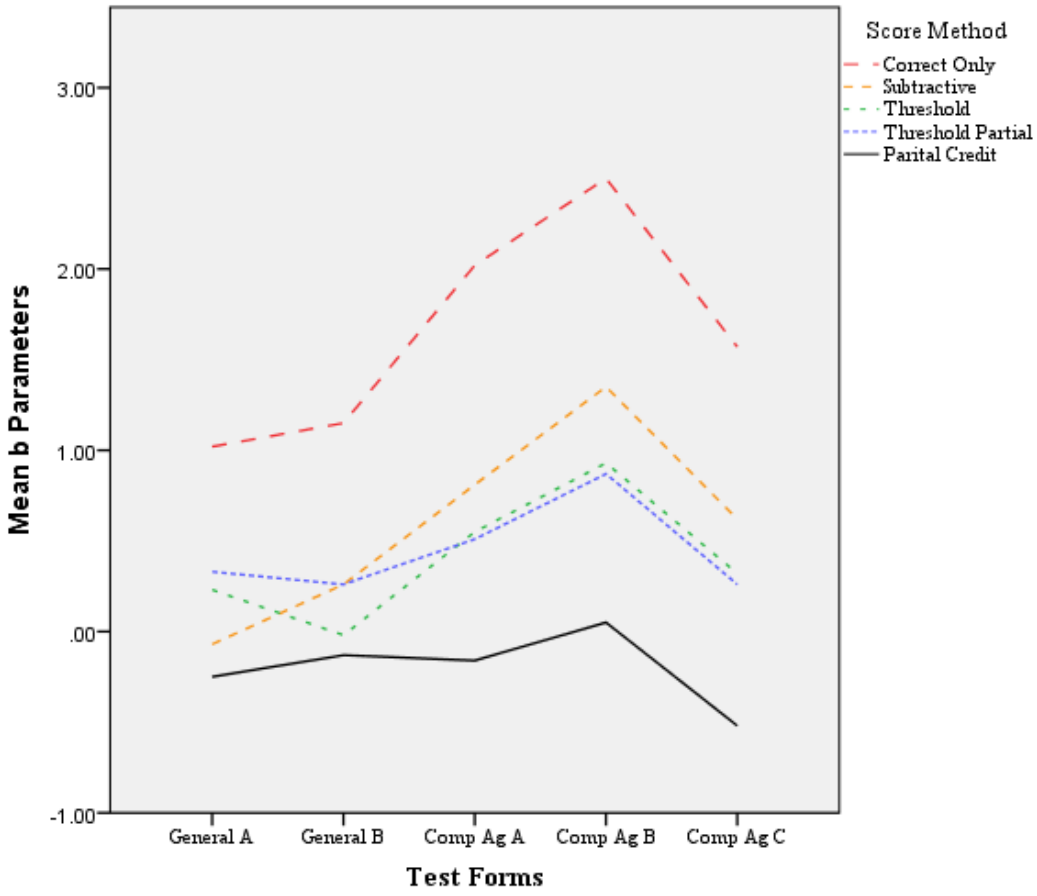


Figure 3. Mean *b* parameters by form and scoring methodology, all items.

Table 17 shows the mean *b* parameters for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments when only TE items within each form were utilized for calibration. As with previous analysis, correct-only scoring had the highest mean *b* parameters (General Form A: *M* = 0.96, *SD* = 1.31; General Form B: *M* = 1.06, *SD* = 1.26; Comprehensive Agriculture Form A: *M* = 1.98, *SD* = 2.98; Comprehensive Agriculture Form B: *M* = 2.63, *SD* = 2.98; Comprehensive Agriculture Form C: *M* = 1.32, *SD* = 2.09). The lowest mean *b* parameters were from partial-credit scoring (General Form A: *M* = -0.07, SD=0.41; General Form B: *M* = -0.58, SD=0.76; Comprehensive Agriculture Form A: *M* = -0.11, *SD* = 0.55; Comprehensive Agriculture Form B: *M* = 0.10, *SD* = 0.51; Comprehensive Agriculture Form C: *M* = -0.43, *SD* = 0.53).

Table 17. *Mean* b *Parameters, Tech Only*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	0.96	1.31	15	1.06	1.26	19	1.98	2.98	16	2.63	2.90	18	1.32	2.09
Partial Credit	15	-0.07	0.41	15	-0.58	0.76	19	-0.11	0.55	16	0.10	0.51	18	-0.43	0.53
Subtractive	15	0.36	0.69	15	0.27	1.18	19	0.77	1.18	16	1.07	1.43	18	0.54	1.00
Threshold	15	0.26	0.56	15	-0.04	0.89	19	0.49	0.95	16	0.74	0.99	18	0.28	0.85
Threshold Partial	15	0.33	0.67	15	-0.09	0.86	19	0.47	0.98	16	0.72	1.02	18	0.23	0.79
Testlet Response	15	-0.69	0.74	15	0.20	0.51	19	0.02	0.67	16	0.00	0.70	18	-0.20	0.59

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the *b* parameters of the different scoring methodologies for all forms on both assessments when only TE items were utilized in calibration. The Mauchly's Test of Sphericity was significant for all within-subjects ANOVAs conducted. All five omnibus tests with a Greenhouse-Geisser

correction indicate a significant difference ($p < 01$) between their mean *b* parameters when utilizing different scoring strategies. Additionally, four out of the five omnibus tests have large partial eta squares, with General CTE Form B having the largest (partial eta squared = .648).

Table 18. b *Parameter Repeated Measures ANOVA, Tech Only*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	22.321	2.445	9.130	17.525	.000	.556
	Error	17.831	34.229	.521			
General Form B	Scoring	22.066	2.650	8.326	25.724	.000	.648
	Error	12.009	37.102	.324			
Comp. Agriculture Form A	Scoring	53.270	1.079	49.364	11.399	.003	.388
	Error	84.117	19.424	4.331			
Comp. Agriculture Form B	Scoring	72.451	1.213	59.721	16.713	.000	.527
	Error	65.023	18.197	3.573			
Comp. Agriculture Form C	Scoring	33.890	1.224	27.696	15.142	.000	.471
	Error	38.049	20.802	1.829			

Each form showed a significant difference between the different scoring strategies' *b* parameters. To further clarify which scoring strategies were responsible for the overall difference, a test of the main effects for each form was conducted. To reduce the chance of a type I error occurring due to multiple comparisons, a Bonferroni adjustment was utilized.

Results from the test of main effects for General Form A can be found in Table 19. The largest *b* parameter mean difference (1.650) can be found between testlet response and correct-only scoring ($p < .01$). Additionally, correct-only and partial-credit scoring also had a significant difference ($p = .014$). The mean difference in *b* parameters between correct-only and partial-credit scoring was 1.030. Subtractive scoring did not have a statistically significant difference when compared to any of the other scoring methods other than testlet response ($p = .002$).

Finally, testlet response had *b* parameters that were statistically different than all other scoring methods.

Table 19. *General CTE Form A* b *Parameter Test of Main Effects, Tech Only*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	1.030	.247	.014	.158	1.903
	Subtractive	.599	.288	.850	-.419	1.616
	Threshold	.698	.224	.115	-.094	1.489
	Threshold Partial	.632	.260	.434	-.285	1.549
	Testlet Response	1.650	.208	.000	.915	2.384
Partial Credit	Correct Only	-1.030	.247	.014	-1.903	-.158
	Subtractive	-.432	.161	.271	-1.001	.138
	Threshold	-.333	.067	.003	-.570	-.095
	Threshold Partial	-.398	.104	.028	-.766	-.030
	Testlet Response	.619	.092	.000	.296	.943
Subtractive	Correct Only	-.599	.288	.850	-1.616	.419
	Partial Credit	.432	.161	.271	-.138	1.001
	Threshold	.099	.188	1.000	-.566	.763
	Threshold Partial	.034	.204	1.000	-.687	.754
	Testlet Response	1.051	.203	.002	.336	1.766
Threshold	Correct Only	-.698	.224	.115	-1.489	.094
	Partial Credit	.333	.067	.003	.095	.570
	Subtractive	-.099	.188	1.000	-.763	.566
	Threshold Partial	-.065	.101	1.000	-.423	.293
	Testlet Response	.952	.104	.000	.587	1.318
Threshold Partial	Correct Only	-.632	.260	.434	-1.549	.285
	Partial Credit	.398	.104	.028	.030	.766
	Subtractive	-.034	.204	1.000	-.754	.687
	Threshold	.065	.101	1.000	-.293	.423
	Testlet Response	1.017	.124	.000	.580	1.455
Testlet Response	Correct Only	-1.650	.208	.000	-2.384	-.915
	Partial Credit	-.619	.092	.000	-.943	-.296
	Subtractive	-1.051	.203	.002	-1.766	-.336
	Threshold	-.952	.104	.000	-1.318	-.587
	Threshold Partial	-1.017	.124	.000	-1.455	-.580

Results from the test of main effects for General Form B can be found in Table 20. The largest *b* parameter mean difference (1.642) can be found between partial-credit and correct-only scoring ($p < .01$). Additionally, correct-only scoring and threshold-partial scoring also had a significant difference ($p < .01$). The mean difference in *b* parameters between correct-only scoring and testlet response was 1.148. Testlet response was not statistically significant when compared to subtractive scoring ($p = 1.00$) and threshold scoring ($p = 1.00$).

Table 20. *General CTE Form B* b *Parameter Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	1.642	.198	.000	.943	2.341
	Subtractive	.794	.172	.006	.187	1.400
	Threshold	1.100	.157	.000	.545	1.656
	Threshold Partial	1.148	.171	.000	.545	1.751
	Testlet Response	.864	.216	.020	.103	1.626
Partial Credit	Correct Only	-1.642	.198	.000	-2.341	-.943
	Subtractive	-.848	.176	.004	-1.471	-.226
	Threshold	-.542	.104	.002	-.909	-.174
	Threshold Partial	-.494	.092	.001	-.818	-.171
	Testlet Response	-.778	.090	.000	-1.096	-.459
Subtractive	Correct Only	-.794	.172	.006	-1.400	-.187
	Partial Credit	.848	.176	.004	.226	1.471
	Threshold	.307	.164	1.000	-.272	.886
	Threshold Partial	.354	.158	.625	-.203	.912
	Testlet Response	.071	.184	1.000	-.579	.721
Threshold	Correct Only	-1.100	.157	.000	-1.656	-.545
	Partial Credit	.542	.104	.002	.174	.909
	Subtractive	-.307	.164	1.000	-.886	.272
	Threshold Partial	.047	.030	1.000	-.058	.153
	Testlet Response	-.236	.123	1.000	-.670	.198
Threshold Partial	Correct Only	-1.148	.171	.000	-1.751	-.545
	Partial Credit	.494	.092	.001	.171	.818
	Subtractive	-.354	.158	.625	-.912	.203
	Threshold	-.047	.030	1.000	-.153	.058
	Testlet Response	-.283	.116	.435	-.695	.128
Testlet Response	Correct Only	-.864	.216	.020	-1.626	-.103
	Partial Credit	.778	.090	.000	.459	1.096
	Subtractive	-.071	.184	1.000	-.721	.579
	Threshold	.236	.123	1.000	-.198	.670
	Threshold Partial	.283	.116	.435	-.128	.695

Results from the test of main effects for Comprehensive Agriculture Form A can be found in Table 21. The largest *b* parameter mean difference (2.090) can be found between partial-credit and correct-only scoring ($p = .032$). Additionally, correct-only scoring and testlet response theory also had a significant difference ($p = .042$). The mean difference in *b* parameters between correct-only scoring and testlet response was 1.960. Subtractive scoring was not statistically significant when compared to correct-only scoring ($p = 0.302$) and threshold scoring ($p = 0.158$). Finally, testlet response was significantly different than all other scoring methods except partial-credit scoring ($p = 0.120$)

Table 21. *Comprehensive Agriculture Form A* b *Parameter Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	2.090	.585	.032	.114	4.066
	Subtractive	1.212	.476	.302	-.395	2.820
	Threshold	1.496	.501	.120	-.199	3.190
	Threshold Partial	1.513	.489	.095	-.142	3.167
	Testlet Response	1.960	.566	.042	.047	3.873
Partial Credit	Correct Only	-2.090	.585	.032	-4.066	-.114
	Subtractive	-.878	.156	.000	-1.404	-.351
	Threshold	-.594	.099	.000	-.930	-.258
	Threshold Partial	-.577	.106	.001	-.937	-.218
	Testlet Response	-.130	.044	.120	-.278	.017
Subtractive	Correct Only	-1.212	.476	.302	-2.820	.395
	Partial Credit	.878	.156	.000	.351	1.404
	Threshold	.283	.099	.158	-.052	.619
	Threshold Partial	.300	.087	.043	.006	.595
	Testlet Response	.747	.128	.000	.315	1.180
Threshold	Correct Only	-1.496	.501	.120	-3.190	.199
	Partial Credit	.594	.099	.000	.258	.930
	Subtractive	-.283	.099	.158	-.619	.052
	Threshold Partial	.017	.030	1.000	-.084	.118
	Testlet Response	.464	.082	.000	.187	.741
Threshold Partial	Correct Only	-1.513	.489	.095	-3.167	.142
	Partial Credit	.577	.106	.001	.218	.937
	Subtractive	-.300	.087	.043	-.595	-.006
	Threshold	-.017	.030	1.000	-.118	.084
	Testlet Response	.447	.089	.001	.147	.747
Testlet Response	Correct Only	-1.960	.566	.042	-3.873	-.047
	Partial Credit	.130	.044	.120	-.017	.278
	Subtractive	-.747	.128	.000	-1.180	-.315
	Threshold	-.464	.082	.000	-.741	-.187
	Threshold Partial	-.447	.089	.001	-.747	-.147

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 22. The largest *b* parameter mean difference (2.628) can be found between testlet response and correct-only scoring ($p = .006$). Additionally, correct-only and partial-credit scoring also had a significant difference ($p = .012$). The mean difference in *b* parameters between correct only and partial credit was 2.532. Subtractive scoring was only statistically significant when compared to partial-credit scoring ($p = 0.020$) and testlet response ($p = 0.002$). Finally, testlet response was significantly different than all other scoring methods except partial credit ($p = 1.00$)

Table 22. *Comprehensive Agriculture Form B* b *Parameter Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	2.532	.608	.012	.413	4.651
	Subtractive	1.560	.457	.058	-.032	3.151
	Threshold	1.892	.494	.024	.172	3.611
	Threshold Partial	1.916	.500	.025	.173	3.659
	Testlet Response	2.628	.575	.006	.624	4.631
Partial Credit	Correct Only	-2.532	.608	.012	-4.651	-.413
	Subtractive	-.972	.247	.020	-1.833	-.112
	Threshold	-.640	.131	.003	-1.097	-.184
	Threshold Partial	-.616	.142	.009	-1.109	-.123
	Testlet Response	.096	.064	1.000	-.129	.320
Subtractive	Correct Only	-1.560	.457	.058	-3.151	.032
	Partial Credit	.972	.247	.020	.112	1.833
	Threshold	.332	.164	.915	-.239	.903
	Threshold Partial	.357	.152	.498	-.173	.886
	Testlet Response	1.068	.205	.002	.354	1.783
Threshold	Correct Only	-1.892	.494	.024	-3.611	-.172
	Partial Credit	.640	.131	.003	.184	1.097
	Subtractive	-.332	.164	.915	-.903	.239
	Threshold Partial	.025	.037	1.000	-.104	.153
	Testlet Response	.736	.111	.000	.351	1.121
Threshold Partial	Correct Only	-1.916	.500	.025	-3.659	-.173
	Partial Credit	.616	.142	.009	.123	1.109
	Subtractive	-.357	.152	.498	-.886	.173
	Threshold	-.025	.037	1.000	-.153	.104
	Testlet Response	.712	.120	.000	.294	1.130
Testlet Response	Correct Only	-2.628	.575	.006	-4.631	-.624
	Partial Credit	-.096	.064	1.000	-.320	.129
	Subtractive	-1.068	.205	.002	-1.783	-.354
	Threshold	-.736	.111	.000	-1.121	-.351
	Threshold Partial	-.712	.120	.000	-1.130	-.294

Results from the test of main effects for Comprehensive Agriculture Form C can be found in Table 23. The largest *b* parameter mean difference (1.745) can be found between partial-credit and correct-only scoring ($p = .006$). Additionally, correct-only scoring and testlet response also had a significant difference ($p = .022$). The mean difference in *b* parameters between correct-only scoring and testlet response was 1.522. Subtractive scoring was only statistically significant when compared to partial-credit scoring ($p < .01$) and testlet response ($p = .005$). Finally, testlet response was significantly different than all other scoring methods.

Table 23. *Comprehensive Agriculture Form C* b *Parameter Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	1.745	.400	.006	.381	3.108
	Subtractive	.774	.284	.218	-.196	1.744
	Threshold	1.043	.330	.085	-.081	2.166
	Threshold Partial	1.090	.333	.067	-.045	2.225
	Testlet Response	1.522	.403	.022	.149	2.895
Partial Credit	Correct Only	-1.745	.400	.006	-3.108	-.381
	Subtractive	-.971	.163	.000	-1.525	-.416
	Threshold	-.702	.095	.000	-1.026	-.377
	Threshold Partial	-.654	.085	.000	-.944	-.365
	Testlet Response	-.223	.042	.001	-.367	-.078
Subtractive	Correct Only	-.774	.284	.218	-1.744	.196
	Partial Credit	.971	.163	.000	.416	1.525
	Threshold	.269	.136	.968	-.195	.733
	Threshold Partial	.316	.121	.274	-.097	.729
	Testlet Response	.748	.168	.005	.174	1.322
Threshold	Correct Only	-1.043	.330	.085	-2.166	.081
	Partial Credit	.702	.095	.000	.377	1.026
	Subtractive	-.269	.136	.968	-.733	.195
	Threshold Partial	.047	.031	1.000	-.058	.153
	Testlet Response	.479	.094	.001	.158	.800
Threshold Partial	Correct Only	-1.090	.333	.067	-2.225	.045
	Partial Credit	.654	.085	.000	.365	.944
	Subtractive	-.316	.121	.274	-.729	.097
	Threshold	-.047	.031	1.000	-.153	.058
	Testlet Response	.432	.090	.003	.123	.740
Testlet Response	Correct Only	-1.522	.403	.022	-2.895	-.149
	Partial Credit	.223	.042	.001	.078	.367
	Subtractive	-.748	.168	.005	-1.322	-.174
	Threshold	-.479	.094	.001	-.800	-.158
	Threshold Partial	-.432	.090	.003	-.740	-.123

Figure 4 shows the change in mean *b* parameters across the six different scoring methods for each of the test forms utilizing only TE items. All forms, except General Form B and Comprehensive Agriculture Form B, show a consistent pattern of *b* parameters across the six different scoring methodologies. General Form B had higher *b* parameters for testlet-response scoring than the other forms. Additionally, Comprehensive Agriculture Form B had lower *b* parameters than partial-credit scoring. Regardless of the test or form, scoring an item correct only will provide the hardest items. Conversely, partial-credit scoring and testlet response consistently provide the lowest *b* parameters. The other scoring methodologies are also consistent, with subtractive being more difficult than threshold and threshold-partial scoring.

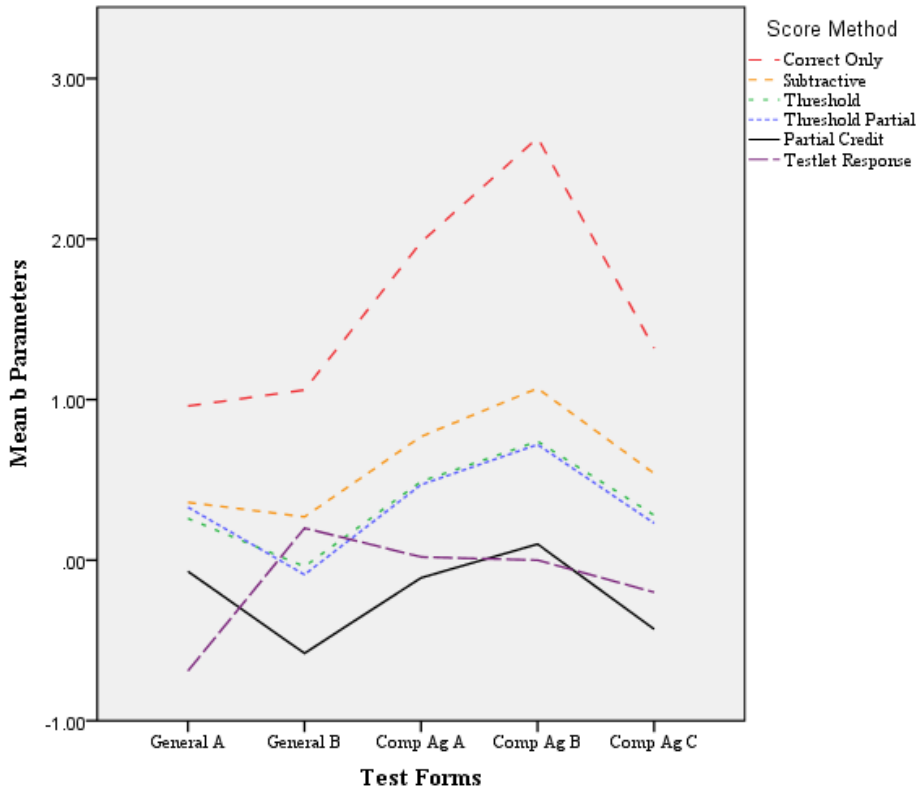


Figure 4. Mean *b* parameters by form and scoring methodology, tech only.

Item Discrimination

Item-total correlations. Item-total correlations were calculated for each TE item on each form of the General CTE and Comprehensive Agriculture assessments for each scoring methodology. Each TE item was correlated with both the total score of all items on the form, and the total score of all TE items within each scoring methodology. Table 24 shows the mean item-total correlations for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments using all items. The correct-only scoring methodology had the lowest mean item-total correlation for all forms on both assessments when all items were used (General Form A: M = -0.44, SD = 0.14; General Form B: M = 0.43, SD = 0.15; Comprehensive

Agriculture Form A: *M* = .42, *SD* = 0.10; Comprehensive Agriculture Form B: *M* = 0.35, *SD* = 0.17; Comprehensive Agriculture Form C: *M* = .45, *SD* = 0.18). Conversely, the highest mean item-total correlation utilized the partial-credit methodology for all forms on both assessments when all items were used (General Form A: *M* = 0.63, *SD* = 0.14; General Form B: *M* = 0.59, *SD* = 0.09; Comprehensive Agriculture Form A: *M* = .47, *SD* = 0.13; Comprehensive Agriculture Form B: *M* = 0.59, *SD* = 0.09; Comprehensive Agriculture Form C: *M* = 0.55, *SD* = 0.11).

Table 24. *Mean Item-Total Correlations, All Items*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	0.44	0.14	15	0.43	0.15	20	0.42	0.10	16	0.35	0.17	18	0.45	0.18
Partial Credit	15	0.63	0.06	15	0.59	0.09	20	0.47	0.13	16	0.59	0.09	18	0.55	0.11
Subtractive	15	0.55	0.10	15	0.53	0.11	20	0.45	0.13	16	0.45	0.16	18	0.51	0.14
Threshold	15	0.58	0.07	15	0.55	0.09	20	0.46	0.12	16	0.51	0.12	18	0.53	0.12
Threshold Partial	15	0.59	0.07	15	0.57	0.09	20	0.46	0.13	16	0.52	0.12	18	0.53	0.12

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the item-total correlations (calculated with all items) of the different scoring methodologies for all forms on both assessments. The Mauchly's Test of Sphericity was significant for all within-subjects ANOVAs conducted. All five omnibus tests (Table 25) with a Greenhouse-Geisser correction indicate a significant difference (p<01) between their mean item-total correlations when utilizing different scoring strategies. Additionally, all five of the omnibus tests have medium to large partial eta squared, with Comprehensive Agriculture Form B having the largest (partial eta squared= .701).

Table 25. *Item-Total Correlation, All Items Repeated Measures ANOVA*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	.299	1.592	.188	26.267	.000	.652
	Error	.160	22.289	.007			
General Form B	Scoring	.239	1.645	.145	19.410	.000	.581
	Error	.172	23.030	.007			
Comp. Agriculture Form A	Scoring	.028	2.183	.013	8.476	.001	.308
	Error	.063	41.470	.002			
Comp. Agriculture Form B	Scoring	.507	1.720	.295	35.168	.000	.701
	Error	.216	25.799	.008			
Comp. Agriculture Form C	Scoring	.124	1.357	.091	10.846	.001	.389
	Error	.194	23.062	.008			

Each form showed a significant difference between the different scoring strategies' item-total correlations. To further clarify which scoring strategies were responsible for the overall difference, a test of the main effects for each form was conducted. To reduce the chance of a type I error occurring due to multiple comparisons, a Bonferroni adjustment was utilized.

Results from the test of main effects for General Form A can be found in Table 26 The largest item-total correlation mean difference (.185) can be found between correct-only and partial-credit scoring ($p < .01$). Additionally, correct-only and threshold-partial scoring also had a significant difference ($p = .001$). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.151. Subtractive scoring did not have a statistically significant difference when compared to threshold scoring ($p = .322$).

Table 26. *General CTE Form A Item-Total Correlation, All Items Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.185*	.032	.000	-.292	-.079
	Subtractive	-.106*	.025	.007	-.187	-.024
	Threshold	-.136*	.025	.001	-.219	-.053
	Threshold Partial	-.151*	.026	.001	-.239	-.063
Partial Credit	Correct Only	.185*	.032	.000	.079	.292
	Subtractive	.079*	.017	.004	.023	.136
	Threshold	.049*	.012	.014	.008	.091
	Threshold Partial	.034	.012	.113	-.005	.073
Subtractive	Correct Only	.106*	.025	.007	.024	.187
	Partial Credit	-.079*	.017	.004	-.136	-.023
	Threshold	-.030	.013	.322	-.072	.012
	Threshold Partial	-.045*	.010	.005	-.078	-.012
Threshold	Correct Only	.136*	.025	.001	.053	.219
	Partial Credit	-.049*	.012	.014	-.091	-.008
	Subtractive	.030	.013	.322	-.012	.072
	Threshold Partial	-.015*	.004	.020	-.029	-.002
Threshold Partial	Correct Only	.151*	.026	.001	.063	.239
	Partial Credit	-.034	.012	.113	-.073	.005
	Subtractive	.045*	.010	.005	.012	.078
	Threshold	.015*	.004	.020	.002	.029

Results from the test of main effects for General Form B can be found in Table 27 The largest item-total correlation mean difference (.165) can be found between correct-only and partial-credit scoring (p = .002). Additionally, correct-only and threshold-partial scoring also had a significant difference (p = .002). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.136. Subtractive scoring did not have a statistically significant difference when compared to threshold scoring (p = 1.00).

Table 27. *General CTE Form B Item-Total Correlation, All Items Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.165*	.032	.002	-.272	-.058
	Subtractive	-.095*	.026	.022	-.180	-.010
	Threshold	-.121*	.025	.003	-.205	-.036
	Threshold Partial	-.136*	.028	.002	-.229	-.044
Partial Credit	Correct Only	.165*	.032	.002	.058	.272
	Subtractive	.070*	.017	.013	.012	.128
	Threshold	.044*	.011	.012	.008	.080
	Threshold Partial	.029	.009	.085	-.002	.060
Subtractive	Correct Only	.095*	.026	.022	.010	.180
	Partial Credit	-.070*	.017	.013	-.128	-.012
	Threshold	-.025	.016	1.00	-.078	.027
	Threshold Partial	-.041	.015	.155	-.090	.008
Threshold	Correct Only	.121*	.025	.003	.036	.205
	Partial Credit	-.044*	.011	.012	-.080	-.008
	Subtractive	.025	.016	1.00	-.027	.078
	Threshold Partial	-.015*	.004	.018	-.029	-.002
Threshold Partial	Correct Only	.136*	.028	.002	.044	.229
	Partial Credit	-.029	.009	.085	-.060	.002
	Subtractive	.041	.015	.155	-.008	.090
	Threshold	.015*	.004	.018	.002	.029

Results from the test of main effects for Comprehensive Agriculture Form A can be found in Table 28 The largest item-total correlation mean difference (.045) can be found between correct-only and partial-credit scoring (p = .004). Additionally, correct-only and threshold-partial scoring also had a significant difference (p = .031). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.039. Subtractive scoring

was only significantly different from correct-only scoring (p = .015); all other comparisons were not significant.

Table 28. *Comprehensive Agriculture Form A Item-Total Correlation, All Items Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.049*	.011	.004	-.085	-.013
	Subtractive	-.033*	.009	.015	-.062	-.005
	Threshold	-.037*	.010	.014	-.069	-.006
	Threshold Partial	-.039*	.012	.031	-.076	-.002
Partial Credit	Correct Only	.049*	.011	.004	.013	.085
	Subtractive	.016	.009	1.00	-.014	.046
	Threshold	.011	.006	.586	-.007	.030
	Threshold Partial	.010	.006	1.00	-.009	.028
Subtractive	Correct Only	.033*	.009	.015	.005	.062
	Partial Credit	-.016	.009	1.00	-.046	.014
	Threshold	-.004	.010	1.00	-.037	.029
	Threshold Partial	-.006	.010	1.00	-.038	.026
Threshold	Correct Only	.037*	.010	.014	.006	.069
	Partial Credit	-.011	.006	.586	-.030	.007
	Subtractive	.004	.010	1.00	-.029	.037
	Threshold Partial	-.002	.003	1.00	-.012	.008
Threshold Partial	Correct Only	.039*	.012	.031	.002	.076
	Partial Credit	-.010	.006	1.00	-.028	.009
	Subtractive	.006	.010	1.00	-.026	.038
	Threshold	.002	.003	1.00	-.008	.012

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 29. The largest item-total correlation mean difference (.240) can be found between correct-only and partial-credit scoring (p < .01). Additionally, correct-only and threshold-partial scoring also had a significant difference (p < .01). The mean difference in item-

total correlations between correct-only and threshold-partial scoring was 0.168. Subtractive scoring was significantly different from all other scoring methods.

Table 29. *Comprehensive Agriculture Form B Item-Total Correlation, All Items Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.240*	.035	.000	-.355	-.125
	Subtractive	-.103*	.021	.002	-.173	-.034
	Threshold	-.155*	.025	.000	-.236	-.073
	Threshold Partial	-.168*	.026	.000	-.252	-.084
Partial Credit	Correct Only	.240*	.035	.000	.125	.355
	Subtractive	.136*	.024	.001	.056	.217
	Threshold	.085*	.016	.001	.033	.138
	Threshold Partial	.072*	.017	.008	.016	.128
Subtractive	Correct Only	.103*	.021	.002	.034	.173
	Partial Credit	-.136*	.024	.001	-.217	-.056
	Threshold	-.051*	.015	.031	-.099	-.003
	Threshold Partial	-.065*	.014	.003	-.110	-.019
Threshold	Correct Only	.155*	.025	.000	.073	.236
	Partial Credit	-.085*	.016	.001	-.138	-.033
	Subtractive	.051*	.015	.031	.003	.099
	Threshold Partial	-.013*	.003	.003	-.022	-.004
Threshold Partial	Correct Only	.168*	.026	.000	.084	.252
	Partial Credit	-.072*	.017	.008	-.128	-.016
	Subtractive	.065*	.014	.003	.019	.110
	Threshold	.013*	.003	.003	.004	.022

Results from the test of main effects for Comprehensive Agriculture Form C can be found in Table 30 The largest item-total correlation mean difference (0.108) can be found between correct-only and partial-credit scoring (p = .022). Additionally, correct-only and threshold-partial scoring also had a significant difference (p = .026). The mean difference in

item-total correlations between correct-only and threshold-partial scoring was 0.089. Subtractive scoring was only significantly different from correct-only scoring (p = .020).

Table 30. *Comprehensive Agriculture Form C Item- Total-Correlation, All Items Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.108*	.030	.022	-.206	-.011
	Subtractive	-.064*	.018	.020	-.121	-.008
	Threshold	-.081*	.023	.026	-.156	-.007
	Threshold Partial	-.089*	.025	.026	-.170	-.008
Partial Credit	Correct Only	.108*	.030	.022	.011	.206
	Subtractive	.044	.016	.153	-.009	.097
	Threshold	.027	.011	.218	-.008	.062
	Threshold Partial	.020	.009	.396	-.009	.048
Subtractive	Correct Only	.064*	.018	.020	.008	.121
	Partial Credit	-.044	.016	.153	-.097	.009
	Threshold	-.017	.012	1.00	-.057	.023
	Threshold Partial	-.025	.013	.712	-.066	.017
Threshold	Correct Only	.081*	.023	.026	.007	.156
	Partial Credit	-.027	.011	.218	-.062	.008
	Subtractive	.017	.012	1.00	-.023	.057
	Threshold Partial	-.008	.003	.297	-.018	.003
Threshold Partial	Correct Only	.089*	.025	.026	.008	.170
	Partial Credit	-.020	.009	.396	-.048	.009
	Subtractive	.025	.013	.712	-.017	.066
	Threshold	.008	.003	.297	-.003	.018

Figure 5 shows the change in mean item-total correlations across the five different scoring methods for each of the test forms. All forms, except Comprehensive Agriculture Form B, showed a consistent pattern of item-total correlations across the five different scoring methodologies. Comprehensive Agriculture Form B had a larger difference in item-total

correlations between partial-credit scoring and the other scoring methodologies. Regardless of the test or form, scoring an item correct only will provide the lowest item-total correlations. Conversely, partial-credit scoring consistently provides the highest item-total correlations. The other scoring methodologies are also consistent, with threshold-partial scoring having larger item-total correlations than threshold and subtractive scoring.

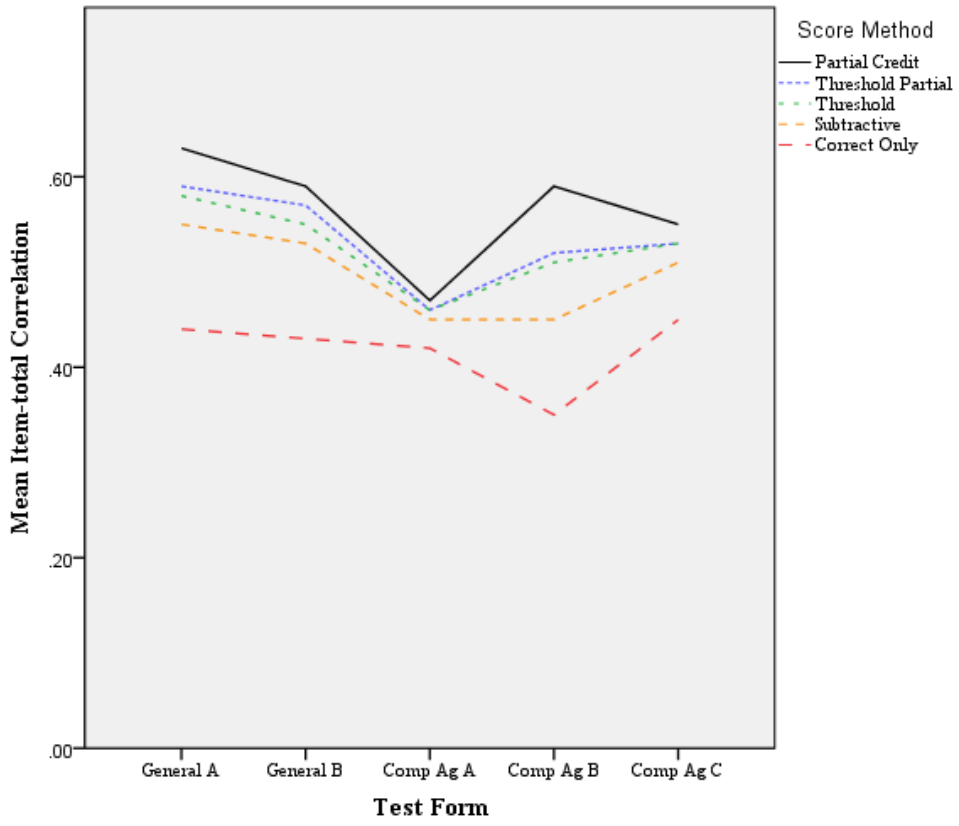


Figure 5. Mean item-total correlation, all items.

Table 31 shows the mean item-total correlations for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments using only TE items to calculate the total score. The correct-only scoring methodology had the lowest mean item-total correlation for all forms on both assessments (General Form A: M = -0.55, SD = 0.16; General Form B: M =

0.54, *SD* = 0.17; Comprehensive Agriculture Form A: *M* = .60, *SD* = 0.15; Comprehensive Agriculture Form B: *M* = 0.44, *SD* = 0.19; Comprehensive Agriculture Form C: *M* = .51, *SD* = 0.19). Conversely, the highest mean item-total correlation utilized the partial-credit methodology for all forms on both assessments (General Form A: *M* = 0.78, *SD* = 0.09; General Form B: *M* = 0.76, *SD* = 0.10; Comprehensive Agriculture Form A: *M* = .63, *SD* = 0.14; Comprehensive Agriculture Form B: *M* = 0.74, *SD* = 0.09; Comprehensive Agriculture Form C: *M* = 0.64, *SD* = 0.11).

Table 31. *Mean Item-Total Correlations, Tech Only*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	.55	.16	15	.54	.17	20	.60	.15	16	.44	.19	18	.51	.19
Partial Credit	15	.78	.09	15	.76	.10	20	.63	.14	16	.74	.09	18	.64	.11
Subtractive	15	.64	.14	15	.65	.14	20	.60	.13	16	.55	.18	18	.57	.15
Threshold	15	.70	.09	15	.70	.10	20	.61	.13	16	.62	.12	18	.60	.13
Threshold Partial	15	.72	.10	15	.71	.11	20	.60	.13	16	.63	.13	18	.60	.12

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the item-total correlations (calculated with only TE items) of the different scoring methodologies for all forms on both assessments. The Mauchly's Test of Sphericity was significant for all within-subjects ANOVAs conducted. Four of the five omnibus tests (Table 32) utilizing a Greenhouse-Geisser correction indicated a significant difference ($p < .01$) between their mean item-total correlations when utilizing different scoring strategies. Comprehensive Agriculture Form B was not significant ($p = .234$). This form was not further analyzed for main effects. The four significant omnibus tests had medium effect sizes with partial eta squared as high as .736.

Table 32. *Item-Total Correlation, Tech Only Repeated Measures ANOVA*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	.455	1.973	.231	29.819	.000	.681
	Error	.214	27.626	.008			
General Form B	Scoring	.410	1.649	.249	30.068	.000	.682
	Error	.191	23.087	.008			
Comp. Agriculture Form A	Scoring	.012	1.416	.009	1.530	.234	.075
	Error	.151	26.913	.006			
Comp. Agriculture Form B	Scoring	.747	1.876	.398	41.776	.000	.736
	Error	.268	28.143	.010			
Comp. Agriculture Form C	Scoring	.149	1.261	.118	12.340	.001	.421
	Error	.205	21.442	.010			

Results from the test of main effects for General CTE Form A can be found in Table 33. The largest item-total correlation mean difference (.227) was found between correct-only and partial-credit scoring ($p < .01$). Additionally, correct-only and threshold-partial scoring also had a significant difference ($p < .01$). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.172. Subtractive scoring was only significantly different from partial scoring ($p < .01$) and threshold-partial scoring ($p = .001$).

Table 33. *General CTE Form A Item-Total Correlation, Tech Only Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.227*	.033	.000	-.336	-.118
	Subtractive	-.089	.033	.162	-.197	.019
	Threshold	-.154*	.026	.000	-.240	-.068
	Threshold Partial	-.172*	.029	.000	-.267	-.078
Partial Credit	Correct Only	.227*	.033	.000	.118	.336
	Subtractive	.138*	.021	.000	.069	.207
	Threshold	.073*	.015	.002	.024	.121
	Threshold Partial	.054*	.013	.010	.011	.098
Subtractive	Correct Only	.089	.033	.162	-.019	.197
	Partial Credit	-.138*	.021	.000	-.207	-.069
	Threshold	-.065	.020	.055	-.132	.001
	Threshold Partial	-.084*	.015	.001	-.133	-.035
Threshold	Correct Only	.154*	.026	.000	.068	.240
	Partial Credit	-.073*	.015	.002	-.121	-.024
	Subtractive	.065	.020	.055	-.001	.132
	Threshold Partial	-.018	.006	.116	-.039	.003
Threshold Partial	Correct Only	.172*	.029	.000	.078	.267
	Partial Credit	-.054*	.013	.010	-.098	-.011
	Subtractive	.084*	.015	.001	.035	.133
	Threshold	.018	.006	.116	-.003	.039

Results from the test of main effects for General CTE Form B can be found in Table 34. The largest item-total correlation mean difference (.219) was found between correct-only and partial-credit scoring ($p < .01$). Additionally, correct-only and threshold-partial scoring also had a significant difference ($p = .001$). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.167. Subtractive scoring was significantly different from correct-only ($p = .012$), partial-credit ($p < .01$), and threshold-partial scoring ($p = .015$).

Table 34. *General CTE Form B Item-Total Correlation, Tech Only Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.219*	.033	.000	-.330	-.109
	Subtractive	-.109*	.027	.012	-.199	-.019
	Threshold	-.156*	.028	.001	-.251	-.062
	Threshold Partial	-.167*	.030	.001	-.265	-.068
Partial Credit	Correct Only	.219*	.033	.000	.109	.330
	Subtractive	.110*	.015	.000	.060	.161
	Threshold	.063*	.013	.002	.021	.105
	Threshold Partial	.053*	.011	.003	.016	.090
Subtractive	Correct Only	.109*	.027	.012	.019	.199
	Partial Credit	-.110*	.015	.000	-.161	-.060
	Threshold	-.047	.016	.120	-.102	.007
	Threshold Partial	-.058*	.015	.015	-.106	-.009
Threshold	Correct Only	.156*	.028	.001	.062	.251
	Partial Credit	-.063*	.013	.002	-.105	-.021
	Subtractive	.047	.016	.120	-.007	.102
	Threshold Partial	-.010	.004	.216	-.023	.003
Threshold Partial	Correct Only	.167*	.030	.001	.068	.265
	Partial Credit	-.053*	.011	.003	-.090	-.016
	Subtractive	.058*	.015	.015	.009	.106
	Threshold	.010	.004	.216	-.003	.023

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 35. The largest item-total correlation mean difference (.292) was found between correct-only and partial-credit scoring ($p < .01$). Additionally, correct-only and threshold-partial scoring also had a significant difference ($p < .01$). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.188.

Table 35. *Comprehensive Agriculture Form B Item-Total Correlation, Tech Only Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.292*	.038	.000	-.419	-.166
	Subtractive	-.106*	.023	.003	-.182	-.031
	Threshold	-.175*	.025	.000	-.258	-.091
	Threshold Partial	-.188*	.027	.000	-.275	-.100
Partial Credit	Correct Only	.292*	.038	.000	.166	.419
	Subtractive	.186*	.029	.000	.092	.280
	Threshold	.118*	.020	.000	.050	.185
	Threshold Partial	.105*	.022	.002	.034	.176
Subtractive	Correct Only	.106*	.023	.003	.031	.182
	Partial Credit	-.186*	.029	.000	-.280	-.092
	Threshold	-.068*	.017	.010	-.123	-.013
	Threshold Partial	-.081*	.015	.001	-.130	-.032
Threshold	Correct Only	.175*	.025	.000	.091	.258
	Partial Credit	-.118*	.020	.000	-.185	-.050
	Subtractive	.068*	.017	.010	.013	.123
	Threshold Partial	-.013	.005	.141	-.028	.002
Threshold Partial	Correct Only	.188*	.027	.000	.100	.275
	Partial Credit	-.105*	.022	.002	-.176	-.034
	Subtractive	.081*	.015	.001	.032	.130
	Threshold	.013	.005	.141	-.002	.028

Results from the test of main effects for Comprehensive Agriculture Form C can be found in Table 36. The largest item-total correlation mean difference (.122) was found between correct-only and partial-credit scoring ($p = 0.015$). Additionally, correct-only scoring and threshold-partial also had a significant difference ($p = .028$). The mean difference in item-total correlations between correct-only and threshold-partial scoring was 0.089.

Table 36. *Comprehensive Agriculture Form C Item-Total Correlation, Tech Only Test of Main Effects*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.122*	.032	.015	-.227	-.018
	Subtractive	-.060*	.017	.029	-.115	-.004
	Threshold	-.082*	.022	.018	-.154	-.010
	Threshold Partial	-.089*	.025	.028	-.171	-.007
Partial Credit	Correct Only	.122*	.032	.015	.018	.227
	Subtractive	.063*	.018	.032	.004	.121
	Threshold	.040*	.012	.034	.002	.079
	Threshold Partial	.034*	.010	.036	.002	.066
Subtractive	Correct Only	.060*	.017	.029	.004	.115
	Partial Credit	-.063*	.018	.032	-.121	-.004
	Threshold	-.022	.011	.557	-.057	.013
	Threshold Partial	-.029	.012	.320	-.069	.011
Threshold	Correct Only	.082*	.022	.018	.010	.154
	Partial Credit	-.040*	.012	.034	-.079	-.002
	Subtractive	.022	.011	.557	-.013	.057
	Threshold Partial	-.007	.004	1.00	-.021	.007
Threshold Partial	Correct Only	.089*	.025	.028	.007	.171
	Partial Credit	-.034*	.010	.036	-.066	-.002
	Subtractive	.029	.012	.320	-.011	.069
	Threshold	.007	.004	1.00	-.007	.021

Figure 6 shows the change in mean item-total correlations across the five different scoring methodologies for each of the test forms. All forms, except Comprehensive Agriculture Form A, show a consistent pattern of item-total correlations across the five different scoring methodologies. Comprehensive Agriculture Form A had a smaller difference in item-total correlations between partial-credit scoring and the other scoring methodologies. Regardless of the test or form, scoring an item correct only will provide the lowest item-total correlations. Conversely, partial-credit scoring consistently provides the highest item-total correlations. The

other scoring methodologies are also consistent, with threshold-partial scoring having larger item-total correlations than threshold and subtractive scoring.

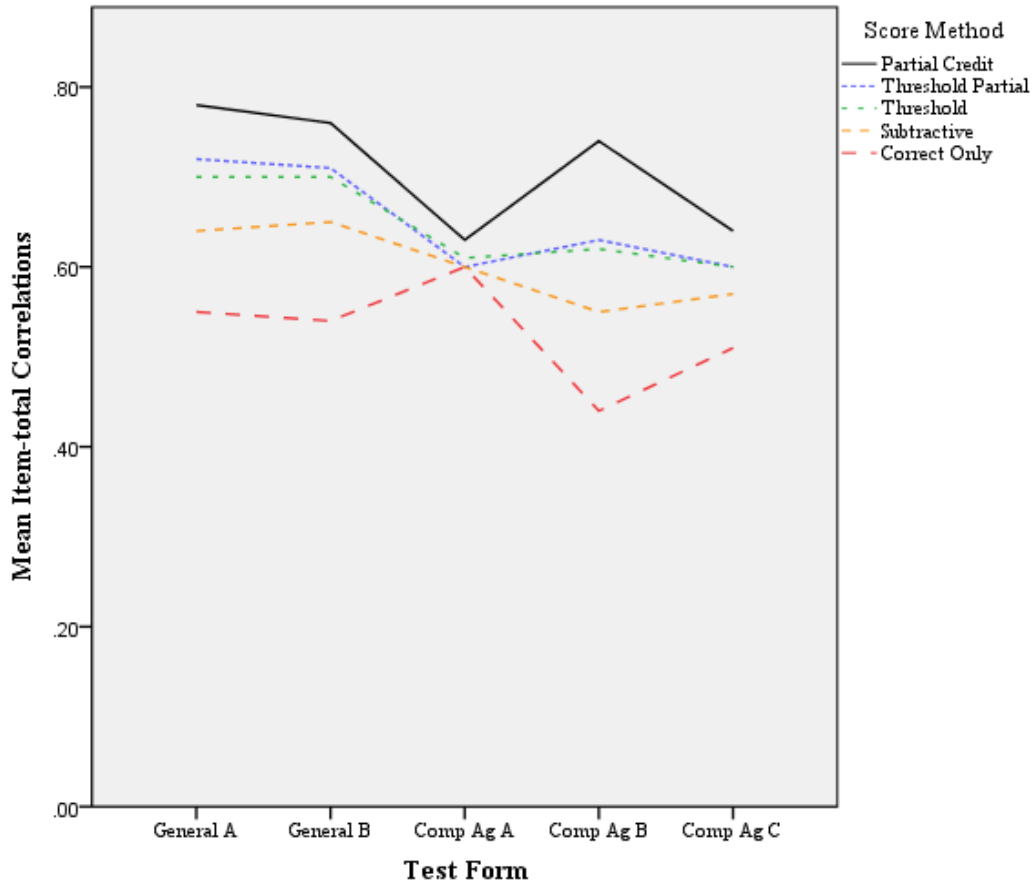


Figure 6. Mean item-total correlations, tech only by form and score methodology.

***a* parameters.** Mixed IRT models were estimated across all forms and all assessments. A combination of a 2-PL and GRM were utilized to estimate the *a* parameter for each item. When calibrating with all items, partial-credit scoring had the highest *a* parameters (General Form A: $M = 1.75$, $SD = 0.308$; General Form B: $M = 1.63$, $SD = 0.368$; Comprehensive Agriculture Form A: $M = 1.61$, $SD = 0.444$; Comprehensive Agriculture Form B: $M = 1.55$, $SD = 0.361$; Comprehensive Agriculture Form C: $M = 1.19$, $SD = 0.411$). The lowest *a* parameters were from

correct-only scoring (General Form A: *M* = 1.34, *SD* = 0.340; General Form B: *M* = 1.39, *SD* = 0.339; Comprehensive Agriculture Form A: *M* = 1.04, *SD* = 0.373; Comprehensive Agriculture Form B: *M* = 1.03, *SD* = 0.409; Comprehensive Agriculture Form C: *M* = 1.17, *SD* = 0.540), except for on General Form B. Table 37 shows the mean *a* parameters for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments when all items within each form were utilized for calibration.

Table 37. *Mean* a *Parameters, All Items*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	1.34	.340	15	1.39	.339	19	1.04	.373	16	1.03	.409	18	1.17	.540
Partial Credit	15	1.75	.308	15	1.63	.368	19	1.61	.444	16	1.55	.361	18	1.19	.411
Subtractive	15	1.51	.323	15	1.42	.341	19	1.24	.462	16	1.09	.406	18	1.20	.529
Threshold	15	1.55	.289	15	1.38	.443	19	1.35	.394	16	1.23	.378	18	1.18	.471
Threshold Partial	15	1.55	.282	15	1.49	.328	19	1.39	.426	16	1.23	.397	18	1.19	.465

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the *a* parameters of the different scoring methodologies for all forms on both assessments. The Mauchly's Test of Sphericity was significant for all within-subjects ANOVAs conducted. Four out of the five omnibus tests with a Greenhouse-Geisser correction indicated a significant difference ($p < .05$) between their mean *a* parameters when utilizing different scoring strategies (Table 38). Comprehensive Agriculture Form C was not significant ($p = .932$), therefore no additional analysis was conducted on that form.

Table 38. *a Parameter Repeated Measures ANOVA, All Items*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	1.261	1.876	.672	23.358	.000	.625
	Error	.756	26.265	.029			
General Form B	Scoring	.612	2.221	.276	3.843	.028	.215
	Error	2.230	31.092	.072			
Comp. Agriculture Form A	Scoring	3.399	2.036	1.669	22.396	.000	.554
	Error	2.732	36.653	.075			
Comp. Agriculture Form B	Scoring	2.634	1.687	1.561	41.251	.000	.733
	Error	.958	25.310	.038			
Comp. Agriculture Form C	Scoring	.009	2.401	.004	.101	.932	.006
	Error	1.491	40.814	.037			

Each form (excluding Comprehensive Form C) showed a significant difference between the different scoring strategies' *a* parameters when all items were utilized for calibration. To further clarify which scoring strategies were responsible for the overall difference, a test of the main effects for each form was conducted. To reduce the chance of a type I error occurring due to multiple comparisons, a Bonferroni adjustment was utilized.

Results from the test of main effects for General Form A can be found in Table 39. The largest *a* parameter mean difference (0.408) can be found between partial-credit and correct-only scoring ($p < .01$). Additionally, correct-only scoring and subtractive scoring also had a significant difference ($p = .012$). The mean difference in *a* parameters between correct-only scoring and subtractive was 0.167. Subtractive scoring was not statistically different than threshold ($p = 1.00$) or threshold-partial scoring ($p = 1.00$).

Table 39. *General CTE Form A* a *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.408	.057	.000	-.598	-.218
	Subtractive	-.167	.041	.012	-.305	-.030
	Threshold	-.202	.054	.022	-.382	-.022
	Threshold Partial	-.203	.060	.043	-.402	-.005
Partial Credit	Correct Only	.408	.057	.000	.218	.598
	Subtractive	.240	.046	.001	.089	.392
	Threshold	.206	.024	.000	.127	.284
	Threshold Partial	.204	.027	.000	.115	.293
Subtractive	Correct Only	.167	.041	.012	.030	.305
	Partial Credit	-.240	.046	.001	-.392	-.089
	Threshold	-.034	.038	1.000	-.161	.092
	Threshold Partial	-.036	.040	1.000	-.168	.097
Threshold	Correct Only	.202	.054	.022	.022	.382
	Partial Credit	-.206	.024	.000	-.284	-.127
	Subtractive	.034	.038	1.000	-.092	.161
	Threshold Partial	-.002	.012	1.000	-.041	.038
Threshold Partial	Correct Only	.203	.060	.043	.005	.402
	Partial Credit	-.204	.027	.000	-.293	-.115
	Subtractive	.036	.040	1.000	-.097	.168
	Threshold	.002	.012	1.000	-.038	.041

Results from the test of main effects for General Form B can be found in Table 40. The largest *a* parameter mean difference (0.242) can be found between partial-credit and threshold scoring (p = .029). Additionally, partial-credit and subtractive scoring also had a significant difference (p = .004). The mean difference in *a* parameters between partial-credit and subtractive

scoring was 0.209. Other than partial-credit scoring, threshold scoring was not statistically different than any other scoring method.

Table 40. *General CTE Form B* a *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.233	.077	.092	-.489	.024
	Subtractive	-.024	.078	1.000	-.283	.235
	Threshold	.009	.107	1.000	-.346	.364
	Threshold Partial	-.102	.069	1.000	-.331	.127
Partial Credit	Correct Only	.233	.077	.092	-.024	.489
	Subtractive	.209	.046	.004	.057	.361
	Threshold	.242	.067	.029	.018	.465
	Threshold Partial	.131	.031	.008	.028	.234
Subtractive	Correct Only	.024	.078	1.000	-.235	.283
	Partial Credit	-.209	.046	.004	-.361	-.057
	Threshold	.033	.095	1.000	-.282	.348
	Threshold Partial	-.078	.035	.403	-.194	.037
Threshold	Correct Only	-.009	.107	1.000	-.364	.346
	Partial Credit	-.242	.067	.029	-.465	-.018
	Subtractive	-.033	.095	1.000	-.348	.282
	Threshold Partial	-.111	.085	1.000	-.393	.172
Threshold Partial	Correct Only	.102	.069	1.000	-.127	.331
	Partial Credit	-.131	.031	.008	-.234	-.028
	Subtractive	.078	.035	.403	-.037	.194
	Threshold	.111	.085	1.000	-.172	.393

Results from the test of main effects for Comprehensive Agriculture Form A can be found in Table 41. The largest *a* parameter mean difference (0.578) can be found between partial-credit and correct-only scoring ($p < .01$). Additionally, partial-credit and subtractive

scoring also had a significant difference ($p < .01$). The mean difference in *a* parameters between partial-credit and subtractive scoring was 0.376. Threshold scoring was not statistically different than threshold-partial scoring ($p = 1.00$), but was statistically different than correct only ($p = .009$) and partial credit ($p < .01$).

Table 41. *Comprehensive Agriculture Form A* a *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.578	.093	.000	-.875	-.280
	Subtractive	-.202	.069	.090	-.422	.019
	Threshold	-.318	.080	.009	-.574	-.061
	Threshold Partial	-.350	.079	.003	-.601	-.099
Partial Credit	Correct Only	.578	.093	.000	.280	.875
	Subtractive	.376	.065	.000	.169	.583
	Threshold	.260	.042	.000	.126	.395
	Threshold Partial	.228	.038	.000	.107	.349
Subtractive	Correct Only	.202	.069	.090	-.019	.422
	Partial Credit	-.376	.065	.000	-.583	-.169
	Threshold	-.116	.060	.704	-.308	.077
	Threshold Partial	-.148	.049	.077	-.305	.010
Threshold	Correct Only	.318	.080	.009	.061	.574
	Partial Credit	-.260	.042	.000	-.395	-.126
	Subtractive	.116	.060	.704	-.077	.308
	Threshold Partial	-.032	.021	1.000	-.100	.036
Threshold Partial	Correct Only	.350	.079	.003	.099	.601
	Partial Credit	-.228	.038	.000	-.349	-.107
	Subtractive	.148	.049	.077	-.010	.305
	Threshold	.032	.021	1.000	-.036	.100

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 42. The largest *a* parameter mean difference (0.523) can be found between partial credit and correct only ($p<.01$). Additionally, partial-credit and subtractive scoring also had a significant difference ($p<.01$). The mean difference in *a* parameters between partial-credit and subtractive scoring was 0.466. Threshold scoring was not statistically different than threshold-partial scoring ($p = 1.00$), but was statistically different than subtractive ($p < .01$) and partial credit ($p < .01$).

Table 42. *Comprehensive Agriculture Form B* a *Parameter Test of Main Effects, All Items*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Correct Only	Partial Credit	-.523	.047	.000	-.677	-.369
	Subtractive	-.057	.058	1.000	-.247	.133
	Threshold	-.206	.064	.056	-.415	.004
	Threshold Partial	-.201	.066	.083	-.419	.016
Partial Credit	Correct Only	.523	.047	.000	.369	.677
	Subtractive	.466	.042	.000	.327	.605
	Threshold	.317	.037	.000	.196	.438
	Threshold Partial	.321	.039	.000	.192	.451
Subtractive	Correct Only	.057	.058	1.000	-.133	.247
	Partial Credit	-.466	.042	.000	-.605	-.327
	Threshold	-.149	.023	.000	-.225	-.072
	Threshold Partial	-.144	.023	.000	-.221	-.068
Threshold	Correct Only	.206	.064	.056	-.004	.415
	Partial Credit	-.317	.037	.000	-.438	-.196
	Subtractive	.149	.023	.000	.072	.225
	Threshold Partial	.004	.013	1.000	-.038	.047
Threshold Partial	Correct Only	.201	.066	.083	-.016	.419
	Partial Credit	-.321	.039	.000	-.451	-.192
	Subtractive	.144	.023	.000	.068	.221
	Threshold	-.004	.013	1.000	-.047	.038

Figure 7 shows the change in mean *a* parameters across the five different scoring methods for each of the test forms. All forms, except Comprehensive Agriculture Form C (omnibus test was not significant) and General Form B, shows a consistent pattern of mean *a* parameters across the remaining four different scoring methodologies. Threshold scoring has lower mean a parameters than subtractive scoring for General Form B. This is different than the other forms. Regardless of the test of form, scoring an item correct-only will provide the lowest *a*

parameters. Conversely, partial-credit scoring consistently provides the highest *a* parameters. The other scoring methodologies are also consistent, with threshold-partial scoring having larger *a* parameters than threshold and subtractive scoring.

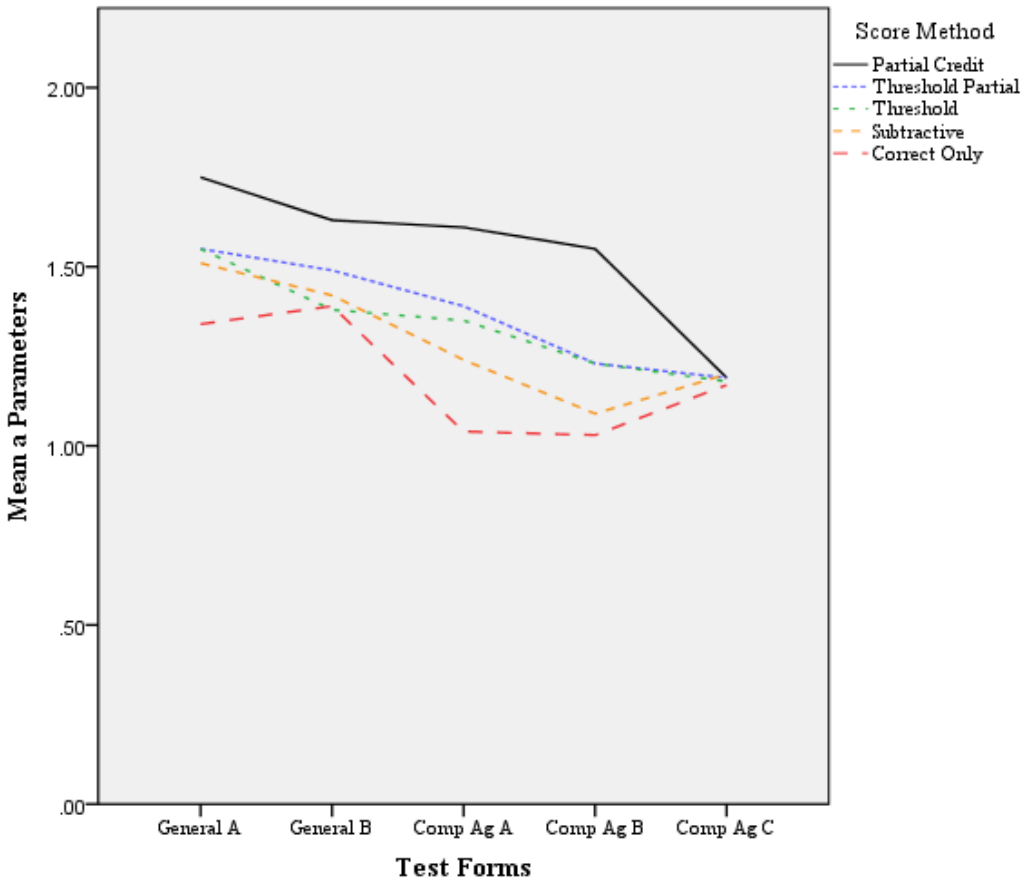


Figure 7. Mean *a* parameters by form and scoring methodology, all items.

Table 43 shows the mean *a* parameters for each scoring methodology on both the General CTE and Comprehensive Agriculture assessments using only TE items for calibration. The correct-only scoring methodology had the lowest mean *a* parameters for all forms on both assessments (General Form A: M=1.69, SD=0.60; General Form B: M= 1.42, SD=0.50; Comprehensive Agriculture Form A: *M*= 1.03, *SD*= 0.43; Comprehensive Agriculture Form B:

M= 1.07, *SD*= 0.59; Comprehensive Agriculture Form C: *M*= 1.47, *SD*= 0.99). Conversely, the highest mean *a* parameter utilized testlet response theory methodology for all forms on both assessments, except for on the General Form A (General Form A: M=2.68, SD=1.77; General Form B: M= 4.19, SD=1.19; Comprehensive Agriculture Form A: *M*= 2.59, *SD*= 1.49; Comprehensive Agriculture Form B: *M*= 2.84, *SD*= 1.38; Comprehensive Agriculture Form C: *M*= 2.35, *SD*= 1.27).

Table 43. *Mean* a *Parameters, Tech Only*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Correct Only	15	1.69	0.60	14	1.42	0.50	19	1.03	0.43	16	1.07	0.59	18	1.47	0.99
Partial Credit	15	2.73	0.69	14	1.70	0.63	19	2.00	0.56	16	2.23	0.48	18	1.48	0.55
Subtractive	15	2.25	0.65	14	1.64	0.63	19	1.40	0.52	16	1.48	0.54	18	1.51	0.79
Threshold	15	2.27	0.61	14	1.64	0.58	19	1.60	0.49	16	1.73	0.52	18	1.43	0.61
Threshold Partial	15	2.27	0.61	14	1.66	0.60	19	1.61	0.51	16	1.72	0.55	18	1.44	0.60
Testlet Response	15	2.68	1.77	14	4.19	1.19	19	2.59	1.49	16	2.84	1.38	18	2.35	1.27

Multiple within-subjects Analysis of Variance (ANOVA) were calculated on the *a* parameters of the different scoring methodologies for all forms on both assessments utilizing only TE items. The Mauchly's Test of Sphericity was significant for all within subjects ANOVA's conducted. All six omnibus tests with a Greenhouse-Geisser correction indicate a significant difference ($p<.05$) between their mean *a* parameters when utilizing different scoring strategies (Table 44).

Table 44. *a Parameter Repeated Measures ANOVA, Tech Only*

Test Form	Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
General Form A	Scoring	10.572	1.185	8.923	4.276	.049	.234
	Error	34.611	16.586	2.087			
General Form B	Scoring	77.862	1.053	73.970	26.289	.000	.669
	Error	38.503	13.684	2.814			
Comp. Agriculture Form A	Scoring	27.295	1.155	23.633	10.746	.003	.374
	Error	45.718	20.789	2.199			
Comp. Agriculture Form B	Scoring	30.232	1.300	23.256	19.663	.000	.567
	Error	23.063	19.499	1.183			
Comp. Agriculture Form C	Scoring	11.922	1.759	6.778	10.126	.001	.373
	Error	20.015	29.902	.669			

Each form showed a significant difference between the different scoring strategies' *a* parameters when only TE items were utilized for calibration. To further clarify which scoring strategies were responsible for the overall difference, a test of the main effects for each form was conducted. To reduce the chance of a type I error occurring due to multiple comparisons, a Bonferroni adjustment was utilized.

Results from the test of main effects for General Form A can be found in Table 45. The largest *a* parameter mean difference (1.042) was found between partial-credit and correct-only scoring ($p<.01$). Additionally, correct-only and threshold scoring also had a significant difference ($p=.001$). The mean difference in *a* parameters between correct-only and subtractive scoring was 0.577. Testlet response theory was not statistically different than any of the other methods ($p=1.00$).

Table 45. *General CTE Form A a Parameter Test of Main Effects, Tech Only*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-1.042	.132	.000	-1.507	-.578
	Subtractive	-.558	.092	.000	-.884	-.233
	Threshold	-.577	.105	.001	-.946	-.208
	Threshold Partial	-.571	.114	.003	-.972	-.171
	Testlet Response	-.833	.423	1.000	-2.325	.659
Partial Credit	Correct Only	1.042	.132	.000	.578	1.507
	Subtractive	.484	.077	.000	.214	.754
	Threshold	.465	.065	.000	.235	.696
	Threshold Partial	.471	.067	.000	.235	.707
	Testlet Response	.209	.435	1.000	-1.327	1.746
Subtractive	Correct Only	.558	.092	.000	.233	.884
	Partial Credit	-.484	.077	.000	-.754	-.214
	Threshold	-.019	.051	1.000	-.198	.161
	Threshold Partial	-.013	.055	1.000	-.206	.180
	Testlet Response	-.275	.410	1.000	-1.721	1.171
Threshold	Correct Only	.577	.105	.001	.208	.946
	Partial Credit	-.465	.065	.000	-.696	-.235
	Subtractive	.019	.051	1.000	-.161	.198
	Threshold Partial	.006	.016	1.000	-.050	.061
	Testlet Response	-.256	.417	1.000	-1.728	1.215
Threshold Partial	Correct Only	.571	.114	.003	.171	.972
	Partial Credit	-.471	.067	.000	-.707	-.235
	Subtractive	.013	.055	1.000	-.180	.206
	Threshold	-.006	.016	1.000	-.061	.050
	Testlet Response	-.262	.416	1.000	-1.729	1.205
Testlet Response	Correct Only	.833	.423	1.000	-.659	2.325
	Partial Credit	-.209	.435	1.000	-1.746	1.327
	Subtractive	.275	.410	1.000	-1.171	1.721
	Threshold	.256	.417	1.000	-1.215	1.728
	Threshold Partial	.262	.416	1.000	-1.205	1.729

Results from the test of main effects for General Form B can be found in Table 46. The largest *a* parameter mean difference (2.763) was found between testlet response theory and correct-only scoring (p=.001). Additionally, correct-only and partial-credit scoring also had a significant difference (p=.039). The mean difference in *a* parameters between correct-only and partial-credit scoring was 0.279. Subtractive, threshold, and threshold-partial scoring were not significantly different from each other. Finally, testlet response theory was significantly different from all other scoring methodologies.

Table 46. *General CTE Form B* a *Parameter Test of Main Effects, Tech Only*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.279	.075	.039	-.547	-.010
	Subtractive	-.215	.074	.183	-.480	.050
	Threshold	-.218	.060	.048	-.434	-.001
	Threshold Partial	-.242	.065	.037	-.473	-.010
	Testlet Response	-2.763	.496	.001	-4.540	-.985
Partial Credit	Correct Only	.279	.075	.039	.010	.547
	Subtractive	.063	.053	1.000	-.126	.253
	Threshold	.061	.033	1.000	-.056	.178
	Threshold Partial	.037	.034	1.000	-.083	.157
	Testlet Response	-2.484	.505	.004	-4.293	-.676
Subtractive	Correct Only	.215	.074	.183	-.050	.480
	Partial Credit	-.063	.053	1.000	-.253	.126
	Threshold	-.003	.049	1.000	-.179	.173
	Threshold Partial	-.026	.049	1.000	-.201	.149
	Testlet Response	-2.548	.483	.002	-4.279	-.817
Threshold	Correct Only	.218	.060	.048	.001	.434
	Partial Credit	-.061	.033	1.000	-.178	.056
	Subtractive	.003	.049	1.000	-.173	.179
	Threshold Partial	-.024	.013	1.000	-.072	.024
	Testlet Response	-2.545	.500	.003	-4.337	-.753
Threshold Partial	Correct Only	.242	.065	.037	.010	.473
	Partial Credit	-.037	.034	1.000	-.157	.083
	Subtractive	.026	.049	1.000	-.149	.201
	Threshold	.024	.013	1.000	-.024	.072
	Testlet Response	-2.521	.506	.004	-4.336	-.707
Testlet Response	Correct Only	2.763	.496	.001	.985	4.540
	Partial Credit	2.484	.505	.004	.676	4.293
	Subtractive	2.548	.483	.002	.817	4.279
	Threshold	2.545	.500	.003	.753	4.337
	Threshold Partial	2.521	.506	.004	.707	4.336

Results from the test of main effects for Comprehensive Agriculture Form A can be found in Table 47. The largest *a* parameter mean difference (1.560) was found between testlet response theory and correct-only scoring (p=.005). Additionally, correct-only and partial-credit scoring also had a significant difference (p<.01). The mean difference in *a* parameters between correct-only and partial-credit scoring was 0.969. Subtractive, threshold, and threshold-partial scoring were not significantly different from each other. Finally, testlet response theory was only significantly different from correct-only scoring.

Table 47. *Comprehensive Agriculture Form A* a *Parameter Test of Main Effects, Tech Only*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.969	.117	.000	-1.364	-.574
	Subtractive	-.373	.072	.001	-.616	-.130
	Threshold	-.575	.087	.000	-.869	-.282
	Threshold Partial	-.578	.079	.000	-.845	-.311
	Testlet Response	-1.560	.351	.005	-2.746	-.375
Partial Credit	Correct Only	.969	.117	.000	.574	1.364
	Subtractive	.596	.094	.000	.277	.915
	Threshold	.394	.061	.000	.187	.600
	Threshold Partial	.391	.059	.000	.192	.590
	Testlet Response	-.591	.416	1.000	-1.999	.816
Subtractive	Correct Only	.373	.072	.001	.130	.616
	Partial Credit	-.596	.094	.000	-.915	-.277
	Threshold	-.202	.078	.279	-.466	.062
	Threshold Partial	-.204	.064	.076	-.421	.012
	Testlet Response	-1.187	.377	.083	-2.462	.088
Threshold	Correct Only	.575	.087	.000	.282	.869
	Partial Credit	-.394	.061	.000	-.600	-.187
	Subtractive	.202	.078	.279	-.062	.466
	Threshold Partial	-.003	.028	1.000	-.096	.091
	Testlet Response	-.985	.390	.315	-2.302	.332
Threshold Partial	Correct Only	.578	.079	.000	.311	.845
	Partial Credit	-.391	.059	.000	-.590	-.192
	Subtractive	.204	.064	.076	-.012	.421
	Threshold	.003	.028	1.000	-.091	.096
	Testlet Response	-.983	.390	.322	-2.302	.336
Testlet Response	Correct Only	1.560	.351	.005	.375	2.746
	Partial Credit	.591	.416	1.000	-.816	1.999
	Subtractive	1.187	.377	.083	-.088	2.462
	Threshold	.985	.390	.315	-.332	2.302
	Threshold Partial	.983	.390	.322	-.336	2.302

Results from the test of main effects for Comprehensive Agriculture Form B can be found in Table 48. The largest *a* parameter mean difference (1.765) was found between testlet response theory and correct-only scoring ($p<.01$). Additionally, correct-only and partial-credit scoring also had a significant difference ($p<.01$). The mean difference in *a* parameters between correct-only and partial-credit scoring was 1.156. Subtractive scoring was significantly different from all other scoring methods. Threshold and threshold-partial scoring were not significantly different ($p=1.00$). Finally, testlet response theory was not significantly different from partial-credit scoring ($p=1.00$) or threshold scoring ($p=.055$).

Table 48. *Comprehensive Agriculture Form B* a *Parameter Test of Main Effects, Tech Only*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-1.156	.162	.000	-1.721	-.592
	Subtractive	-.411	.110	.029	-.793	-.030
	Threshold	-.660	.112	.000	-1.051	-.269
	Threshold Partial	-.643	.119	.001	-1.056	-.229
	Testlet Response	-1.765	.262	.000	-2.678	-.852
Partial Credit	Correct Only	1.156	.162	.000	.592	1.721
	Subtractive	.745	.081	.000	.463	1.027
	Threshold	.496	.080	.000	.219	.773
	Threshold Partial	.514	.082	.000	.228	.800
	Testlet Response	-.609	.348	1.000	-1.821	.603
Subtractive	Correct Only	.411	.110	.029	.030	.793
	Partial Credit	-.745	.081	.000	-1.027	-.463
	Threshold	-.249	.042	.000	-.396	-.102
	Threshold Partial	-.231	.027	.000	-.325	-.137
	Testlet Response	-1.354	.307	.008	-2.423	-.285
Threshold	Correct Only	.660	.112	.000	.269	1.051
	Partial Credit	-.496	.080	.000	-.773	-.219
	Subtractive	.249	.042	.000	.102	.396
	Threshold Partial	.018	.029	1.000	-.085	.120
	Testlet Response	-1.105	.321	.055	-2.225	.015
Threshold Partial	Correct Only	.643	.119	.001	.229	1.056
	Partial Credit	-.514	.082	.000	-.800	-.228
	Subtractive	.231	.027	.000	.137	.325
	Threshold	-.018	.029	1.000	-.120	.085
	Testlet Response	-1.123	.319	.046	-2.232	-.013
Testlet Response	Correct Only	1.765	.262	.000	.852	2.678
	Partial Credit	.609	.348	1.000	-.603	1.821
	Subtractive	1.354	.307	.008	.285	2.423
	Threshold	1.105	.321	.055	-.015	2.225
	Threshold Partial	1.123	.319	.046	.013	2.232

Results from the test of main effects for Comprehensive Agriculture Form C can be found in Table 49 The largest *a* parameter mean difference (0.886) was found between testlet response theory and correct-only scoring (p=.025). Additionally, testlet response theory and threshold scoring also had a significant difference (p=.018). The mean difference in *a* parameters between correct-only and partial-credit scoring was 0.925. All other scoring methods were not significantly different from each other (p=1.00).

Table 49. *Comprehensive Agriculture Form C* a *Parameter Test of Main Effects, Tech Only*

(I) Scoring Method	(J) Scoring Method	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Correct Only	Partial Credit	-.015	.148	1.000	-.519	.488
	Subtractive	-.042	.101	1.000	-.385	.301
	Threshold	.039	.120	1.000	-.371	.449
	Threshold Partial	.030	.125	1.000	-.395	.456
	Testlet Response	-.886	.238	.025	-1.698	-.074
Partial Credit	Correct Only	.015	.148	1.000	-.488	.519
	Subtractive	-.027	.108	1.000	-.394	.340
	Threshold	.054	.049	1.000	-.111	.220
	Threshold Partial	.046	.047	1.000	-.116	.207
	Testlet Response	-.871	.256	.050	-1.743	.001
Subtractive	Correct Only	.042	.101	1.000	-.301	.385
	Partial Credit	.027	.108	1.000	-.340	.394
	Threshold	.081	.089	1.000	-.223	.385
	Threshold Partial	.072	.084	1.000	-.215	.359
	Testlet Response	-.844	.251	.055	-1.700	.012
Threshold	Correct Only	-.039	.120	1.000	-.449	.371
	Partial Credit	-.054	.049	1.000	-.220	.111
	Subtractive	-.081	.089	1.000	-.385	.223
	Threshold Partial	-.009	.013	1.000	-.054	.036
	Testlet Response	-.925	.238	.018	-1.738	-.113
Threshold Partial	Correct Only	-.030	.125	1.000	-.456	.395
	Partial Credit	-.046	.047	1.000	-.207	.116
	Subtractive	-.072	.084	1.000	-.359	.215
	Threshold	.009	.013	1.000	-.036	.054
	Testlet Response	-.916	.239	.020	-1.732	-.101
Testlet Response	Correct Only	.886	.238	.025	.074	1.698
	Partial Credit	.871	.256	.050	-.001	1.743
	Subtractive	.844	.251	.055	-.012	1.700
	Threshold	.925	.238	.018	.113	1.738
	Threshold Partial	.916	.239	.020	.101	1.732

Figure 8 shows the change in mean *a* parameters across the six different scoring methods for each of the test forms utilizing only TE items in calibration. Comprehensive Agriculture Form C did not show significant differences between scoring methodologies, except when testlet response theory was utilized. All other forms showed a consistent pattern with partial-credit and testlet response scoring providing the highest mean *a* parameters, and correct-only scoring providing the lowest mean *a* parameters. The other scoring methodologies were also consistent, with threshold-partial scoring having larger *a* parameters than threshold and subtractive scoring.

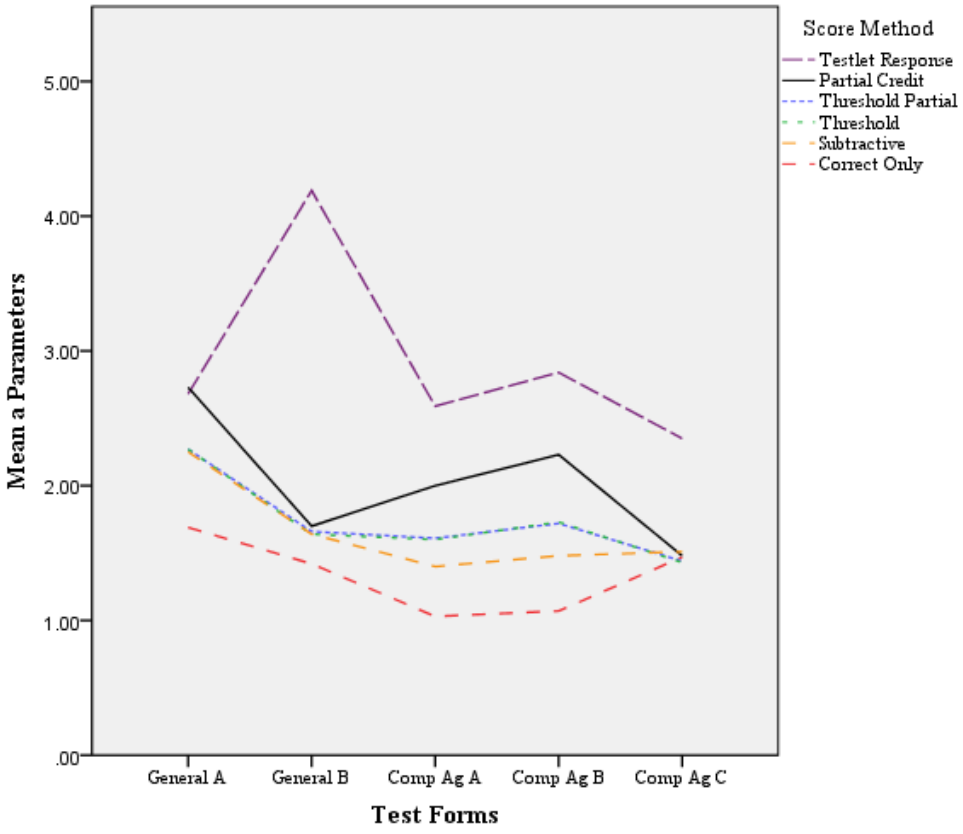


Figure 8. Mean *a* parameters by form and scoring methodology, tech only.

Research Question 2: How does the scoring method of TE items affect reliability of scores from each test form?

Coefficient Alpha

Coefficient alpha was calculated for each form either utilizing all items or just the TE items. The number of test takers and the reliability coefficient for all five forms across the two assessments utilizing all items are shown in Table 50 The two scoring methods that consistently had the highest reliability are partial credit (General Form A: α=.968; General Form B: α =.970; Comprehensive Agriculture Form A: α =.932; Comprehensive Agriculture Form B: α =.943; Comprehensive Agriculture Form C: α =.941) and threshold partial (General Form A: α=.968; General Form B: α =.969; Comprehensive Agriculture Form A: α =.933; Comprehensive Agriculture Form B: α =.941; Comprehensive Agriculture Form C: α =.942).

Table 50. *Coefficient Alpha, All Items*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	α	*SEM*	*N*	α	*SEM*	*N*	α	*SEM*	*N*	α	*SEM*	*N*	α	*SEM*
Correct Only	406	.964	4.02	453	.967	3.85	126	.929	4.25	126	.931	4.13	134	.937	4.24
Partial Credit	406	.968	3.88	453	.970	3.77	126	.932	4.07	126	.943	4.04	134	.941	4.13
Subtractive	406	.966	3.98	453	.968	3.85	126	.932	4.18	126	.937	4.10	134	.940	4.23
Threshold	406	.967	3.92	453	.969	3.82	126	.933	4.11	126	.939	4.09	134	.941	4.19
Threshold Partial	406	.968	3.90	453	.969	3.85	126	.933	4.14	126	.941	4.10	134	.942	4.20

Figure 9 shows the pattern of coefficient alphas for each scoring method by test form for all items. Though the changes are small, there is a pattern prevalent between the test forms. This pattern indicates a consistent effect of scoring methodology on reliability estimates.

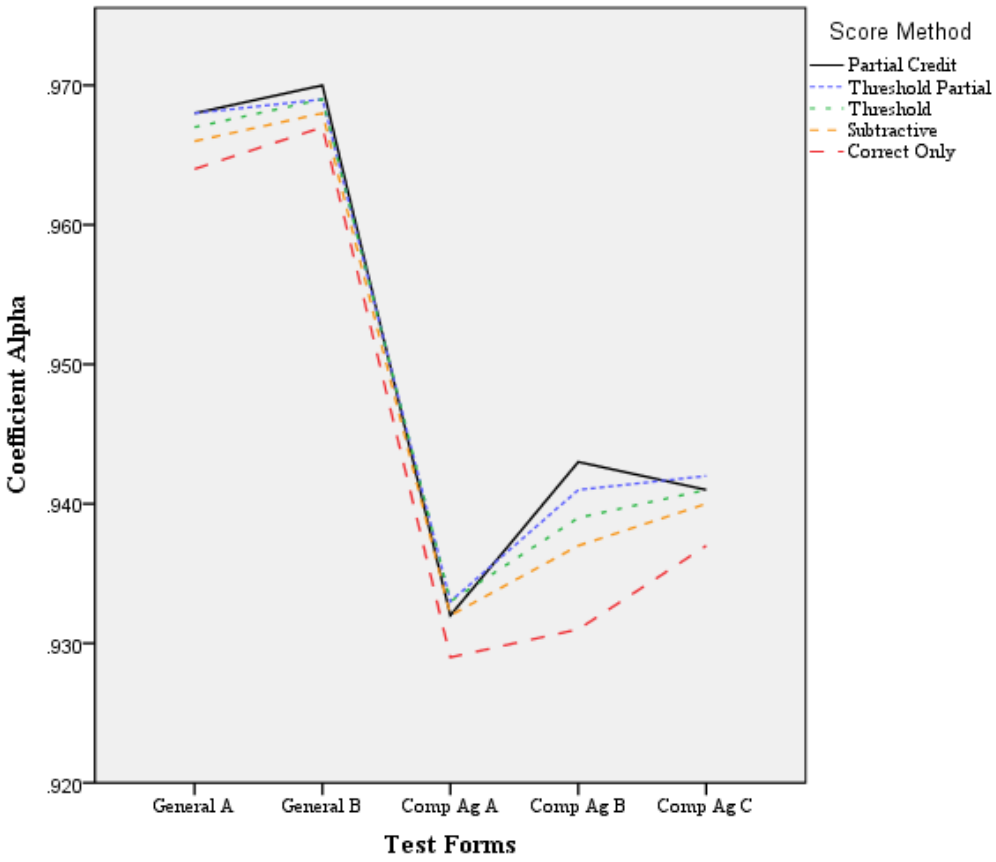


Figure 9. Coefficient alpha, all items.

Coefficient alpha was calculated for each form for each scoring method utilizing only TE items. The number of test takers and the reliability coefficient for all five forms across the two assessments utilizing only TE items are shown in Table 51 The scoring method that consistently had the highest reliability was partial credit (General Form A: α=.952; General Form B: α =.949; Comprehensive Agriculture Form A: α =.914; Comprehensive Agriculture Form B: α =.947; Comprehensive Agriculture Form C: α =.920). The scoring method that produced the lowest coefficient alphas was the correct-only scoring method (General Form A: α=.849; General Form

B: α =.844; Comprehensive Agriculture Form A: α =.898; Comprehensive Agriculture Form B: α =.760; Comprehensive Agriculture Form C: α =.847).

Table 51. *Coefficient Alpha, Tech Only*

	General CTE						Comprehensive Agriculture								
	Form A			Form B			Form A			Form B			Form C		
	N	α	*SEM*	*N*	α	*SEM*	*N*	α	*SEM*	*N*	α	*SEM*	*N*	α	*SEM*
Correct Only	406	.849	1.45	453	.844	1.41	126	.898	1.63	126	.760	1.43	134	.847	1.66
Partial Credit	406	.952	0.97	453	.949	1.02	126	.914	1.18	126	.947	1.07	134	.920	1.27
Subtractive	406	.908	1.26	453	.899	1.26	126	.903	1.50	126	.864	1.34	134	.885	1.57
Threshold	406	.928	1.15	453	.924	1.17	126	.909	1.39	126	.901	1.26	134	.902	1.46
Threshold Partial	406	.935	1.15	453	.929	1.19	126	.905	1.44	126	.910	1.33	134	.905	1.52

Figure 10 shows the pattern of coefficient alphas for each scoring method by test form for only TE items. Due to only having TE items, the pattern is more prevalent. Except for Comprehensive Agriculture Form A, there is a consistent pattern of scoring methodology effects on reliability estimates.

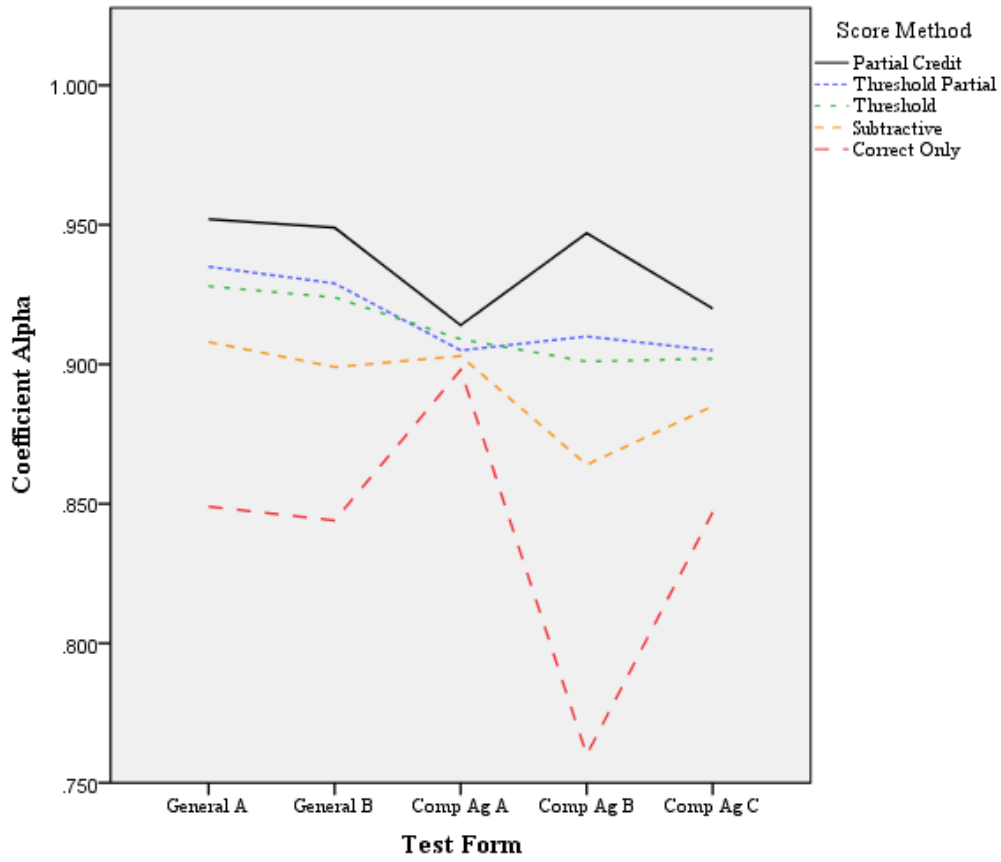


Figure 10. Coefficient alpha, tech only.

Test Information

Test information was calculated for each form on each assessment for all scoring methods. The higher the test information function, the more discriminating the test is at that level of θ_i. Additionally, the test information scale is determined in part by the number of items on each test, therefore there should be no comparison between forms and assessments.

Figure 11 compares the test information functions for all scoring methods on General Form A. It is clear that partial-credit scoring provides the most information of any scoring method. The lowest test information can be found with correct-only scoring. Threshold-partial and threshold scoring provide the next most test information, with subtractive scoring coming in

second to last. Additionally, all test information functions are positioned with peak information just below a θ of 0. Partial-credit scoring provides a little more information for lower level abilities than the other scoring methods.

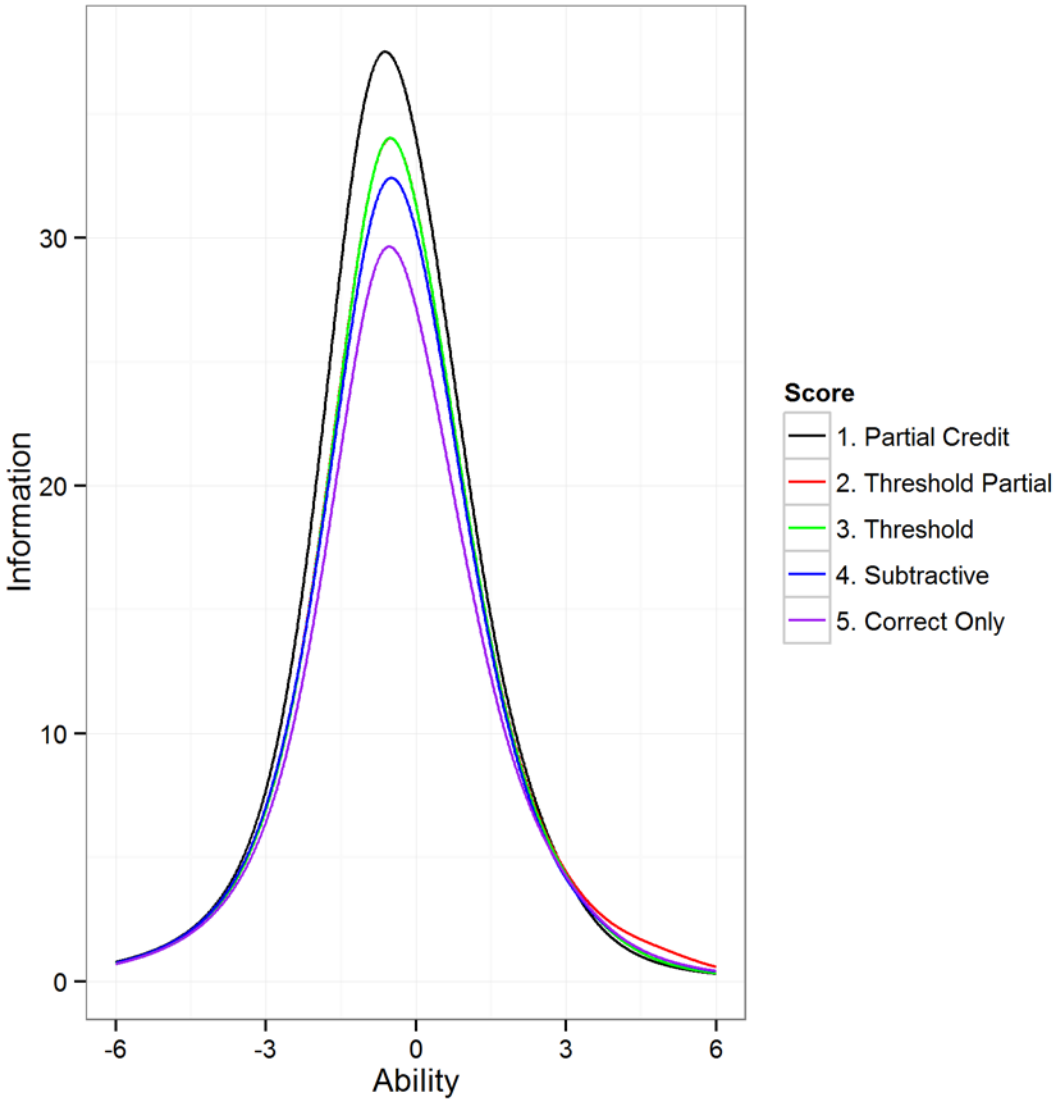


Figure 11. General Form A test information function comparison, all items.

Figure 12 compares the test information functions for all scoring methods on General Form B. Similar to General Form A, it is clear that partial-credit scoring provides the most

information of any scoring method. The lowest amount of test information can be found with correct-only scoring. Threshold-partial and threshold scoring provide the next highest amount of test information, with subtractive scoring coming in second to last. Additionally, all test information functions are positioned with peak information just below a θ of 0. Partial-credit scoring provides a little more information for lower level abilities than the other scoring methods.

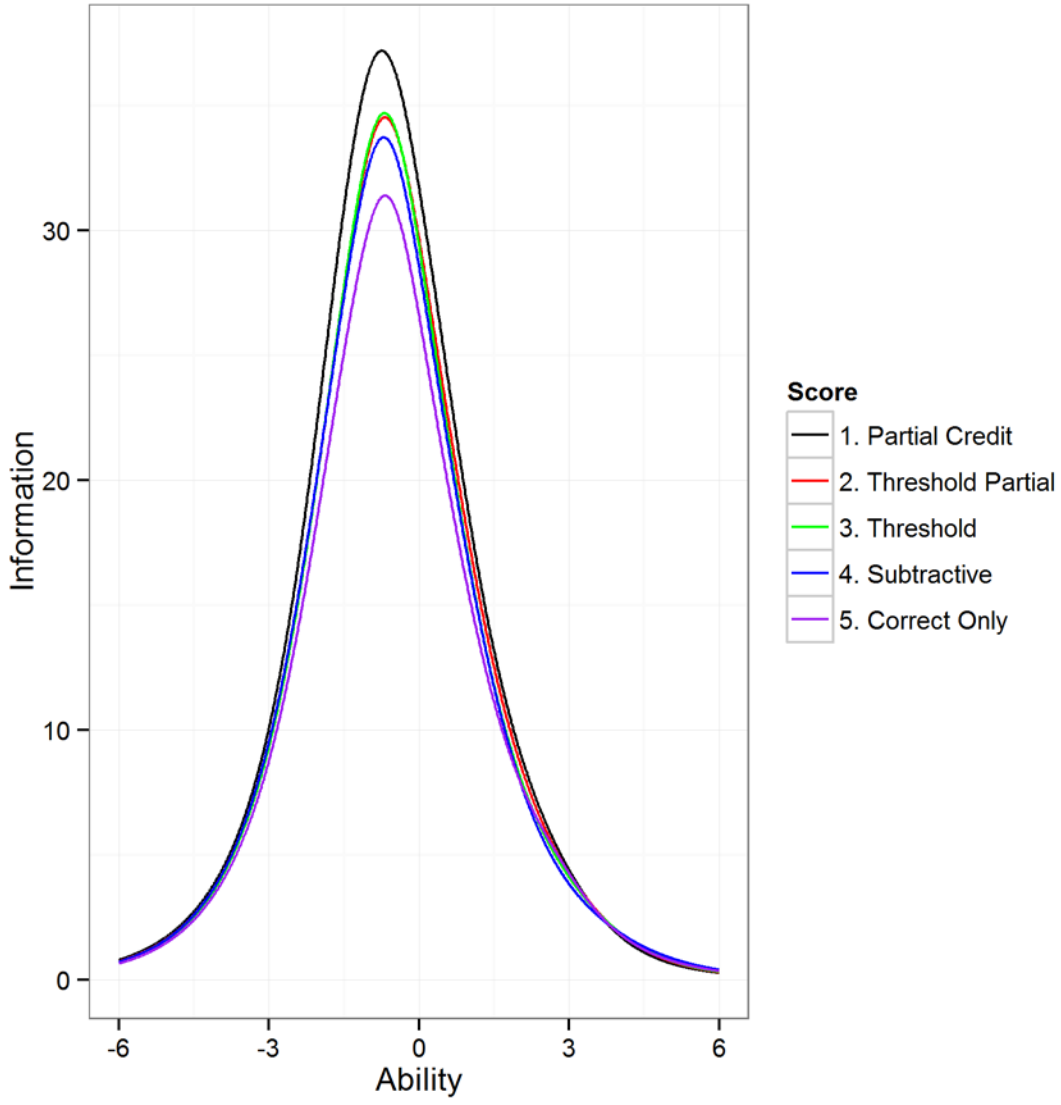


Figure 12. General Form B test information function comparison, all items.

Figure 13 compares the test information functions for all scoring methods on Comprehensive Agriculture Form A. Partial-credit scoring provides the most information of any scoring method. Moreover, partial-credit scoring provides more information at more levels of θ_i than the other scoring methods. The lowest amount of test information can be found with correct-only scoring. Correct-only scoring provides the lowest total test information, but also provides information at the lowest range of θ_i. Threshold-partial and threshold scoring provide the next most test information, with subtractive scoring coming in second to last. Additionally, all test information functions are positioned with peak information just below a θ of 0.

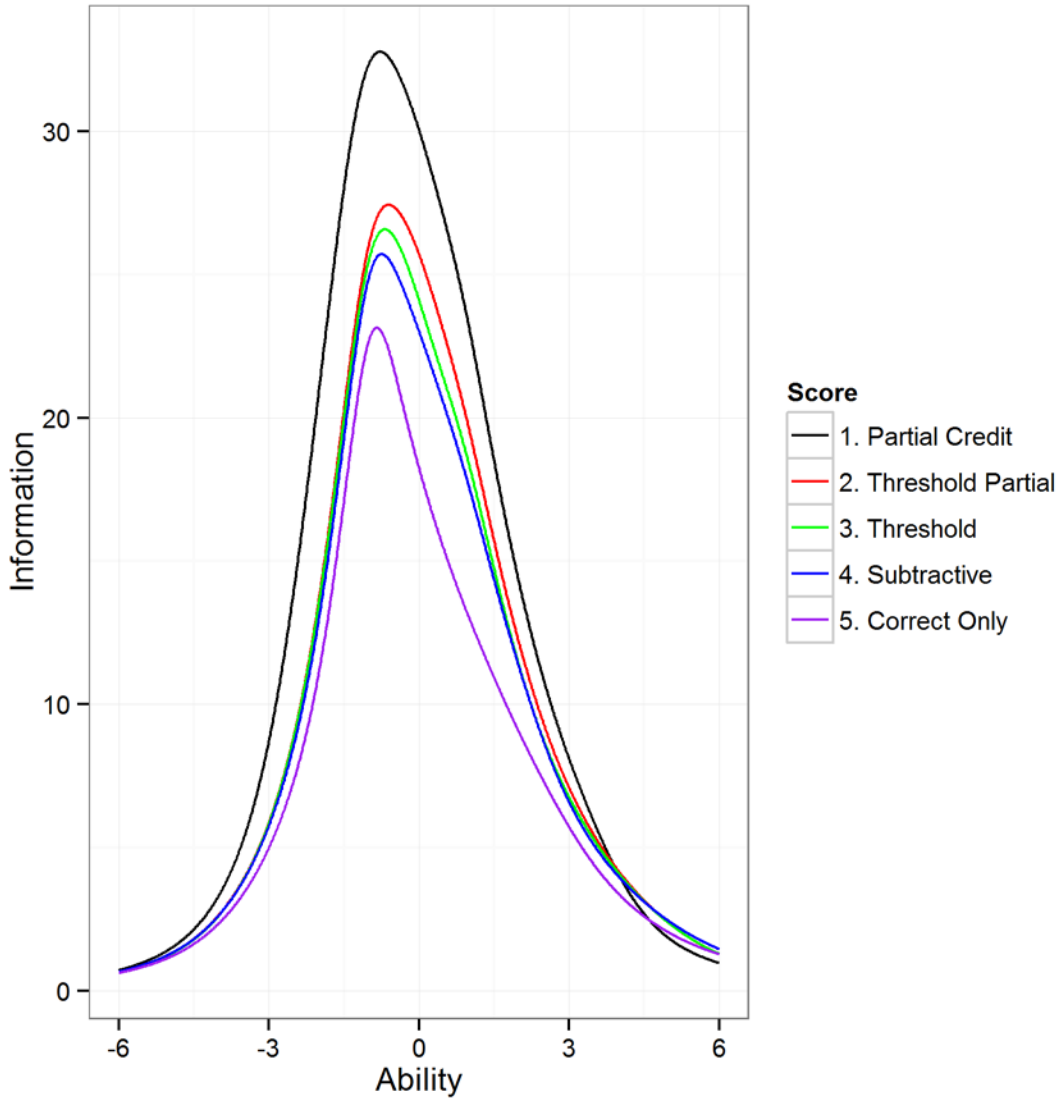


Figure 13. Comprehensive Agriculture Form A test information function comparison, all items.

Figure 14 compares the test information functions for all scoring methods on Comprehensive Agriculture Form B. Similar to Form A, partial-credit scoring provides the most information of any scoring method. Moreover, partial-credit scoring provides more information at more levels of θ_i than the other scoring methods. The lowest test information can be found with correct-only scoring. Correct-only scoring provides the lowest total test information, but

also provides less information at fewer levels of θ_i. Threshold-partial and threshold scoring provide the next most test information, with subtractive scoring coming in second to last. Additionally, all test information functions are positioned with peak information just below a θ of 0.

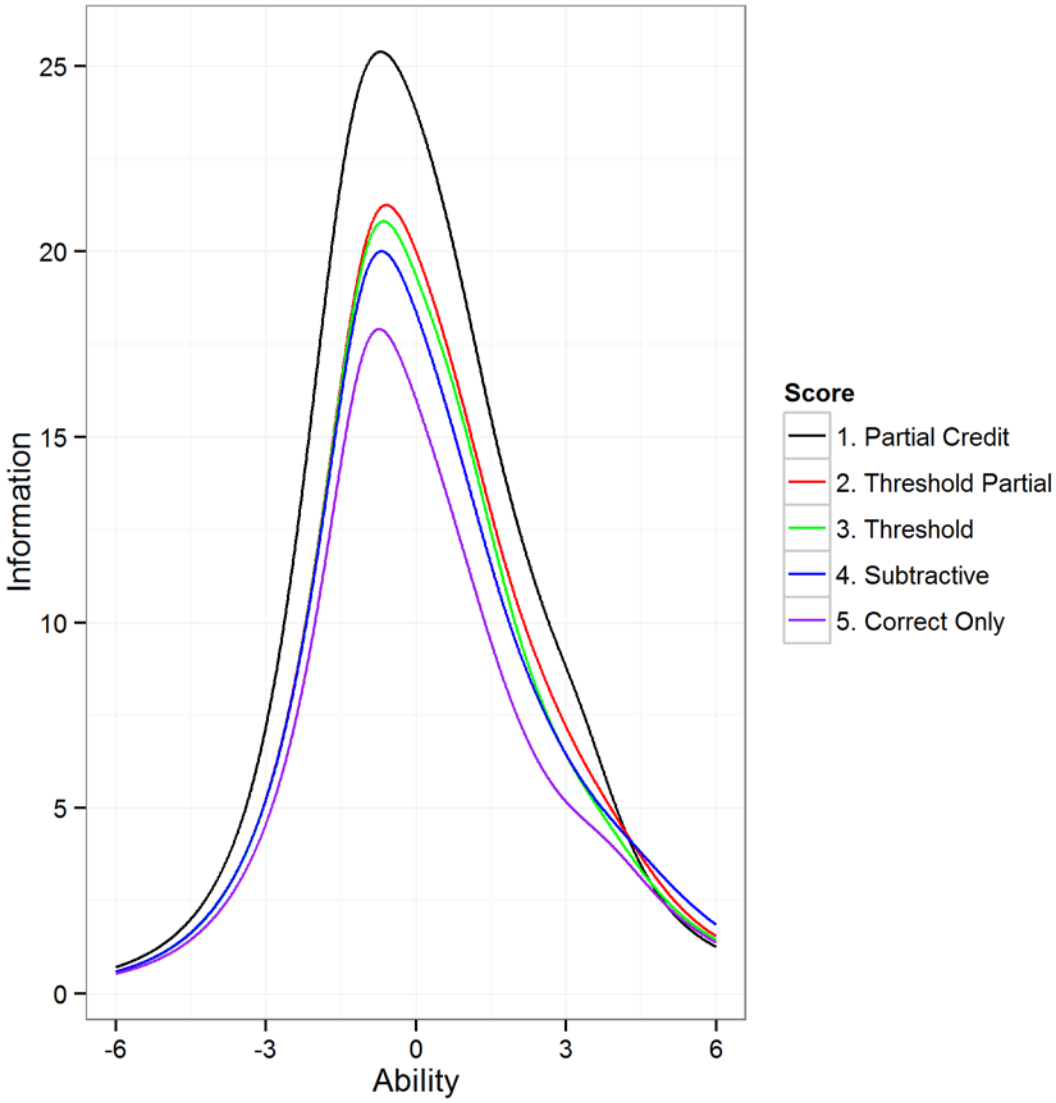


Figure 14. Comprehensive Agriculture Form B test information function comparison, all items.

Figure 15 compares the test information functions for all scoring methods on Comprehensive Agriculture Form C. Comprehensive Agriculture Form C breaks from the pattern of the previous forms. Partial-credit and threshold-partial scoring provide similar levels of test information. The test information function for partial-credit scoring is slightly to the left of the test information function for threshold-partial scoring. This indicates that for higher-ability students on this form, scoring the items as threshold-partial would provide more information than if they were scored as partial-credit. The lowest test information can be found with correct-only scoring. Correct-only scoring provides the lowest total test information, but also provides less information at fewer levels of θ_i. Additionally, all test information functions are positioned with peak information just below a θ_i of 0, except for correct-only scoring, which is positioned at 0.

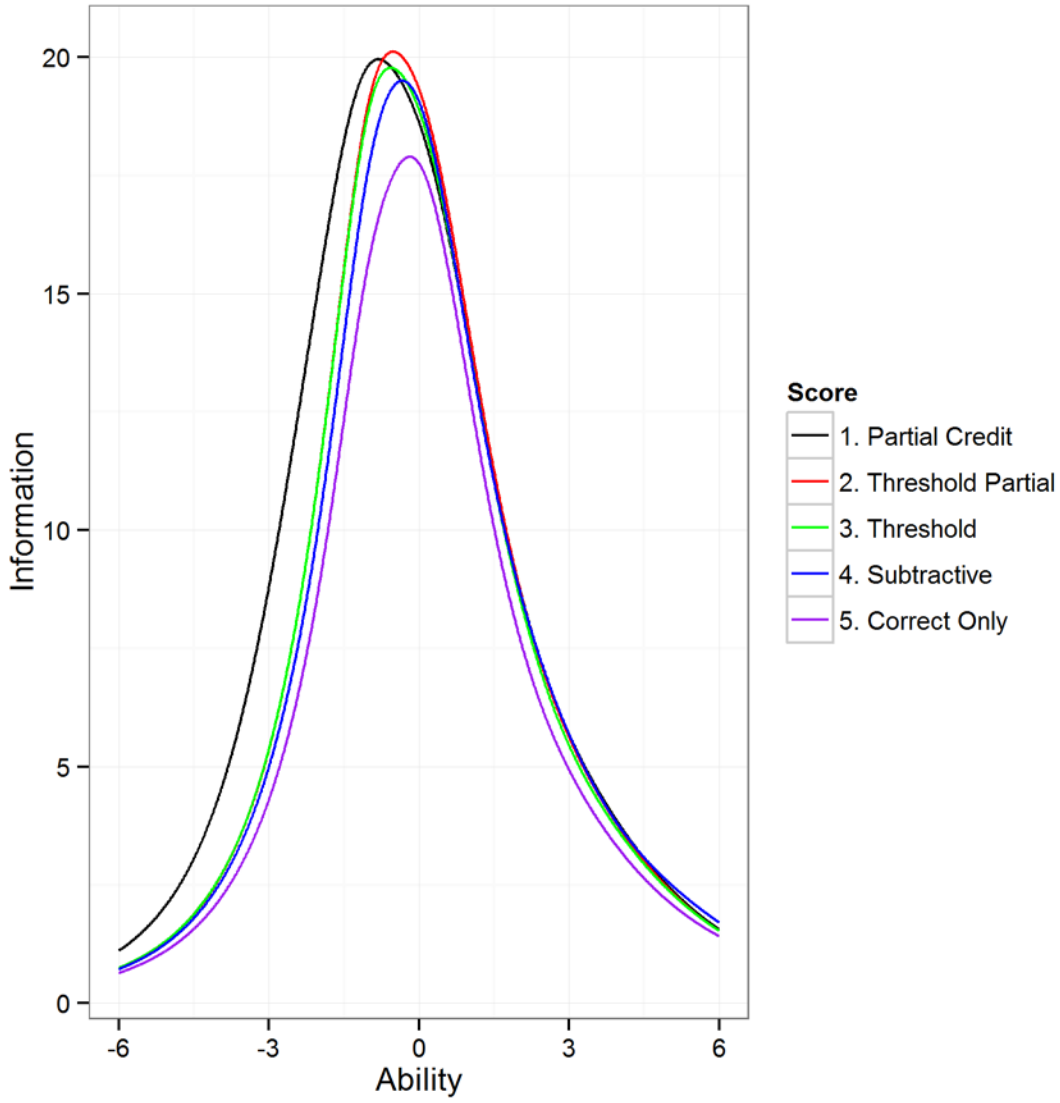


Figure 15. Comprehensive Agriculture Form C test information function comparison, all items.

Test information functions were also computed utilizing only TE items for calibration. Additionally, when only TE items were utilized, testlet response theory was also calibrated. For ease of interpretation, for each form two different figures showing the TIF are reported. Multiple graphs were created as TRT provides drastically more test information due to having more items. In order to enhance the comparison, multiple figures are utilized. The first figure shows just the

original five methods of scoring, and the second figure shows those same TIFs with testlet response added.

Figure 16 compares the test information functions for all scoring methods on General CTE Form A. Partial-credit scoring provides the most test information followed by threshold-partial scoring. The test information function for partial-credit scoring is wider than the function for threshold-partial scoring. This indicates that partial-credit scoring provides more information for a wider range of abilities. The lowest test information can be found with subtractive scoring. Additionally, all test information functions are positioned with peak information just at a θ_i of 0.

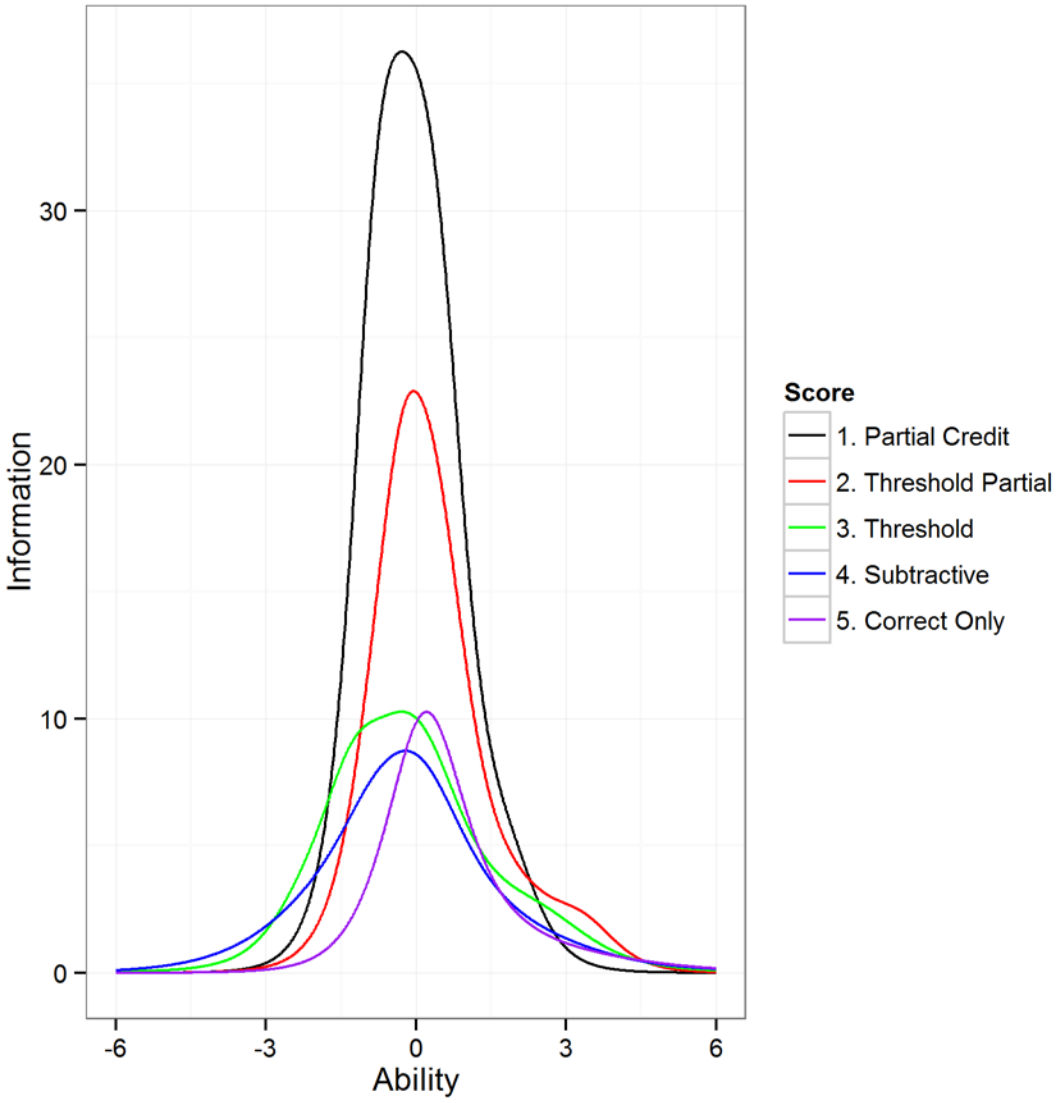


Figure 16. General CTE Form A test information function comparison, tech only (excluding TRT)

Figure 17 adds testlet response theory to Figure 16. Testlet response theory scoring provides much more information than all the other scoring methods. This makes sense, as each scoring point within testlet response is considered a separate item. Therefore, a 15-item TE assessment is actually scored as a 74-item dichotomously scored assessment. The addition of these items will account for more test information. In Figure 17, TRT scoring is multi-modal.

This can happen when an assessment has a large range of *a* parameters, covering different difficulties. Additionally, TRT provides more information on the lower end of the ability scale, than the other scoring methodologies provide.

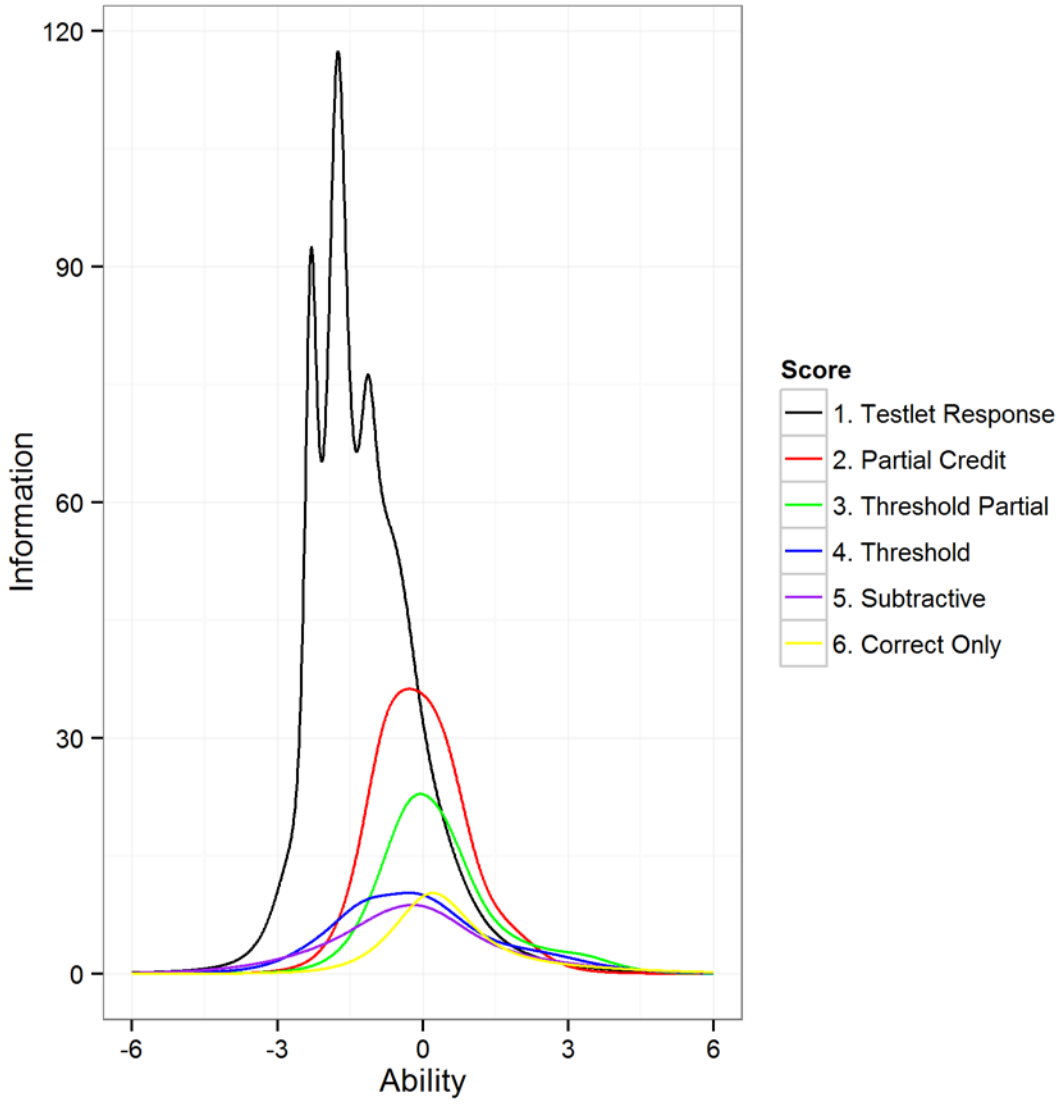


Figure 17. General CTE Form A test information function comparison, tech only.

Figure 18 compares the test information functions for all scoring methods (except testlet response theory) on General CTE Form B. Figure 18 shows that partial-credit scoring provides the most test information for General Form B. The test information function for partial-credit scoring is slightly to the left of the test information function for threshold-partial scoring. This indicates that for higher-ability students on this form, scoring the items as threshold-partial provides more information than scoring as partial-credit. The lowest test information can be found with subtractive and threshold scoring. Correct-only scoring provides more test information for higher-ability test takers. Additionally, all test information functions are positioned with peak information just below a θ_i of 0, except for correct-only scoring, which is positioned at 0.

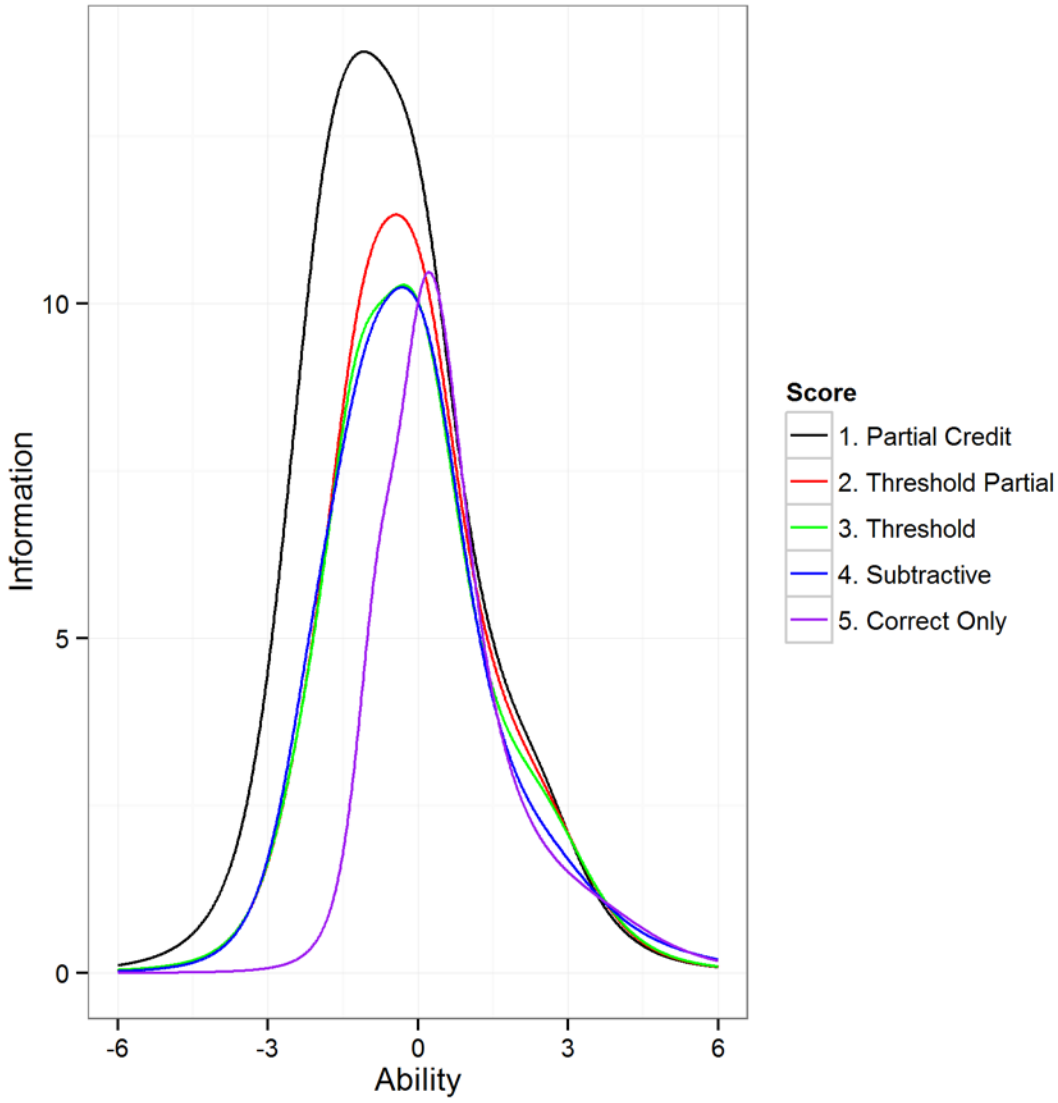


Figure 18. General CTE Form B test information function comparison, tech only (excluding TRT).

Figure 19 adds testlet response theory to Figure 18. Testlet response theory scoring provides much more information than all the other scoring methods. This occurs due to each scoring point within a testlet response model being considered a separate item. Therefore, a 15-item TE assessment is actually scored as a 72-item dichotomously scored assessment. The addition of these items accounts for more test information. In Figure 19, TRT scoring is multi-

modal. Although unlike General CTE Form A, the TRT test information function is centered around 0.

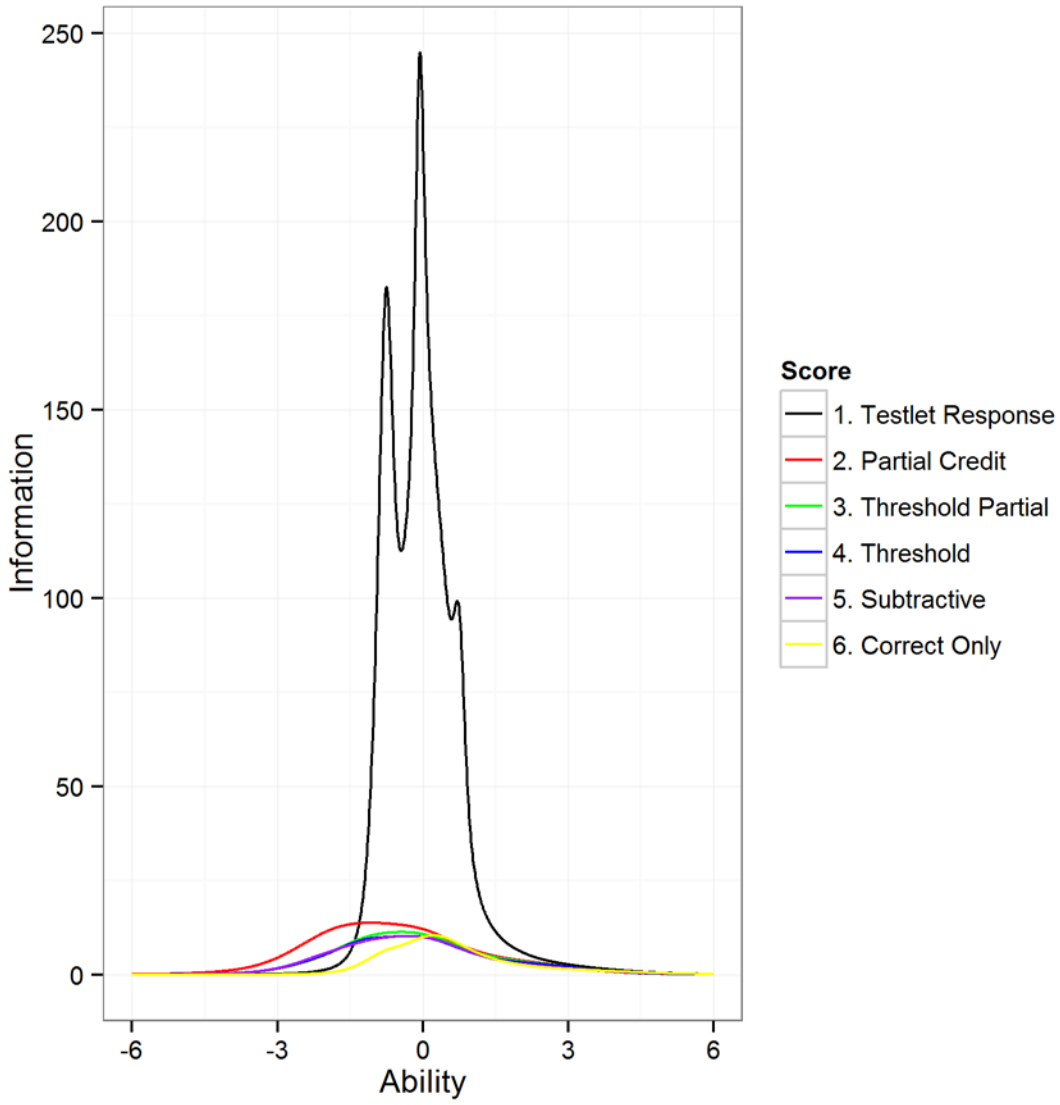


Figure 19. General CTE Form B test information function comparison, tech only.

Figure 20 compares the test information functions for all scoring methods (except testlet response theory) on Comprehensive Agriculture Form A. Partial-credit scoring provides the most test information for Comprehensive Agriculture Form A. The test information function for

partial-credit scoring is wider than all the other scoring methods. This indicates that scoring items as partial-credit provides more information at a wider range of abilities than utilizing the other scoring methods. The lowest test information can be found with correct-only scoring. Additionally, all test information functions, except partial credit, are positioned with peak information just above a θ_i of 0.

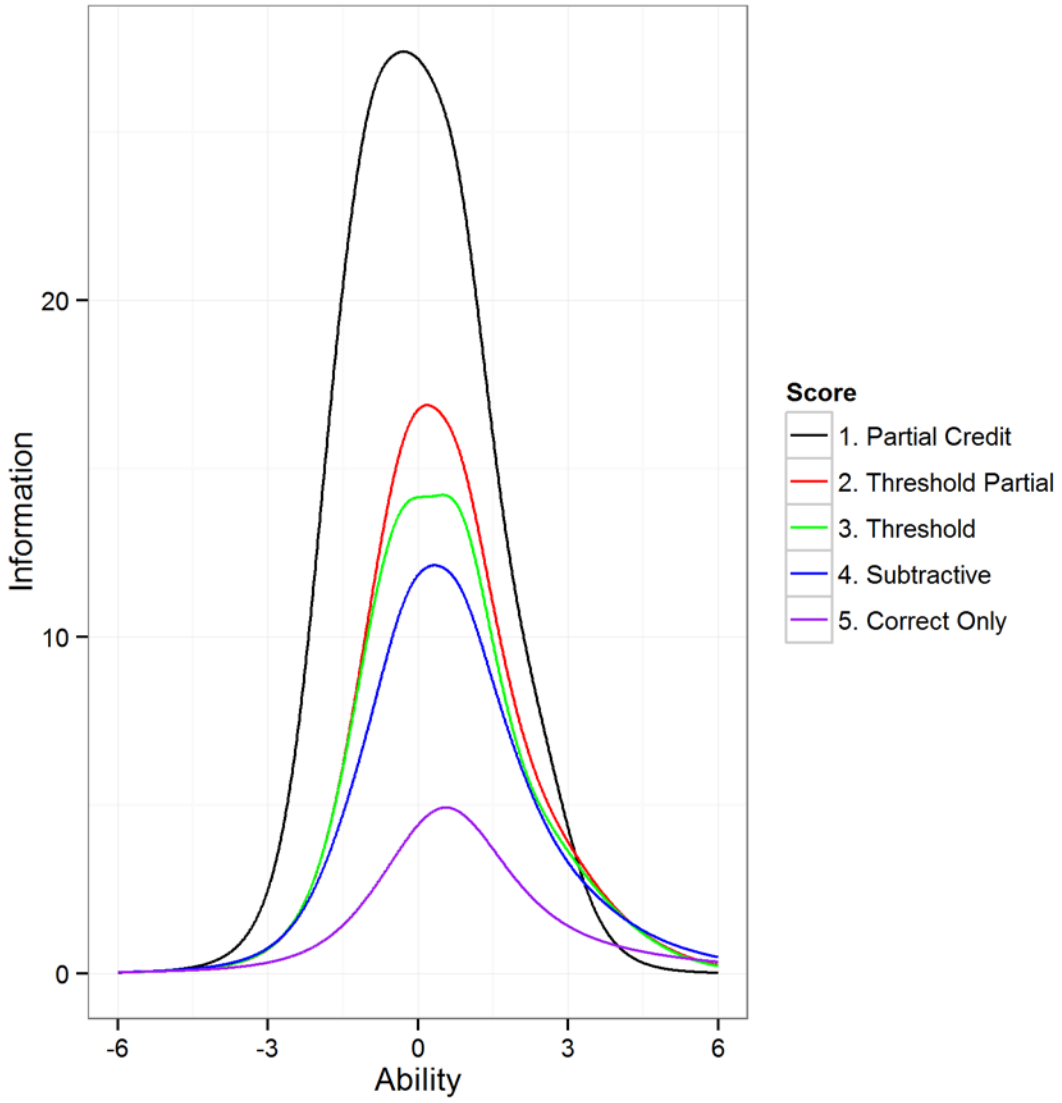


Figure 20. Comprehensive Agriculture Form A test information function comparison, tech only (excluding TRT).

Figure 21 compares all six TIFs for each of the scoring methods. Testlet response theory clearly provides the most test information similar to previous findings. Additionally, the testlet response theory TIF is multi-modal indicating that the test has a broad range of *a* and *b* parameters.

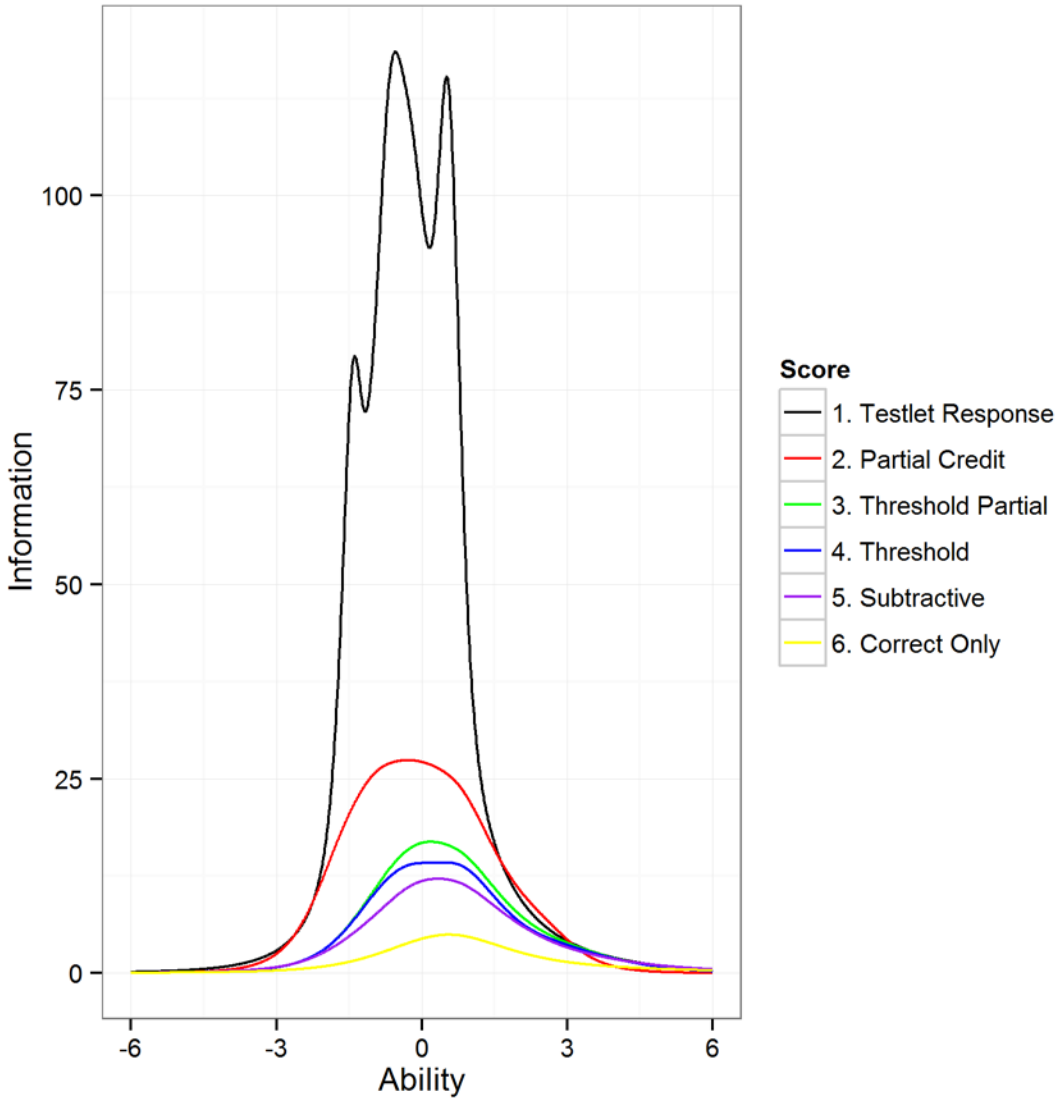


Figure 21. Comprehensive Agriculture Form A test information function comparison, tech only.

Figure 22 compares the test information functions for all scoring methods (except testlet response theory) on Comprehensive Agriculture Form B. Partial-credit scoring provides the most test information for Comprehensive Agriculture Form B. Additionally, the test information function for partial credit is wider than all the other scoring methods. This indicates that scoring items as partial credit provides more information at a wider range of abilities than utilizing the

other scoring methods. The lowest test information can be found with correct-only scoring. Additionally, all test information functions, except partial-credit scoring, are positioned with peak information just above a θ_i of 0.

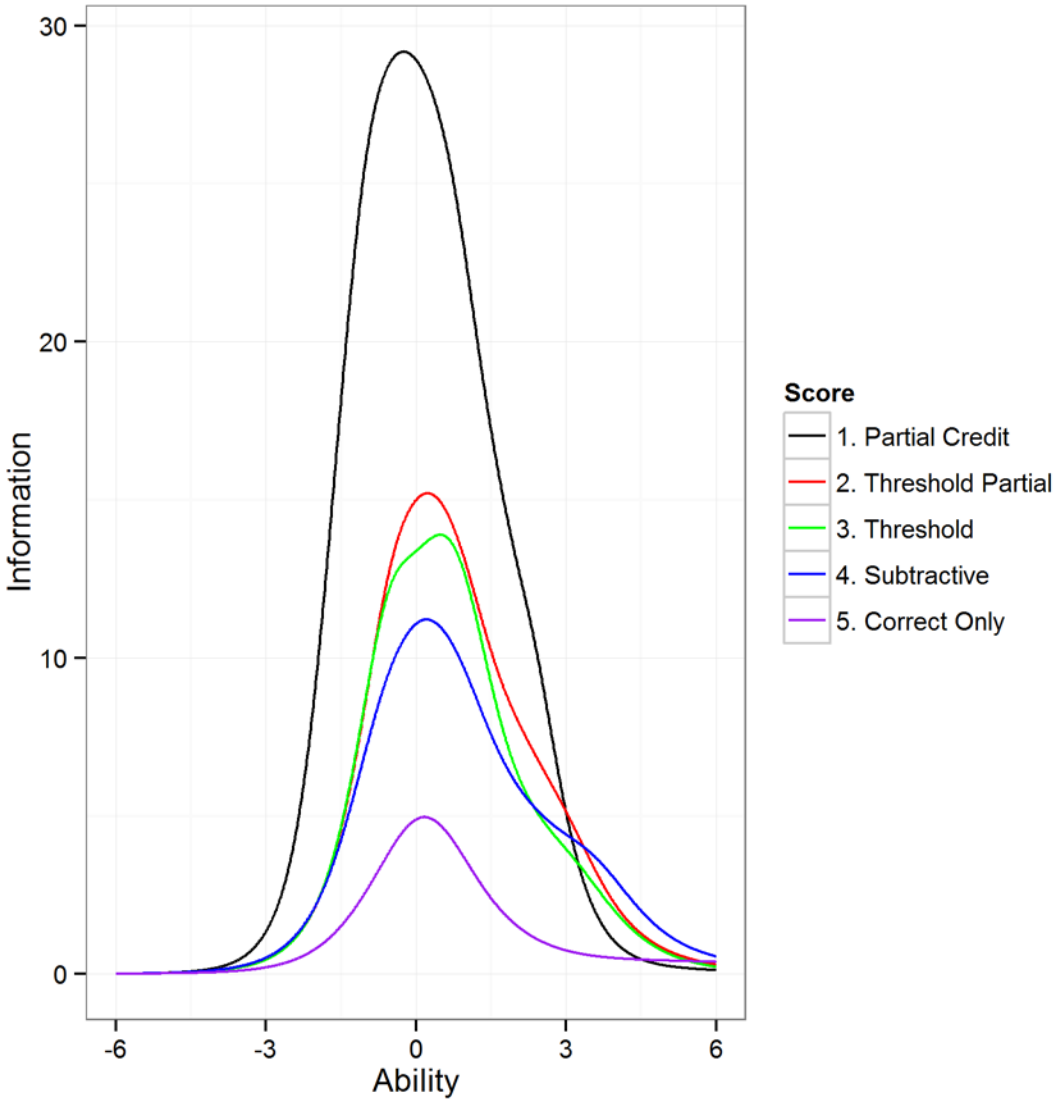


Figure 22. Comprehensive Agriculture Form B test information function comparison, tech only (excluding TRT).

Figure 23 compares all six TIFs for each of the scoring methods. Testlet response theory clearly provides the most test information, similar to previous findings. Unlike previous forms,

the TIF for this testlet response model is unimodal. Similar to previous findings, TRT provides far more test information than all the other scoring methodologies.

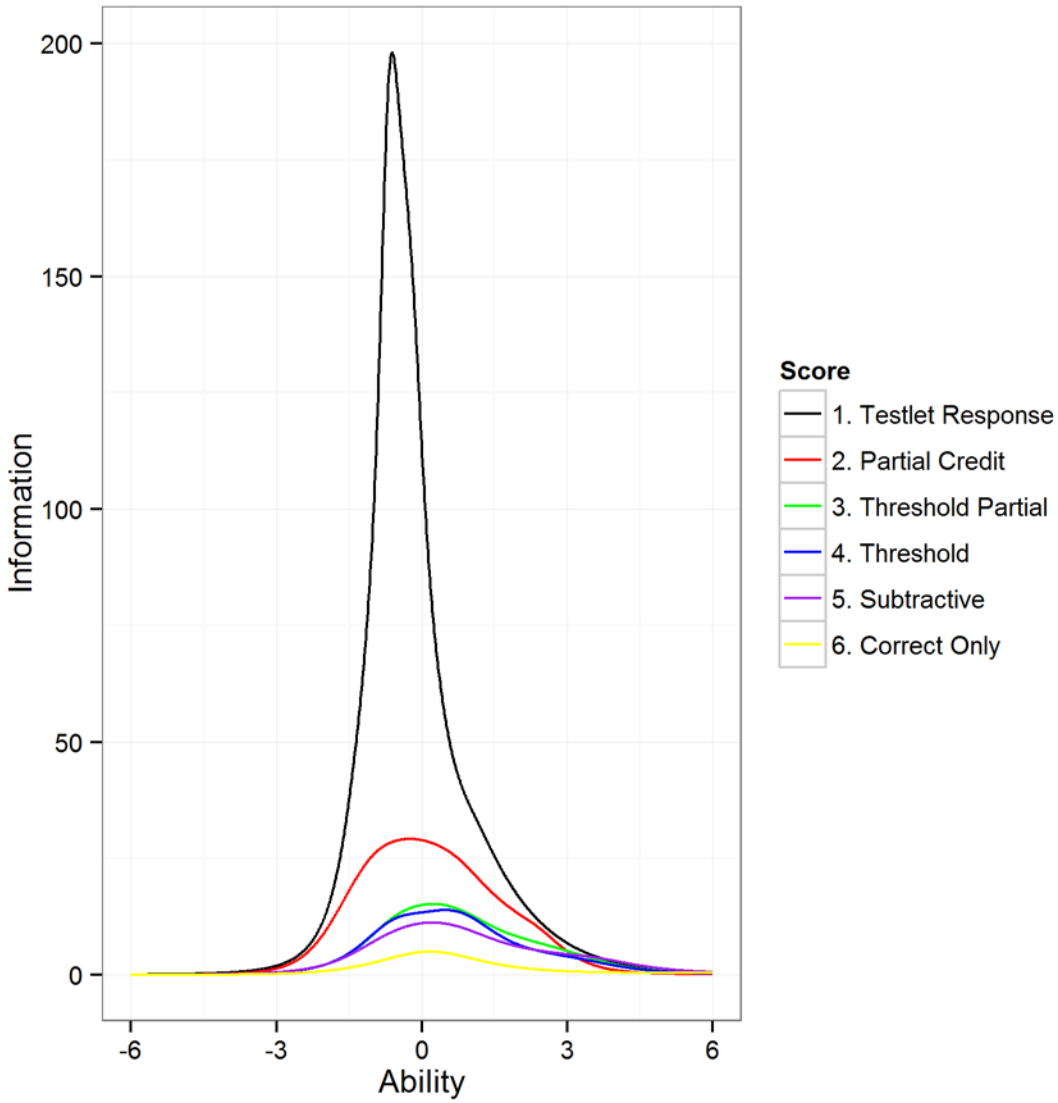


Figure 23. Comprehensive Agriculture Form B test information function comparison, tech only.

Figure 24 compares the test information functions for all scoring methods (except testlet response theory) on Comprehensive Agriculture Form C. Partial-credit scoring provides the most test information for Comprehensive Agriculture Form C. Additionally, the test information

function for partial credit is wider than all the other scoring methods. This indicates that scoring items as partial credit provides more information at a wider range of abilities than utilizing the other scoring methods. The lowest test information can be found with correct-only scoring. Additionally, all test information functions, except partial credit, are positioned with peak information just above a θ_i of 0.

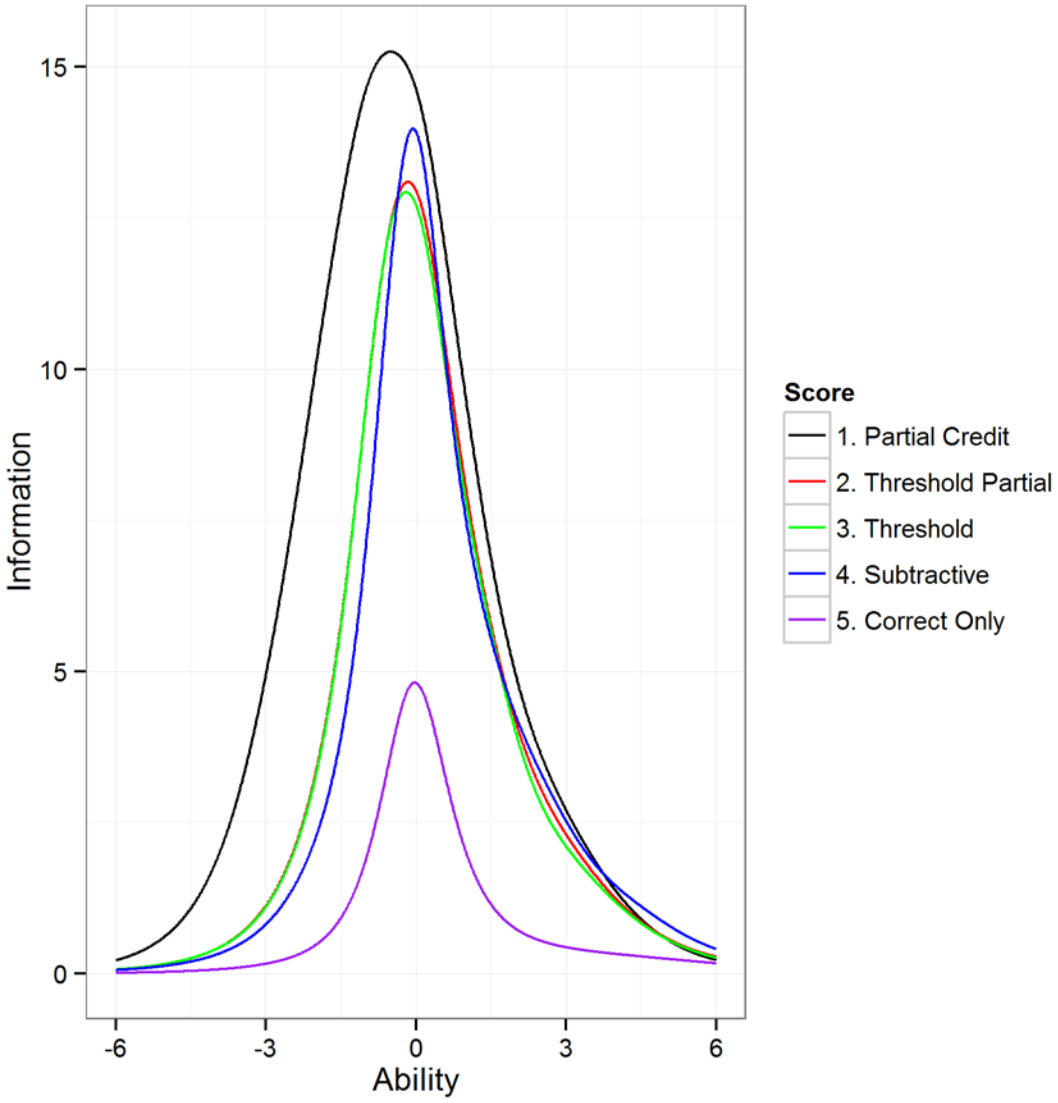


Figure 24. Comprehensive Agriculture Form C test information function comparison, tech only (excluding TRT).

Figure 25 compares all six TIFs for each of the scoring methods. Testlet response theory clearly provides the most test information, similar to previous findings. As with the previous form, the TIF for this testlet response model is unimodal, and very narrow at its peak. This indicates that it provides a lot of test information, but only at a very specific ability level.

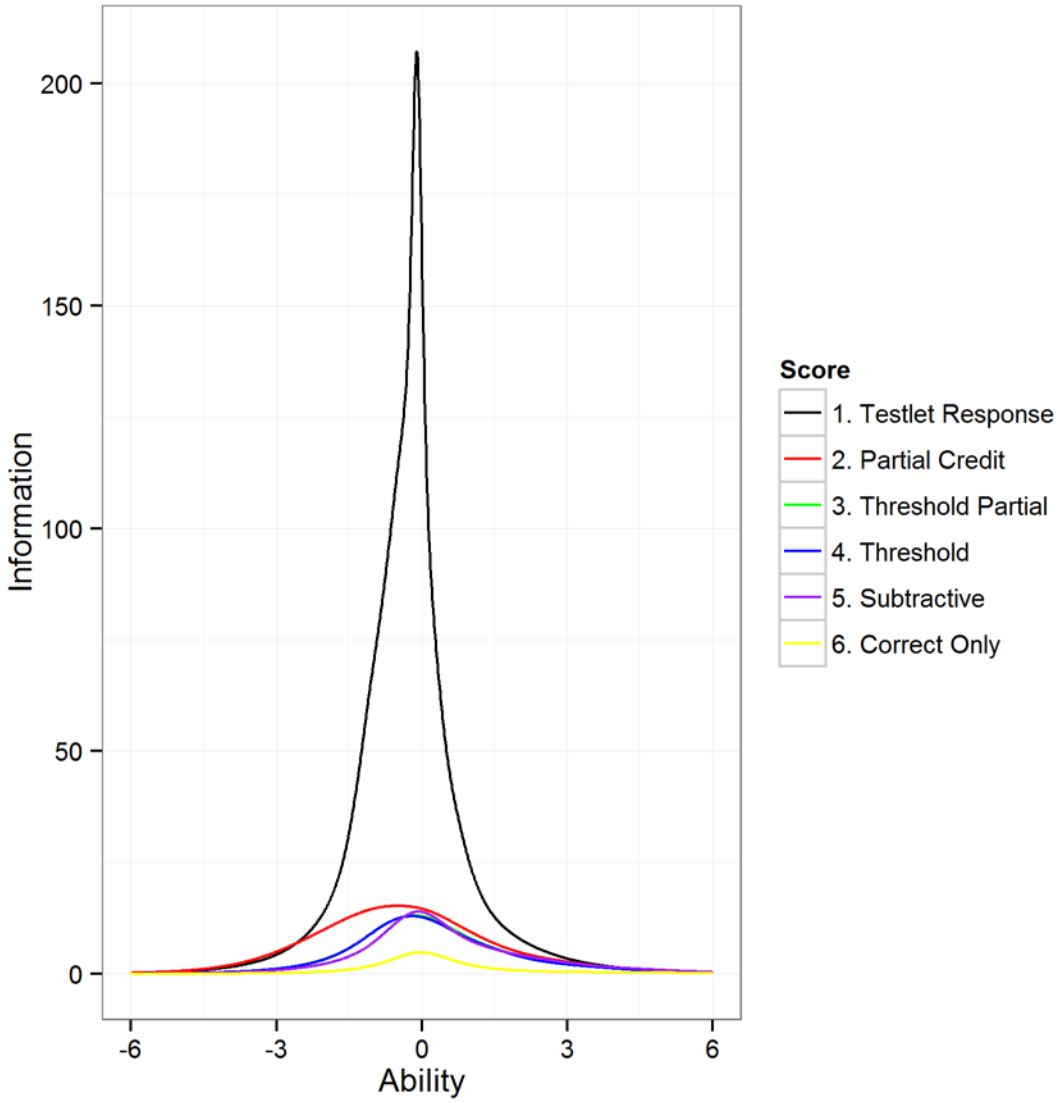


Figure 25. Comprehensive Agriculture Form C test information function comparison, tech only.

Research Question 3: When using IRT, how does adjusting the scoring method affect the model fit?

Model Fit

Standardized residuals were calculated and graphed for each of the IRT models. Standardized residuals represent the distance between what the model predicts, and the actual outcome based on data. The difference between these two are then graphically represented. When calibrating the five different scoring methods, any non-TE item with standardized residuals over 15 was removed, and the model was recalibrated. If TE items had standardized residuals over 15, the item was retained in the analysis. Standardized residuals were not calculated for testlet response theory scoring. The process of removing items due to poor fit could not be utilized in TRT, as all items are part of a testlet and are also all TE items. Since fit could not be improved, it is not advisable to compare TRT standardized residuals to other scoring methods' standardized residuals. For the other five scoring methods, the main point of comparison for standardized residuals is the density at and around 0. The larger the density at 0, the more items had perfect fit between actual and predicted outcomes.

Figure 26 shows a density plot of the standardized residuals for General CTE Form A for all five scoring methods calibrated utilizing all items. Correct-only scoring shows the best model fit with a density above .20 centered at 0. The worst fitting model was partial-credit, followed by threshold-partial and threshold scoring. Some of the misfit associated with partial-credit scoring could be due to the relatively low sample size and the extra parameters that were calibrated.

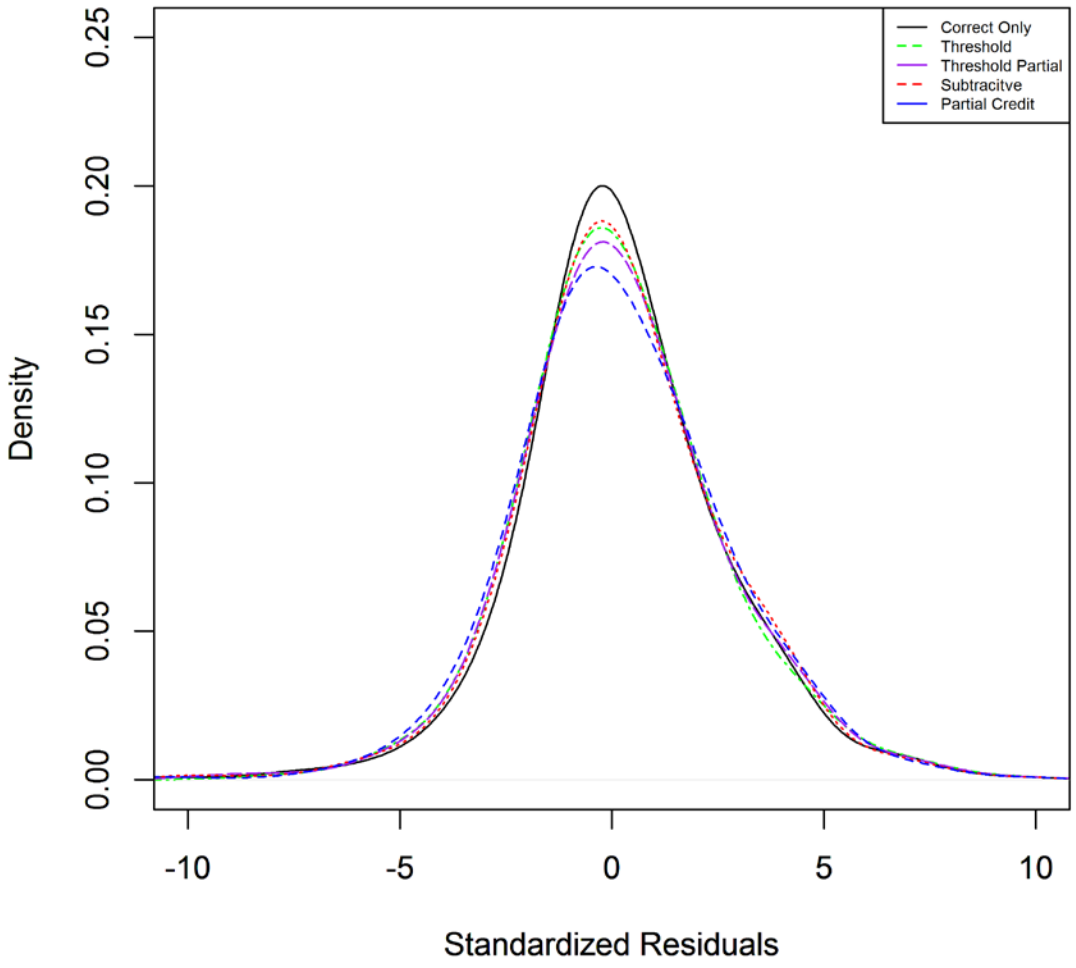


Figure 26. General CTE Form A density plot of standardized residuals.

Figure 27 shows a density plot of the standardized residuals for General CTE Form B for all five scoring methods calibrated utilizing all items. Correct-only scoring shows the best model fit with a density just below .20 centered at 0. The worst fitting model was partial-credit, followed by threshold-partial and threshold scoring.

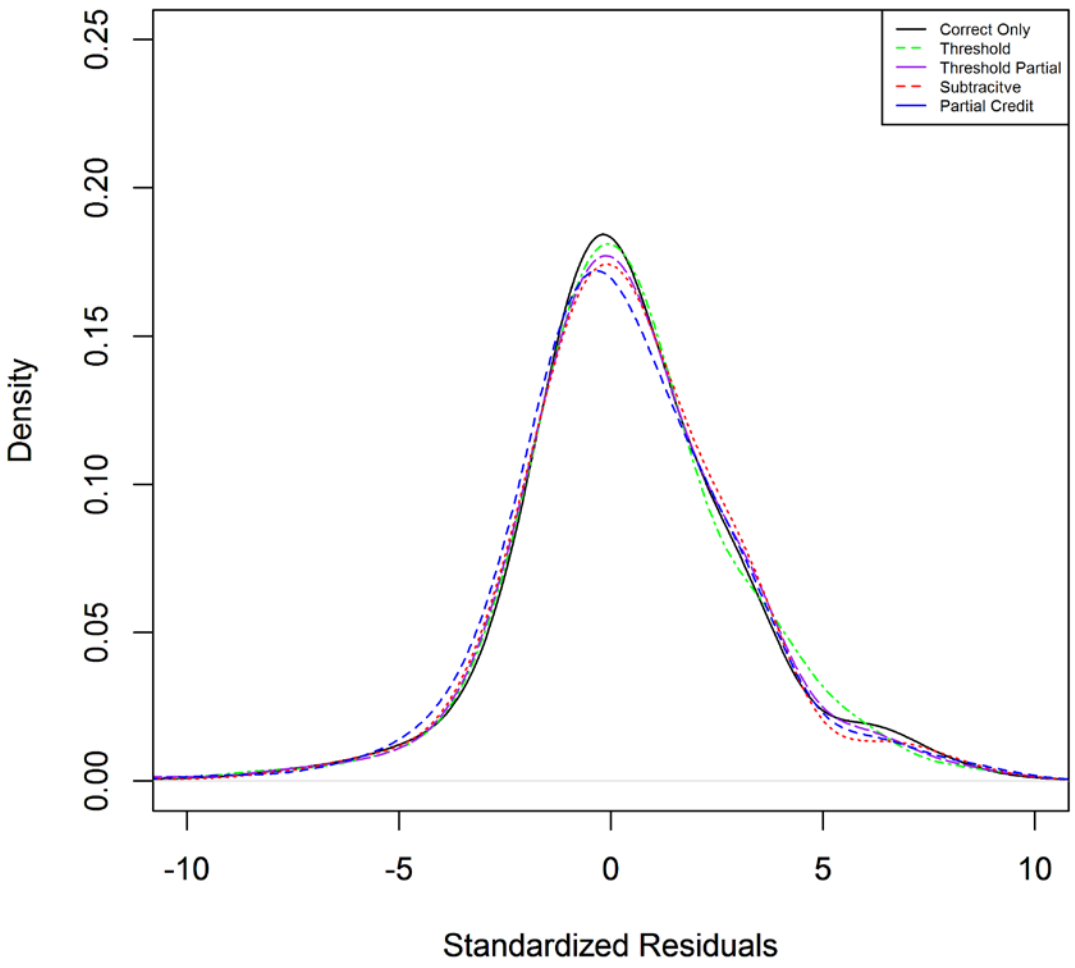


Figure 27. General CTE Form B density plot of standardized residuals.

Figure 28 shows a density plot of the standardized residuals for Comprehensive Agriculture Form A for all five scoring methods calibrated utilizing all items. Correct-only scoring shows the best model fit with a density just at .25 centered at 0. The worst fitting model was partial-credit, followed by threshold-partial and threshold scoring.

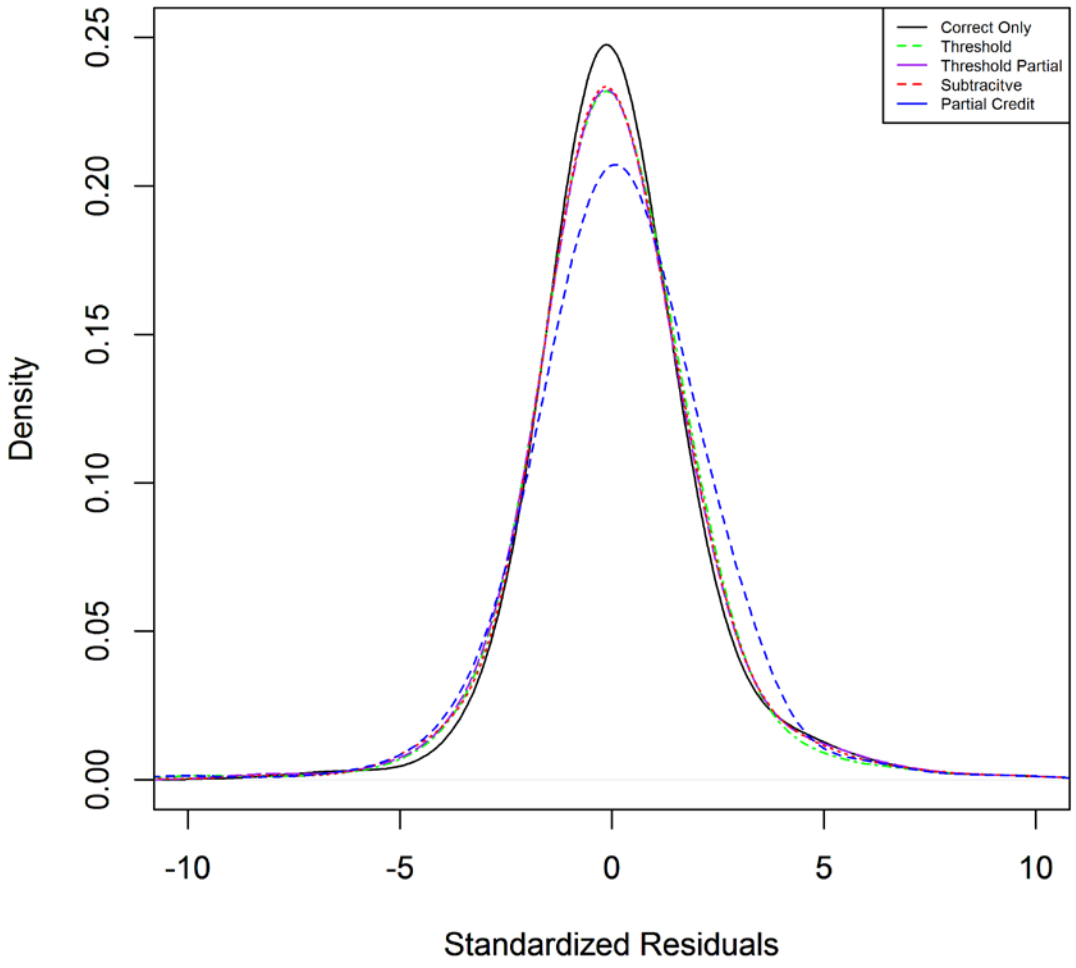


Figure 28. Comprehensive Agriculture Form A density plot of standardized residuals.

Figure 29 shows a density plot of the standardized residuals for Comprehensive Agriculture Form B for all five scoring methods calibrated utilizing all items. Correct-only scoring shows the best model fit with a density between .20 and .25 centered at 0. The worst fitting model was partial-credit, followed by threshold-partial and threshold scoring.

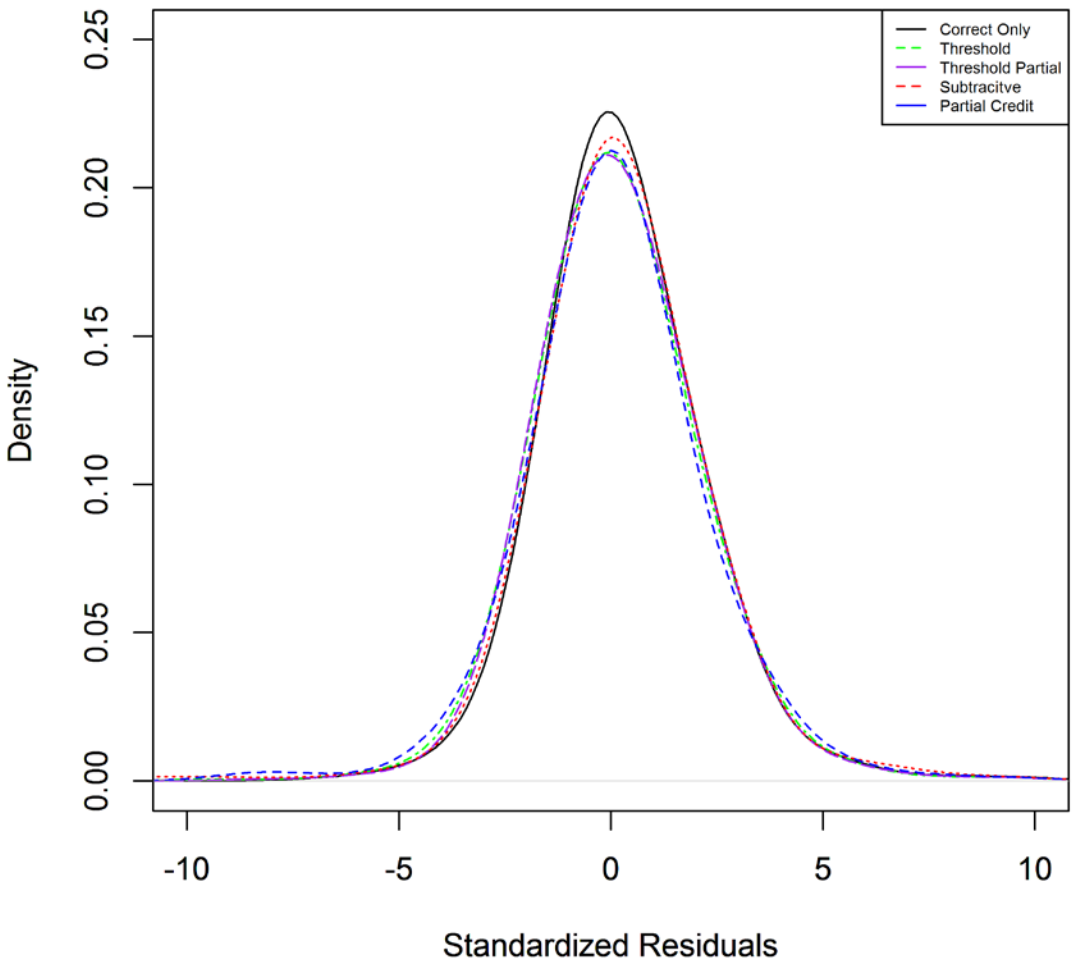


Figure 29. Comprehensive Agriculture Form B density plot of standardized residuals.

Figure 30 shows a density plot of the standardized residuals for Comprehensive Agriculture Form C for all five scoring methods calibrated utilizing all items. Threshold scoring shows the best model fit with a density approaching .25 centered at 0. The worst fitting model was subtractive, followed by partial-credit, threshold-partial, and correct-only scoring.

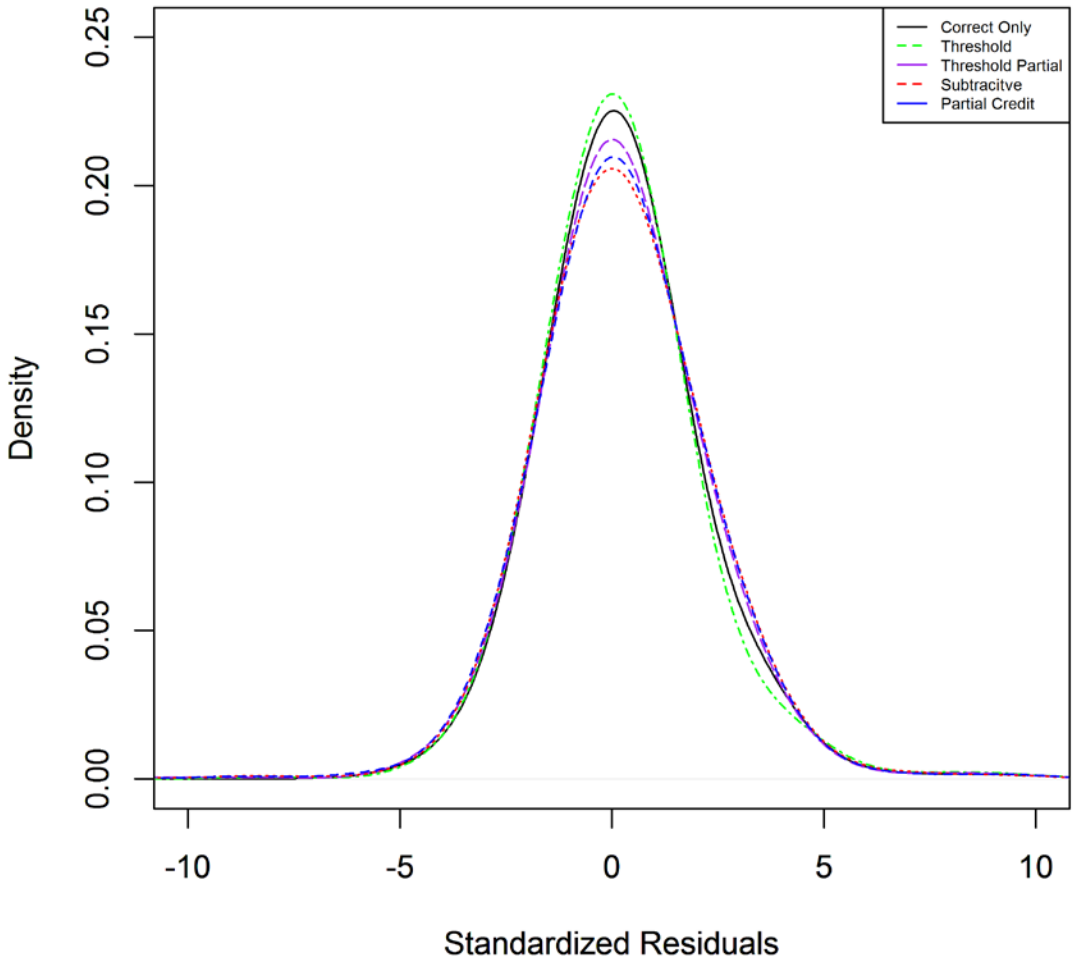


Figure 30. Comprehensive Agriculture Form C density plot of standardized residuals.

Figure 31 shows a density plot of the standardized residuals for General CTE Form A for all five scoring methods calibrated utilizing only TE items. Correct-only scoring shows the best model fit with a density at .25 centered at 0. The worst fitting model was partial-credit, followed by threshold, threshold-partial, and subtractive scoring.

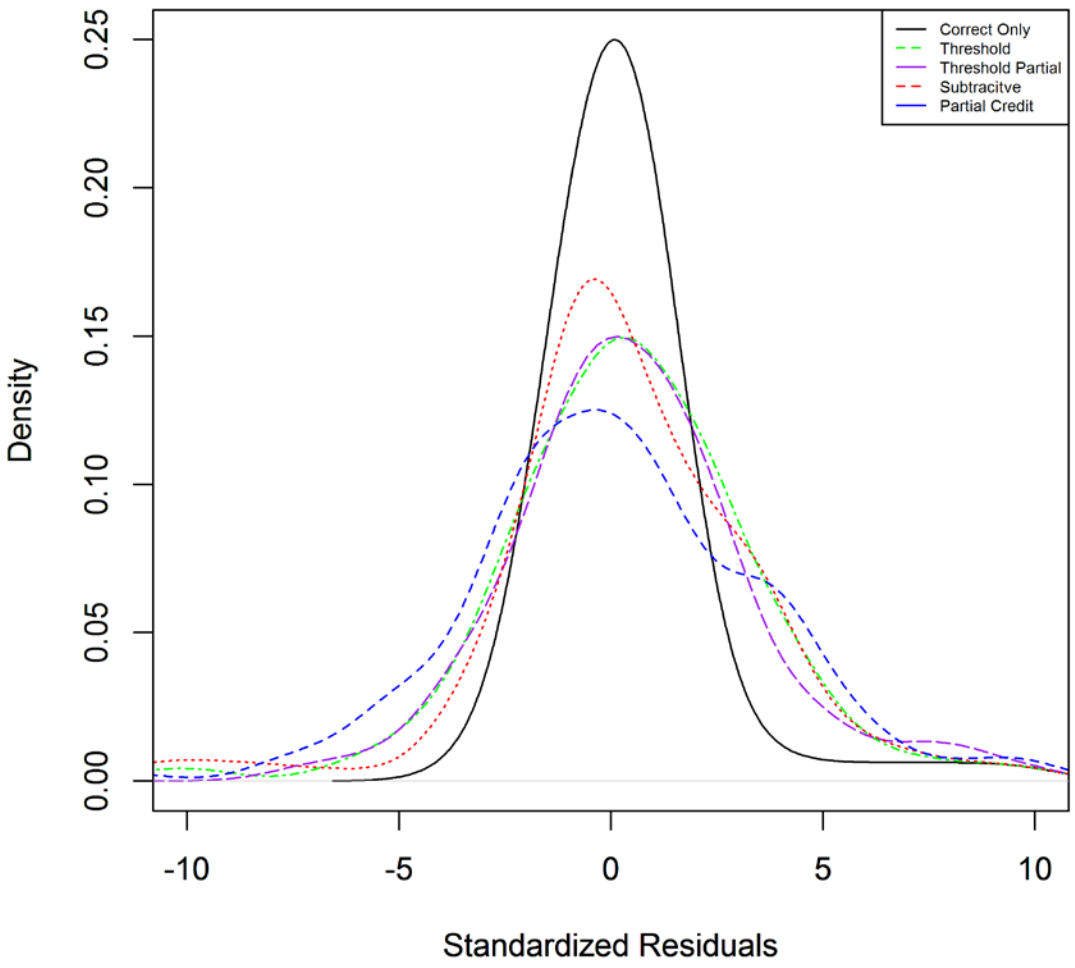


Figure 31. General CTE Form A density plot of standardized residuals, tech only.

Figure 32 shows a density plot of the standardized residuals for General CTE Form B for all five scoring methods calibrated utilizing only TE items. Correct-only scoring shows the best model fit with a density exceeding .25 centered at 0. The worst fitting model was partial-credit, followed by threshold, threshold-partial, and subtractive scoring.

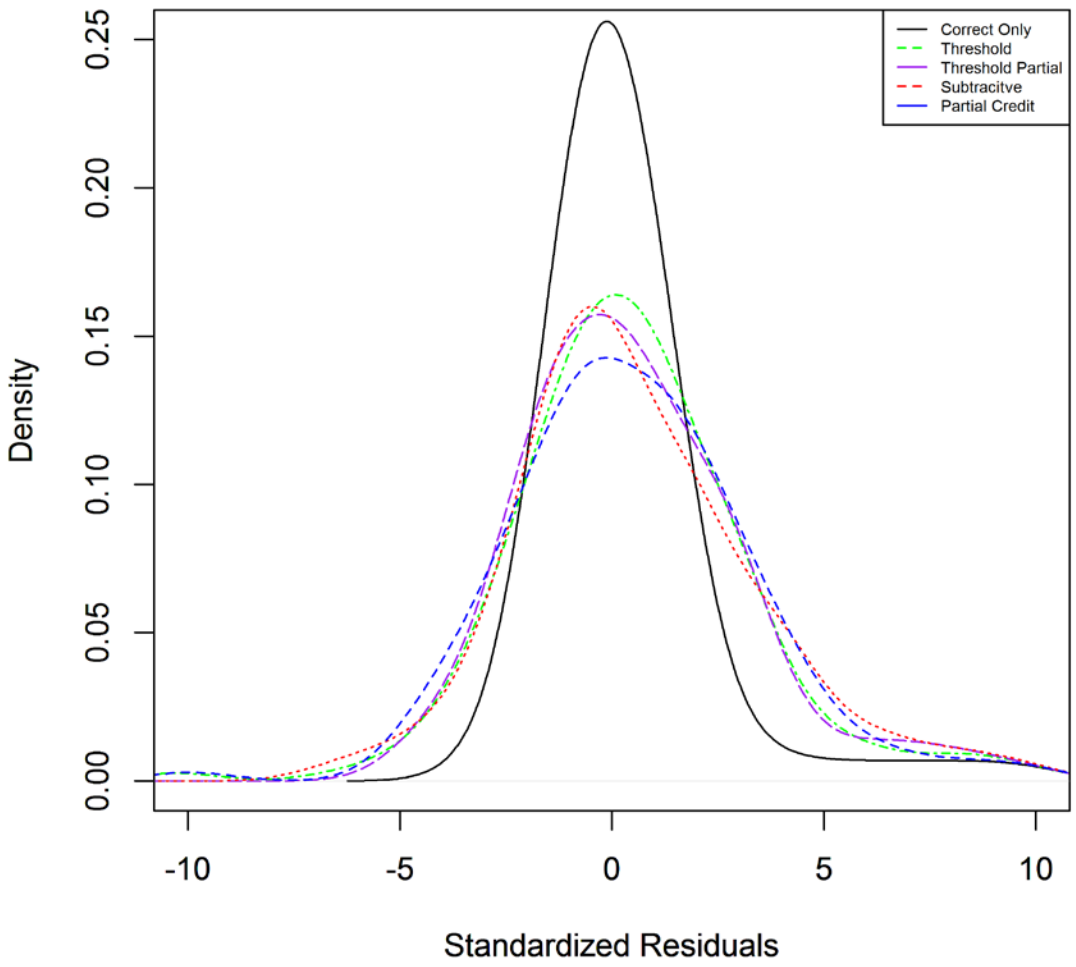


Figure 32. General CTE Form B density plot of standardized residuals, tech only.

Figure 33 shows a density plot of the standardized residuals for Comprehensive Agriculture Form A for all five scoring methods calibrated utilizing only TE items. Correct-only scoring shows the best model fit with a density approaching .30 centered at 0. The worst fitting model was partial-credit, followed by threshold, threshold-partial, and subtractive scoring.

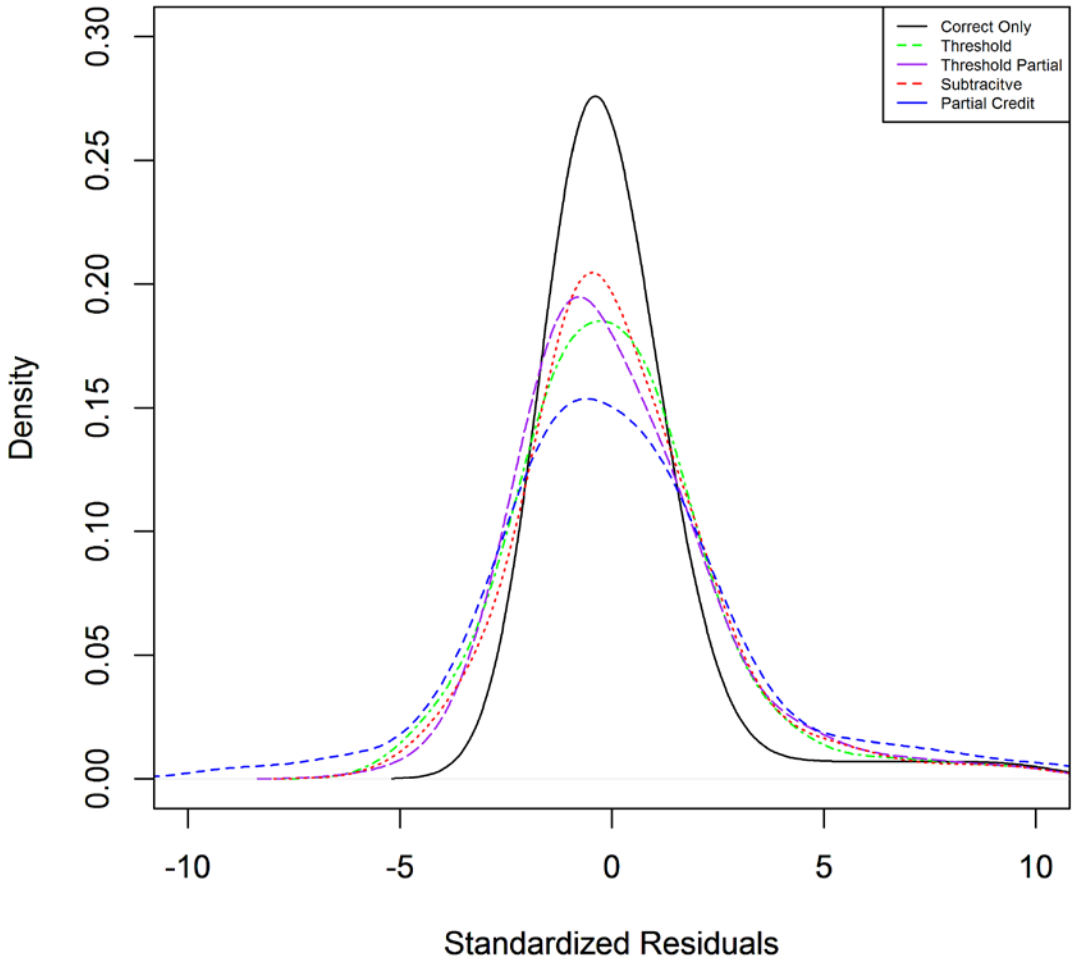


Figure 33. Comprehensive Agriculture Form A density plot of standardized residuals, tech only.

Figure 34 shows a density plot of the standardized residuals for Comprehensive Agriculture Form B for all five scoring methods calibrated utilizing only TE items. Correct-only scoring shows the best model fit with a density approaching .30 centered at 0. The worst fitting model was partial-credit, which has a slightly negative skew. The other three scoring methods are closer to normal, and centered on 0.

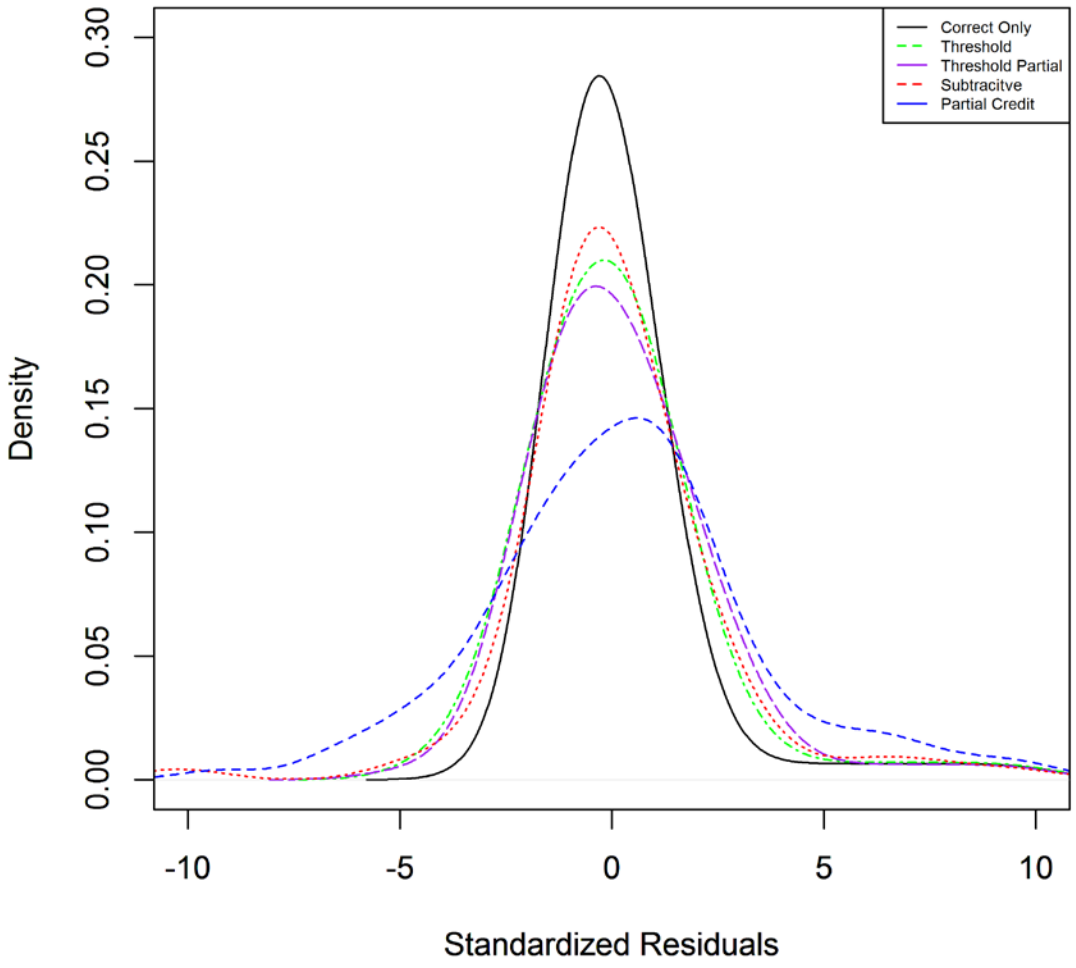


Figure 34. Comprehensive Agriculture Form B density plot of standardized residuals, tech only.

Figure 35 shows a density plot of the standardized residuals for Comprehensive Agriculture Form C for all five scoring methods calibrated utilizing only TE items. Correct-only scoring shows the best model fit with a density approaching .40 centered at 0. The worst fitting model was partial-credit followed by threshold partial, threshold, and subtractive scoring.

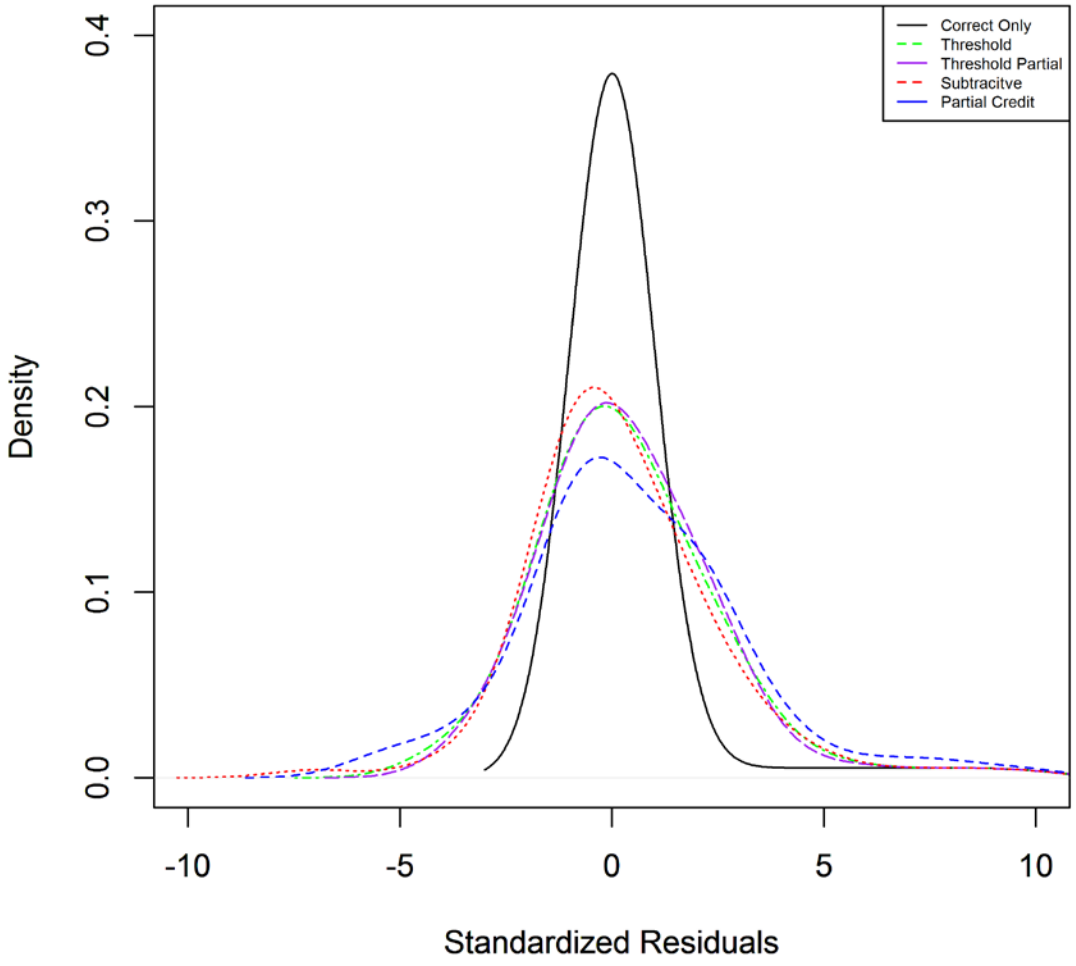


Figure 35. Comprehensive Agriculture Form C density plot of standardized residuals, tech only.

Chapter Five: Discussion

With the increased use of TE items, the understanding of how these items function operationally is important for test developers. Often, determining how items should be scored is an afterthought, and little attention is paid to how different scoring methods can effect basic testing outcomes. This study attempts to clarify which common scoring methods are the best to use when introducing TE items into an operational assessment. The results of this study can help create a baseline from which test developers can make quality decisions when building TE assessments.

The primary research questions of this study were separated by whether all items or only TE items were used to calculate or estimate the statistics. This separation allows for two different interpretations of the data. When including all items, the interpretation of the changes in statistics due to adjustments in scoring can be framed as a real-world effect. Specifically, this research will help test developers determine, at the test level, what changes can be expected when adjusting scoring strategies. Given that a majority of test development will most likely be a mixed-item format, determining the effect of changing a small proportion of items while holding the others constant will help practitioners understand the effects of their test-development scoring decisions. In contrast, calculating these differences only using TE items shows the isolated differences of each scoring method. While it is important to know the real-world effects, it is equally beneficial to understand what each of the scoring methods does to basic item and test statistics. By calculating and estimating these statistics for only TE items, test developers can examine the isolated effect caused by adjusting the scoring of these new item types. Though

there tends to be similarities between these two methods, the magnitude of the differences varies. For simplicity, we will discuss these differences as separate entities.

Item Difficulty

For this study, item difficulty was calculated utilizing both *p*-values and *b* parameters. *P*-values are the only statistic in this study that do not vary based on other items on an assessment. For this reason, *p*-values will only be discussed uniformly and not broken down by whether all items or TE-only items were used in the calculation of the statistic. Item difficulty using *p*-values provided the most consistent results between assessments and forms used for this study. Comparing the mean *p*-values for each of the five scoring methods produced nearly exact patterns. The scoring method with the lowest *p*-value (most difficult items) was correct-only scoring. The highest average *p*-value for correct-only scoring was .36 for Comprehensive Agriculture Form A. This *p*-value indicates that, on average, 36% of students correctly answered the TE items on this form. Conversely, partial-credit scoring provided the highest *p*-values across all forms and all assessments. The highest average *p*-value for partial-credit scoring was .65, which was also on Comprehensive Agriculture Form A. The interpretation of this *p*-value is slightly different because it was a polytomous item; however, the difference in *p*-values between correct-only and partial-credit scoring was statistically and practically significant. The other three scoring methods were less distinct, but still significantly different, with subtractive scoring producing harder items than threshold scoring, and threshold scoring producing harder items than threshold-partial scoring. Because some of these item types have up to 10 different correct responses, missing any of the correct responses results in an incorrect response for the item when scored as correct-only. Whereas, with partial-credit scoring, any correct response provides some score value. Additionally, for items with a large number of correct responses, threshold-partial

scoring becomes more like partial-credit scoring. Because threshold-partial allows for more score variability, threshold-partial scoring would be expected to score items as less difficult than threshold and subtractive scoring. The findings of this study are consistent with these expectations.

Using IRT, *b* parameters were estimated with all items and with only TE items. As with *p*-values, correct-only scoring produced the highest *b* parameters (hardest items), and partial-credit scoring produced the lowest *b* parameters when all items were used in calibration. Additionally, subtractive scoring revealed significantly harder items than partial-credit scoring on the Comprehensive Agriculture forms. For General CTE, the difference between partial-credit and subtractive scoring was not significantly different. Threshold and threshold-partial scoring also held the same pattern for the majority of forms, with threshold-partial scoring estimating lower *b* parameters than threshold scoring.

When calibrating with only TE items, the pattern was very similar to the *b* parameters calibrated using all items. In addition to the scoring strategies used with all item calibration, testlet response theory was also calibrated for the TE only form. The scoring method that calibrated the highest *b* parameters was correct-only scoring. The easiest items were scored with partial-credit or testlet response. Each of these methods produced similar levels of difficulties, and on most forms, were not significantly different from each other. These findings are consistent with expectations. As the number of ways a test taker can receive a higher score increases, the easier the item should become. Thus partial-credit scoring and TRT calibrated the lowest *b* parameters and correct-only, subtractive, threshold, and partial-threshold scoring calibrated the highest *b* parameters.

Overall, for both CTT and IRT, scoring TE items as partial-credit allowed for the easiest items. Given that items scored as partial-credit allow for the most variability of all the scoring methods used in this study, this finding makes sense. Partial-credit scoring is the most forgiving of scoring methods. Test takers can miss one or more of the potential correct response options and still receive a positive score on the item. These differences in item difficulty are even more extreme when calibrating *b* parameters with only TE items. When only TE items are used in the calibration, differences in *b* parameters are only due to the effects of adjusting the scoring methodology. Therefore, the differences seen in *b* parameters are more exaggerated and distinct, whereas the other three scoring methods had less distinct item difficulties.

Across all forms and assessments in this study, scoring items as correct-only resulted in significantly lower item difficulties. This finding was consistent for both *p*-values and *b* parameters. Bauer et al. (2010) reported that scoring MSMC items as correct-only caused the items to become substantially more difficult than when the items were scored as partial-credit. The current study's outcome was consistent with Bauer et al. (2010). For both CTT and IRT, scoring the items as correct-only allowed for significantly more difficult items than when those items were scored as partial-credit.

Item Discrimination

Item-total correlations were determined for each scoring method using all items and then with TE items only. Although the findings were consistent when all items and only TE items were used, they were far more pronounced when only TE items were used in calculations. Partial-credit scoring consistently provided the highest item-total correlations. Theoretically, this makes sense, as items scored as partial credit should have larger variances than items scored as correct only. The more variability there is, the greater ability for two variables to correlate. This

finding was consistent with Ripkey et al. (1996), who found that items scored as correct only had lower item discriminations than items scored polytomously. In the current study, the differences in mean item-total correlations between the different scoring methods were all in the same direction but with slightly different magnitudes. The largest difference in mean item-total correlations was found on Comprehensive Form B when only TE items were included in the calculation. The mean difference between correct-only and partial-credit scoring was .292. This difference in item-total correlations is highly statistically significant. As with item difficulty, the biggest differences were found between correct-only and partial-credit scoring. The differences in the other three methods using CTT were less distinct. Similar to *p*-values, subtractive scoring tended to produce smaller item-total correlations than threshold and threshold-partial scoring, although the significance of these differences depended on the form and the test.

When using IRT, the *a* parameters provided similar outcomes. Partial-credit scoring produced the highest *a* parameters when using all items in calibration. Although not much different than subtractive scoring, correct-only produced the lowest *a* parameters when all items were included in the calibration. When including only TE items for calibration, the *a* parameters were slightly less distinct. While not always significantly different than partial-credit scoring, TRT scoring produced the highest *a* parameters. Similar to previous findings, correct-only scoring produced the lowest *a* parameters when only TE items were included in calibration. Subtractive, threshold, and threshold-partial scoring did not produce distinctive *a* parameters. The significance of the differences between each scoring method was not consistent among different forms and different assessments.

Reliability

Coefficient alpha was calculated on all forms for all assessments using each of the five scoring methods. Additionally, standard error of measurement was also calculated for each scoring method. Across the majority of forms, partial-credit scoring consistently provided the highest reliability estimates. Even when all items were included to estimate reliability, adjusting the scoring strategy had an effect on the overall reliability. For example, on Comprehensive Agriculture Form B, reliability went from .931 for correct-only scoring to .943 for partial-credit scoring. Additionally, *SEM* estimates went from .413 for correct-only scoring to .404 for partial-credit scoring. Given that the test included far more non-TE items than TE items, this effect is substantial. As with item difficulty and discrimination, the differences between subtractive, threshold, and threshold-partial scoring were minimal.

When coefficient alpha was calculated with only TE items, the effects of the scoring adjustment were more dramatic. As with reliability of the whole forms, partial-credit scoring produced the highest reliability estimates. Additionally, correct-only scoring produced the lowest estimates. This finding is consistent with Bauer et al. (2010) and Albanese and Sabers (1988), who found that correct-only scoring of MSMC items provided low estimates of reliability. In the current study, the most dramatic difference between correct-only and partial-credit scoring was found on Comprehensive Agriculture Form B. Reliability went from .760 for correct-only scoring to .947 with partial-credit scoring. The second best method of scoring to increase reliability was threshold-partial scoring, which was superior to subtractive scoring.

According to Jodoin (2003), TE items scored polytomously lead to increased test information when compared to TE items scored dichotomously. In the current study, test information functions were created for forms utilizing all items and TE items only. Findings

showed a similar change in test information between scoring methods. When calibrated with all items, partial-credit scoring provided the most test information. The pattern of these functions was consistent across all forms and all assessments. The test information function for partial-credit scoring across forms tended to have a taller peak with a wider base than the other scoring methods. This indicates that when scoring items using partial credit, there is an increased capacity for estimation at a wider range of ability levels than with the other scoring methods. Correct-only scoring tended to be centered higher on the ability scale, indicating that correct-only scoring is better at estimating ability with higher-ability test takers than with lower-ability test takers. Even though the correct-only TIF is shifted to the right, it still remains within the other scoring methods' functions, meaning that all other scoring methods provide better estimations of a wider range of abilities than correct-only scoring. This finding alone points toward the benefits of scoring items as partial credit rather than using a dichotomous scoring strategy.

When calibrating only TE items, this pattern remains consistent. For TE items only, TRT was also estimated. As predicted, TRT provided far more test information than any other method of scoring. Comparing the other five scoring strategies with TRT is difficult, largely because TRT splits a 15-item TE assessment into a 75-item assessment. Therefore, the amount of possible information is far greater than for the other five scoring methods.

Model Fit

Standardized residuals were calculated for each form on each assessment with different scoring strategies. Standardized residuals indicate how close the model prediction is to the actual data. A perfect model would have standardized residuals of 0 across all items. The more non-zero residuals, the more misfit there is between the model and the data. Density graphs were

created to represent the distribution of standardized residuals between the different scoring methods. Across all scoring methods, correct-only scoring consistently had the best model fit.

This finding is contrary to what was theorized for this study. The standardized residual findings were opposite to all the other findings reported so far. Correct-only scoring provided the best fit to the data when comparing standardized residuals, while partial-credit scoring produced the worst standardized residuals. With partial-credit scoring, the misfit might be due to having to estimate more parameters (i.e., additional *b* parameters) with a smaller sample size than desired. By comparison, correct-only scoring had the least number of parameters to be estimated, thus allowing for better fit. This finding also occurred when utilizing only TE items. With the TE only calibration, the fit of correct-only scoring far exceeded that of the other scoring methods.

Implications

The results of this study are clear; the best methods for scoring TE items are partial-credit and TRT scoring. Scoring TE items with partial-credit scoring produces items that are not too difficult, have higher discrimination (both *p*-values and *a* parameters), have increased coefficient alphas, have lower *SEMs*, and provide more test information at a larger range of abilities. Scoring items utilizing TRT also provides beneficial item and test statistics. Though statistically speaking, TRT provides better results than partial-credit scoring, there are practical considerations that make partial-credit scoring a better overall choice for most testing programs. In contrast, scoring TE items as correct-only provides significantly less information about the test taker and does not utilize all the possible benefits of these new item types.

Statistically speaking, partial-credit and TRT scoring are superior to all other scoring methods evaluated in this study. However, in certain testing situations, non-statistical factors may play a role in selecting scoring strategy. For example, if a testing program is very high

stakes (for example medical licensure), a subtractive, threshold, or partial-threshold method may be a better choice for scoring.

A subtractive scoring method allows for a penalty for incorrect responses. In the case of a high-stakes test, the ability to allow for a penalty for incorrect responses may be desirable. For a test taker who is being tested on medical procedures, answering incorrectly isn't just benign, but rather could indicate a gap in knowledge that would be potentially dangerous to a future patient. For these types of tests, penalizing the test taker for information they don't know, as well as rewarding them for information they do know, may make sense.

Similarly, threshold or threshold-partial scoring may be beneficial in certain testing environments. In some cases, test developers may not be interested in all levels of knowledge. If a test taker can only respond correctly to one or two parts of an item, it may not capture the information test developers are trying to measure. For example, if an item asks the test taker to sort the parts of a formal letter into the correct order, the majority of test takers may correctly order the first and last parts of the letter. Creating a threshold would require test takers to demonstrate a certain level of knowledge before receiving any points toward their total scores. In this example, test takers would have to correctly order more than just the first and last parts of a formal letter to get credit. In cases like this, scoring items with a threshold could be beneficial.

Finally, scoring items with TRT provided high item discrimination and the most test information in the current study. Although it may be the best way to increase test information, scoring with TRT has some practical drawbacks. First, it is difficult to estimate. It also requires a higher sample size than other methods of IRT. Additionally, if an assessment is built to a test blueprint that attempts to balance content based on number of items, utilizing TRT will upset this balance. For example, if a blueprint requires three items from a certain subject, it's not as simple

as selecting three TE items. If those three TE items have five responses each, the amount of items actually estimated using TRT will be fifteen rather than three. To use TRT, test developers would need to adjust how blueprints are defined, as well as potentially limit the number of responses in a TE item.

Overall, the results of this study make practical sense. Items that allow more variability in student responses produce better item and test statistics. Breaking down what is actually happening when adjusting scores provides a different perspective into the results of this study. When we score an assessment, we are trying to predict a test takers ability on the construct of interest. How we score the assessment determines how well we assess that construct.

Scoring methods that reduce the variability of scores, actually cover up information about the test taker that is crucial to understanding their ability level. For example, if there is a TE item with four correct responses, each correct response provides slightly different information about the test taker. With correct-only scoring, that uniqueness is covered up, suggesting that only an understanding of all knowledge within that item is beneficial. In most situations, this is an incorrect way to view a multi-part item. Conversely, partial-credit scoring allows for all pieces of that four-part TE item to count towards the calculation of the test takers ability. These two methods are as opposite as possible, and provide the easiest example of scoring differences. The differences between subtractive, threshold, and threshold-partial scoring are more nuanced.

Subtractive scoring allows for the individual item response to provide information about the test taker, but also covers up that information if they answer an item incorrectly. For example, if the question asks to label the parts of a barn, and the test taker responds correctly to two of the labels but incorrectly to the other two, they would receive a 0 for the item. In terms of interpreting that ability, the conclusion would be that the test taker has no ability in the construct

of interest based on this single item. Obviously, the test taker does have some knowledge, but the subtractive scoring method covers up the score variability. Specifically, there are multiple ways to receive different score points. In the example above, a test taker could receive 0 points for not answering any of the item correctly, or 0 points for answering two items correctly and two items incorrectly. Since there are multiple ways to receive the same score point, it essentially hides the information test developers are interested in determining.

Threshold scoring leads to more difficult items than did partial-credit, and provides less item information and lower reliabilities/test information. With threshold scoring, test takers have to respond correctly to a specific number of items before they receive any credit. In this study, the number of items was 50% of the total number correct. If the test taker responded to less than 50% of the items correctly, they received 0 points, if they answered all items correctly they received 1 point, and if they answered 50% or more of the items (but less than all), they received .5 point. This scoring method provides additional variability than what correct-only scoring provides, and this increase in variability is responsible for the better test statistics reported. Though better than correct-only scoring, threshold scoring still covers up information about the test taker. As mentioned previously, this might be purposeful, and if so, using this scoring method would be a perfectly acceptable strategy. If however, the test developer does not have a theoretical reason to use a threshold, then important information about a test taker is lost with this scoring method. For example, if we have an item with eight possible correct answers, the threshold to receive any points would be to respond correctly to four out of the eight. Any test taker who answered three or less items would be treated exactly the same as a test taker who responded to none of the items correctly. Similarly, test takers who respond to four of the items, receive the same number of points as test takers who respond correctly to seven of the eight

correct answers. Essentially, this causes a sort of range restriction. Using this method takes a scale that would have nine possible point values (including zero), and reduces it to a scale with only three possible point values.

Similar to threshold scoring, partial-threshold scoring improves upon the variability in correct-only scoring. Threshold-partial scoring follows the same rules as threshold scoring. A test taker must answer correctly 50% of the items in order to receive any credit. Threshold-partial scoring is distinct from threshold because after a threshold is reached, it becomes a partial-credit item. This adjustment adds potential for more variability in scores compared to the threshold scoring method, but still covers up some ability levels. As mentioned with threshold scoring, those test takers who do not reach the threshold are all treated the same. Thus lower ability levels are covered up with this method of scoring. As with threshold scoring, this might not be a problem if it fits with the purposes of the assessment.

Testlet response theory also follows this principle. Since each possible correct answer is now treated as its own separate question, the most possible variability is extracted from each item. This is similar to what is happening on partial-credit scoring except for one important factor. With partial-credit scoring test takers can receive the same point value regardless of which correct response is selected. Specifically, each correct response is treated the same. Testlet response theory takes out the interchangeability of the correct responses. If a test taker correctly identifies a harder part of an item that is displayed in the estimation of their ability. Whereas with partial-credit scoring, the easiest correct answer and hardest correct answer will receive the same estimation of ability. This assumption that all correct responses within an item are the same is not likely met in most instances. Thus, TRT helps to pull out the uniqueness of all responses within the item.

Finally, it is important to remember that the differences in some of these scoring methods are directly related to the number of possible correct responses. As the number of correct responses decreases, the more these scoring methods become similar. For example, if an item has three correct responses, then threshold, partial threshold, and partial-credit scoring all become the same. Conversely, the more possible correct responses, the more variability in the different scoring methods.

Limitations

This study helps set a foundation for additional research into the effects of scoring methods on TE items. Unfortunately, the major limitation with this study was the overall sample size for both the General CTE and Comprehensive Agriculture assessments. This was particularly true for the Comprehensive Agriculture assessment. Although fit was acceptable, having a higher sample size would allow for more stable statistics across both assessments. In addition to the sample size, the type and size of the TE items themselves were a possible limitation. Some TE items had too few possible responses. In certain situations, this caused threshold and threshold-partial scoring and/or correct-only and subtractive scoring to have the same score values. The instances where the two scoring methods did not produce a different score clouded possible differences between the scoring strategies because the variation between these methods was reduced. Additionally, in a few instances, correct-only scoring had zero correct responses. When conducting the within-subjects analysis of variance, listwise deletion was utilized. This listwise deletion most likely reduced the differences found between correct-only scoring and all other methods. Therefore, the differences found could actually be greater than was indicated by this study.

Future Research

Future research should focus on isolating TE item types to determine if scoring strategies interact with the type of TE item. It is possible that different types of TE items are better served by different scoring methods. For example, it might be beneficial to score reordering/rearranging items as correct-only, due to their linear dependency. Future research could also focus on the design of TE items. Specifically, how many possible answer responses are ideal? This knowledge could help determine scoring strategies, as more item responses allow more variability between partial-credit, subtractive, threshold, and threshold-partial scoring that may not have occurred in this study. Finally, including a wider variety of TE items in different contexts would strengthen the findings of this study.

Conclusion

This study sought to determine the best scoring strategy for TE item types. The results of this study are consistent with results from scoring studies that used MSMC item types, indicating, that TE items in and of themselves may not be that different from MSMC items when it comes to scoring. As the first study to look at scoring using operational TE items, it can serve as a baseline for future comparisons.

The results of this study strongly indicate that when selecting a scoring strategy for TE item types, partial-credit scoring is statistically and practically the best option. Results also illustrate that the common use of dichotomous scoring is an inferior approach to scoring TE items. Correct-only scoring reduces the information provided by TE items and should be avoided in most circumstances. With the popularity of TE item types and the ability to machine score responses, the prevalence of TE item types will continue to increase. This study highlights the

importance of continuing to research the evidence of validity, reliability, and overall characteristics of these item types. As TE items become more prevalent, these findings can help guide future test development across a wide range of assessments.

References

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement, 25*(2), 111-123. doi: 10.1111/j.1745-3984.1988.tb00296.x

Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology, 57*(3), 221-237. doi: 10.1111/j.2044-8325.1984.tb00164.x

Bauer, D., Holzer, M., Kopp, V., & Fischer, M. (2010). Pick-N multiple choice-exams: A comparison of scoring algorithms. *Advances in Health Sciences Education, 16*(2), 211-221. doi: 10.1007/s10459-010-9256-1

Bennett, R. E., & Sebrechts, M. M. (1997). A computer-based task for measuring the representational component of quantitative proficiency. *Journal of Educational Measurement, 34*(1), 64-77. doi: 10.1111/j.1745-3984.1997.tb00507.x

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (pp. i-35). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1990.tb01348.x

Black, P. J. (1997). *Testing friend or foe?: Theory and practice of assessment and testing*. London, England: Routledge.

Chapman, C. J., & Toops, H. A. (1919). A written trade test: Multiple choice method. *Journal of Applied Psychology, 3*(4), 358-365. doi:10.1037/h0073002

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement, 34*(1), 21-41. doi: 10.1111/j.1745-3984.1997.tb00505.x

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31*(4), 295-311. doi: 10.1111/j.1745-3984.1994.tb00448.x

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review, 15*(1), 83-99. doi: 10.1023/A:1021323509290

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: Development of a rubric for rating partial credit. *Journal of Chemical Education, 90*(10), 1310-1315. doi: 10.1021/ed400247d

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Huff, K. L., & Sireci, S. G. (2001). Validity Issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16-25. doi: 10.1111/j.1745-3992.2001.tb00066.x

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement, 72*(3), 493-509. doi: 10.1177/0013164411422903

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*(1), 1-15. doi: 10.1111/j.1745-3984.2003.tb01093.x

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*(4), 357-372. doi: 10.1177/01466210122032226

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96*(5), 927-940. doi: 10.1037/a0023496

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Pub. Co.

Luecht, R. (2001, April). *Capturing, codifying and scoring complex data for innovative, computer-based items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688-695.

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28*(2), 131-145. doi: 10.1111/j.1745-3984.1991.tb00349.x

Osterlind, S. J. (1997). Style, editorial, and publication guidelines for items in constructed-response/performance formats. *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (pp. 203-214). Hingham, MA: Kluwer Academic Publishers.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129-148). Dordrecht: Kluwer Academic Publishers.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Issues in innovative item types. *Practical considerations in computer-based testing: Statistics for social science and public policy* (pp. 70-91). New York: Springer.

Quellmalz, E., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*(5910), 75-79. doi: 10.1126/science.1168046

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf.

Riddile, M. (2012). What's new about the common core state standards? *Principal Leadership, 12*(7), 38-42.

Ripkey, D. R., Case, S. M., & Swanson, D. B. (1996). A "new" item format for assessing aspects of clinical competence. *Academic Medicine, 71*(10), S34-36.

Rogers, W. T., & Ndalichako, J. (2000). Number-right, item-response, and finite-state scoring: Robustness with respect to lack of equally classifiable options and item option independence. *Educational and Psychological Measurement, 60*(1), 5-19. doi: 10.1177/00131640021970330

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment, 4*(6).

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. doi: 10.1111/j.1745-3984.1991.tb00356.x

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software,* 12(3), 1-16.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Vispoel, W. P., & Kim, H. Y. (2014). Psychometric properties for the balanced inventory of desirable responding: Dichotomous versus polytomous conventional and IRT scoring. *Psychological Assessment 26*(3), 878-891. doi: 10.1037/a0036430

Wainer, H., Brown, L., Bradlow, E., Wang, X., Skorupski, W., Boulet, J., & Mislevy, R. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. Williamson, R. Mislevy, & I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 169-197). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

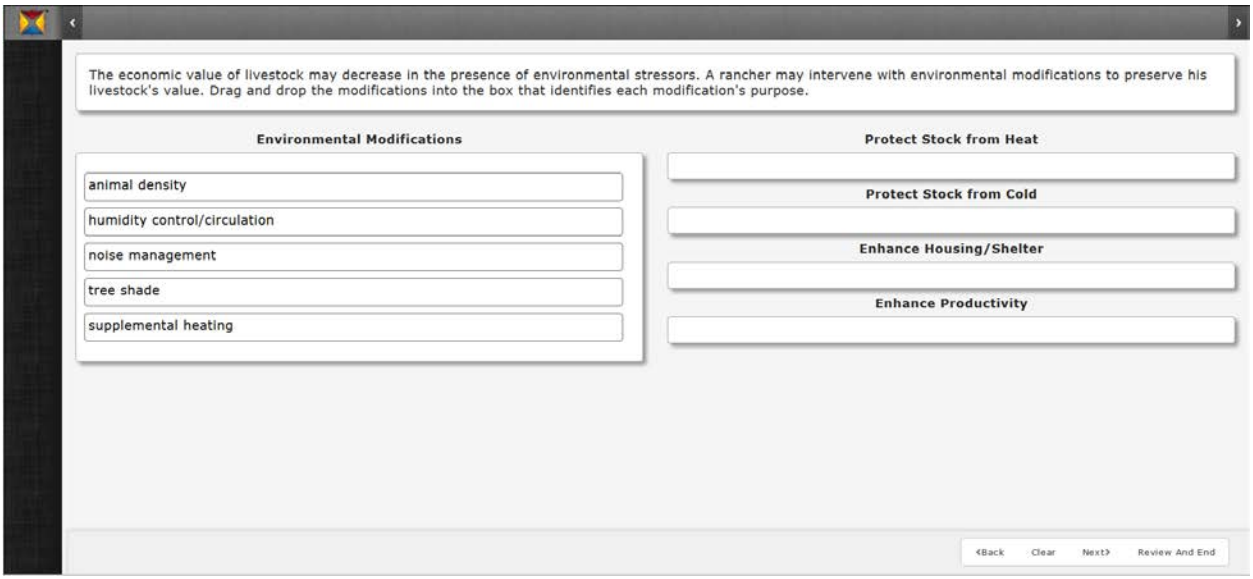
Wan, L., & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education, 25*(1), 58-78. doi: 10.1080/08957347.2012.635507

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(1), 109-128. doi: 10.1177/0146621602026001007

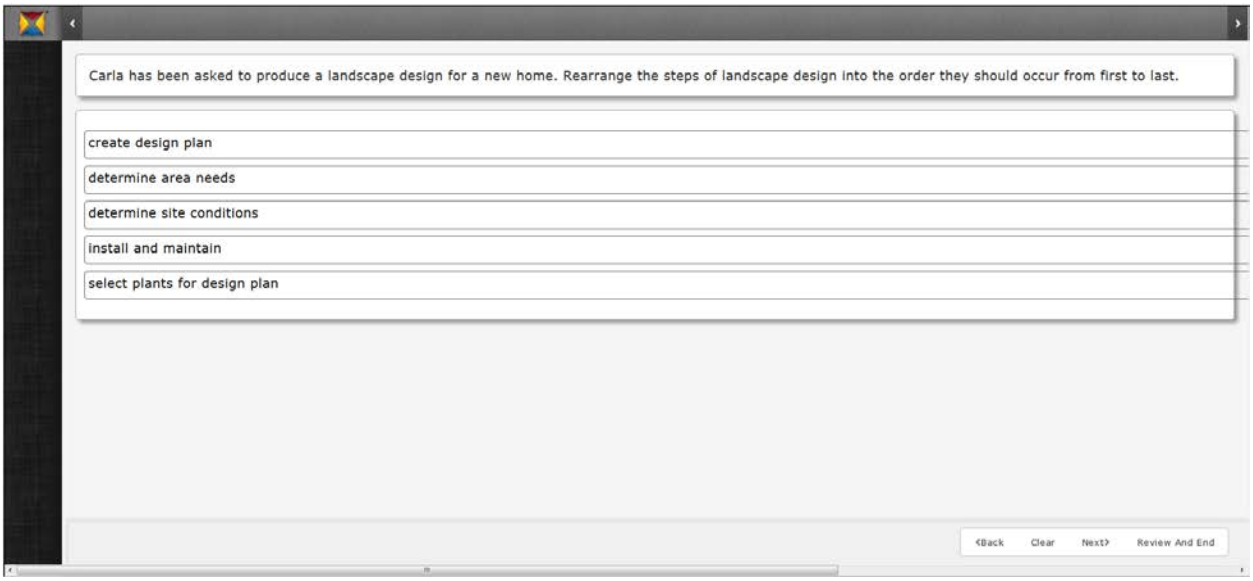
Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362. doi: 10.1207/S15324818AME1504_02

Appendix A

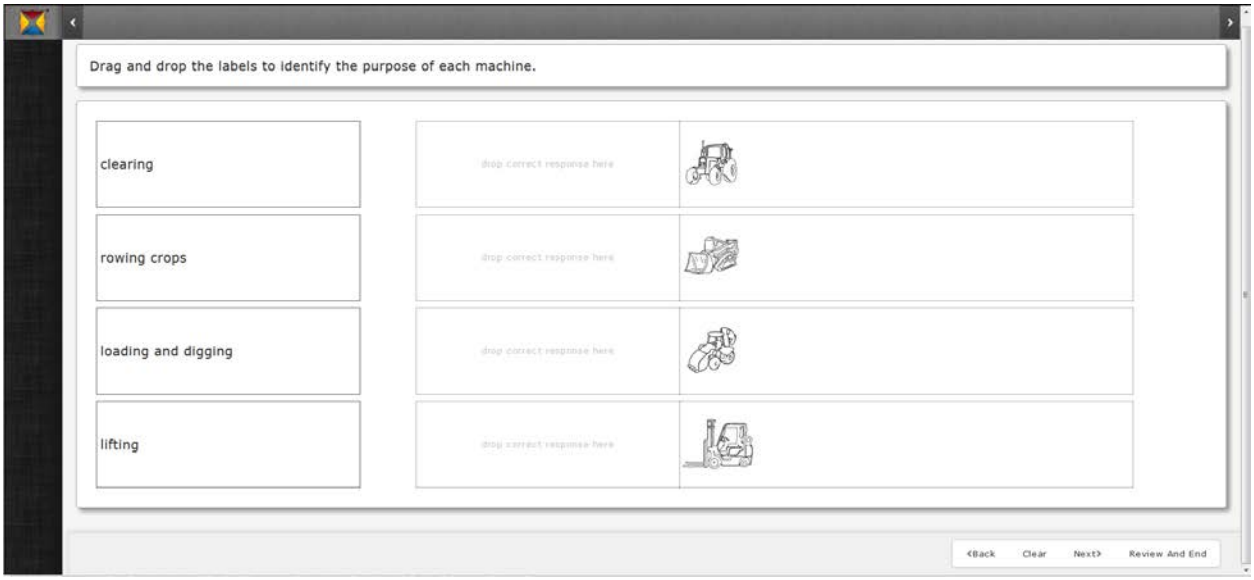
Multiple Drop Buckets Item Type



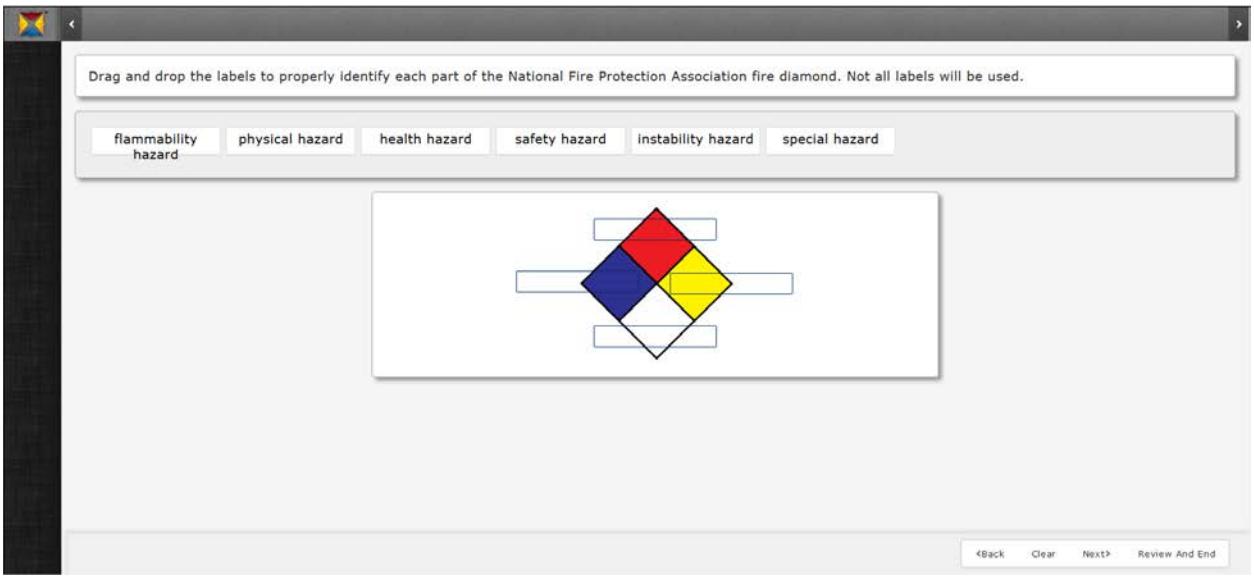
Reordering/rearranging Item Type



Matching Item Type



Figural-response Item Type



Appendix B

General CTE Assessment Test Specification

I.	**Academic Foundations**
A.	Demonstrate language arts knowledge and skills required to pursue the full range of postsecondary education and career opportunities.
1.	Recognize appropriate language for audience, purpose, and situation (e.g., diction/structure and style).
2.	Organize oral and written information.
3.	Create a plan for writing documents (e.g., notes, reports, and forms/documents).
4.	Construct focused copy for a variety of written documents (e.g., notes, reports, and forms/documents).
5.	Edit written documents (e.g., notes, reports, and forms/documents).
6.	Demonstrate comprehension of key elements of oral and written information (e.g., charts/tables/graphs, cause/effect, sequence, summaries, and compare/contrast).
7.	Evaluate oral and written information for accuracy, clarity, and relevancy.
8.	Project potential outcomes and/or solutions based on oral and written information (e.g., trends).
B.	Demonstrate mathematical and quantitative reasoning skills required to pursue the full range of postsecondary education and career opportunities.
1.	Apply basic arithmetic operations using whole numbers, decimals, percentages, and fractions.
2.	Demonstrate use of relational expressions such as equal to, not equal to, greater than, or less than.
3.	Use data and measurements to solve a problem.
4.	Recognize missing and/or irrelevant data in mathematical problem statements.
5.	Interpret charts/tables/graphs.
6.	Interpret and solve basic algebraic equations.
7.	Interpret functions that arise in applications in terms of the context.
8.	Demonstrate knowledge of basic geometry (e.g., area, perimeter, and volume).
9.	Demonstrate knowledge of basic statistics (e.g., mean, median, mode, and range).
10.	Use appropriate calculations in monthly personal budgeting, including income (e.g., net take-home pay) and expenses (e.g., mortgage, car loans, and living expenses).
C.	Demonstrate science knowledge and skills required to pursue the full range of postsecondary education and career opportunities.

General CTE Assessment Test Specification	
	1. Apply scientific reasoning (e.g., observation, data collection, controls, problem identification, and conclusions).
II.	**Information and Communication**
A.	Select and employ appropriate reading and communication strategies to learn and use technical concepts and vocabulary in practice.
	1. Determine the most appropriate reading strategy for identifying the overarching purpose of a text (e.g., skimming, reading for detail, reading for meaning, or critical analysis).
	2. Demonstrate use of content, technical concepts, and vocabulary when analyzing information and following directions.
	3. Interpret and communicate information, data, and observations from reading and apply the information to actual practice.
B.	Demonstrate use of the concepts, strategies, and systems for obtaining and conveying ideas and information to enhance communication in the workplace.
	1. Document information needed to report on a given topic or problem.
	2. Construct appropriate correspondence (e.g., business letter) that conveys and/or obtains information effectively.
C.	Locate, organize, and reference written information from various sources to communicate with coworkers and clients/participants.
	1. Locate written information used to communicate with coworkers and customers.
	2. Organize information to use in written and oral communication.
	3. Reference the sources of information used in communication.
D.	Evaluate and use information resources to accomplish specific occupational tasks.
	1. Review and apply informational sources for occupational tasks (e.g., informational texts, internet sites, and technical materials).
	2. Evaluate the reliability of information (e.g., informational texts, internet sites, and technical materials).
E.	Use appropriate grammar, punctuation, and terminology to prepare and edit documents.
	1. Organize clear, succinct, and accurate multiparagraph documents.
	2. Use descriptions of audience and purpose when preparing and editing documents.
	3. Use appropriate grammar, spelling, punctuation, and capitalization when preparing and editing documents.
F.	Interpret verbal and nonverbal cues/behaviors to enhance communication with coworkers and clients/participants.
	1. Interpret verbal behaviors when communicating with clients and coworkers.
	2. Interpret nonverbal behaviors when communicating with clients and coworkers.
	3. Apply factors and strategies for communicating with a diverse workforce.

General CTE Assessment Test Specification	
G.	Evaluate appropriate visual representations to support written and oral communications (e.g., tables, charts, figures, multimedia presentations, and demonstrations).
1.	Select appropriate visual representations to support written and oral communications (e.g., tables, charts, figures, multimedia presentations, and demonstrations).
2.	Interpret appropriate visual representations to support written and oral communications (e.g., tables, charts, figures, multimedia presentations, and demonstrations).
H.	Employ information management techniques and strategies in the workplace to assist in decision-making.
1.	Describe the nature and scope of information management.
2.	Maintain records to facilitate ongoing business operations.
III.	**Collaboration and Teamwork**
A.	Employ critical-thinking and interpersonal skills to resolve conflicts (e.g., with coworkers, peers, and customers).
1.	Analyze situations and behavior that affect conflict management.
2.	Determine best options/outcomes for conflict resolution using critical-thinking skills.
3.	Analyze the impact of emotions, needs, and concerns of others in an organizational setting (e.g., customers, peers, and coworkers).
4.	Identify stress management techniques.
5.	Identify solutions for resolving conflicts.
IV.	**Safety, Health, and Environment**
A.	Implement personal and jobsite safety rules and regulations to maintain safe and healthy working conditions and environments.
1.	Assess workplace conditions with regard to safety and health.
2.	Align safety issues with appropriate safety standards to ensure a safe workplace/jobsite.
3.	Identify safety hazards common to workplaces.
4.	Identify safety precautions to maintain a safe worksite.
5.	Employ a safety hierarchy and communication system within the workplace/jobsite.
V.	**Leadership**
A.	Employ leadership skills to accomplish organizational goals and objectives.
1.	Identify the various roles of leaders within organizations.
2.	Consider challenges related to leadership (e.g., diversity, environment, and global awareness).

General CTE Assessment Test Specification	
	3. Describe leadership characteristics (e.g., trust, positive attitude, integrity, and responsibility).
VI.	**Employability and Career Development**
A.	Identify work behaviors, personal qualities, activities, and resources that are needed to be employable.
	1. Manage resources in relation to an employee's position (e.g., budget, supplies, and computer).
	2. Identify or demonstrate positive work qualities typically desired.
	3. Manage work roles and responsibilities to balance them with other life roles and responsibilities.
	4. Demonstrate basic proficiency with common technology applications (e.g., spreadsheet, word processor, e-mail, and web browser).
B.	Maintain a career portfolio to document knowledge, skills, and experience in a career field.
	1. Select educational and work history highlights to include in a career portfolio.
	2. Evaluate pre-employment and work-history documents (e.g., résumé, certifications, and job applications).
C.	Identify and evaluate traits for retaining employment.
	1. Demonstrate understanding of required employment forms and documentation (e.g., W-4, I-9 form, work visa, and licensures).
	2. Identify key activities necessary to retain a job.
	3. Analyze positive work behaviors and personal qualities necessary to retain employment.
D.	Recognize and act upon requirements for career advancement to plan for continuing education and training.
	1. Identify opportunities for educational and/or career advancement.

Comprehensive Agriculture Test Specification

I.	**Agribusiness Systems**
A.	Describe agribusinesses and identify global opportunities in agribusiness systems.
	11. Define the types of ownership in an agribusiness.
	12. Identify significant markets in global agribusinesses systems.
B.	Evaluate record-keeping systems to assist in financial management of agribusiness.
	1. Recognize record-keeping and accounting principles.
	2. Use data to manage effectively an agribusiness (e.g., budget, cash flow, income and expense records, and balance sheets).
C.	Understand agriculture issues and important policies and laws in agriculture.
	1. Relate how agricultural laws and policies impact practices in agriculture industry.
D.	Identify principles of agriculture economics within an agriculture business.
	1. Apply the principles of supply and demand.
E.	Demonstrate knowledge of principles of agricultural marketing within an agricultural business.
	1. Illustrate the importance of a marketing chain.
	2. Describe the process of commodity marketing.
	13. Relate the segments of the agriculture industry and their distribution channels.
F.	Demonstrate knowledge of an agribusiness plan.
	1. List the key components of an agribusiness plan.
	2. Recognize the importance of goal setting in an agribusiness.
	3. Determine tax obligations regarding an agribusiness.
II.	**Animal Systems**
A.	Comprehend structure and significance of animal agriculture production systems.
	1. Evaluate the economic and global significance of animal systems.
	2. Describe the history of the animal agriculture industry.
	3. Communicate the process and movement of products from farm to table.
	4. Identify environmental issues relating to animal production.
B.	Comprehend the use of classification and taxonomic principles in animal agriculture.
	1. Recall the historical components of taxonomy in animal agriculture.

Comprehensive Agriculture Test Specification	
	2. Identify the general characteristics used to determine a breed (e.g., hair color, size, ears, etc.).
	3. Organize the components of taxonomy.
C.	Recognize the processes of animal growth and development.
	1. Identify key features and terms related to the process of animal growth and development.
	2. Explain cell structure and function.
	3. Describe the role and components of the following systems: circulatory, endocrine, digestive, muscular, nervous, respiratory, skeletal, and reproductive.
D.	Interpret the role of genetics and reproductive management in animal systems.
	1. Define key terms such as inbred, purebred, line-breeding, cross-breeding, etc.
	2. Summarize the principles of animal reproduction.
	3. Demonstrate the fundamentals of inheritance.
	4. Explore the process of animal selection and the role selection plays in improving animal systems.
	5. Identify current reproductive technologies in an animal breeding program.
E.	Recognize the components of animal health and wellness.
	1. Identify signs of diseases, parasites, and physiological disorders in animals.
	2. Explain the principle of immunity in animals.
	3. List common nutrients involved in animal growth.
	4. Interpret basic animal behaviors.
	5. Diagnose general signs of health in animals.
	6. Summarize environmental conditions on animal production.
	7. Analyze the need for safe, efficient, and industry-recognized standards for handling of animals.
F.	Understand basic principles of meat selection.
	1. Define key terms associated with meat quality and selection.
	2. Differentiate between wholesale and retail cuts.
III.	**Food Products and Processing**
A.	Describe the food products and processing industry.
	3. Determine the meaning and importance of food products and processing.
	4. Demonstrate knowledge of the history and global significance of food systems.
	5. Identify common units of measure as they relate to food processing.

Comprehensive Agriculture Test Specification
B. Identify world food needs.
1. Describe nutrition and the food plate (USDA's MyPlate).
2. Analyze the relationship between diet and population health.
C. Recognize the importance of food safety, sanitation, and quality.
1. Apply principles of food safety and sanitation, including the principles of HACCP.
2. Identify the role of regulating agencies and their responsibilities.
3. Demonstrate understanding of food system procedures as protection from bioterrorism.
4. Identify factors that affect food quality and deterioration.
5. Analyze the role of food product grading to provide consistency in food quality.
6. Analyze the role of inspection in maintaining food safety and quality.
D. Apply knowledge of the science of food products and processing.
1. Identify the role of substances (i.e., water, lipids, proteins, carbohydrates, vitamins, minerals, and food additives) in food chemistry.
2. Identify the role of substances (i.e., water, lipids, proteins, carbohydrates, and food additives) in food processing physics.
E. Identify food production procedures.
1. Describe food preservation procedures.
2. Describe storage and handling procedures.
IV. Natural Resources/Environmental Science
A. Apply the scientific principles of an ecosystem.
1. Describe the organization of life in an ecological system.
2. Differentiate between habitats and niches.
3. Illustrate cycles found in given ecosystems.
4. Identify the aspects of riparian and wetland areas.
5. Describe the effects of diseases and invasive species on ecosystems.
6. Examine the role insects play in ecosystem balance and health.
B. Recognize the importance of navigation and the variety of navigational tools.
1. Identify key terms associated with legal land descriptions.
2. Interpret topographical maps, their features, and their uses.
3. Recognize the importance of the compass and orienteering.
4. Describe the functionality of global positioning systems.

Comprehensive Agriculture Test Specification

C.	Recognize the components of wildlife management.
1.	Relate population dynamics to wildlife management.
2.	Explain wildlife animal adaptations.
3.	Discuss the effects of human interaction on wildlife areas.
4.	Explain the importance of species management and ethics.
D.	Identify the aspects of resource management and their importance.
1.	Differentiate between renewable and nonrenewable resources.
2.	Define terms associated with the management techniques of forestry, soil, land use, water, aquatic/marine resources, and air quality.
3.	Identify the importance and sources of energy resources.
4.	Describe the process of making resource management decisions.
E.	Comprehend the role of governing agencies involved in natural resources.
1.	Generalize issues and regulations related to water, air, land, and outdoor recreation.
2.	Evaluate the effect of waste and pollution on resources.
3.	Defend the use of natural resources for outdoor recreation.
4.	Interpret guidelines established for outdoor recreation areas.
V.	**Plant Systems**
A.	Comprehend structure and significance of plant agriculture systems.
1.	Determine the meaning and importance of plant systems.
2.	Compare and contrast traditional and nontraditional production trends in plant systems (e.g., conventional vs. organic, GMO vs. non-GMO).
3.	Identify plant production industry segments.
B.	Understand plant biology and apply principles in a plant systems production setting.
1.	Use plant classification systems (e.g., taxonomy, plant use, and life cycle).
2.	Identify aspects of plant growth, reproduction, and development.
3.	Identify the anatomy and function of plant parts, including cell structure.
4.	Apply knowledge of photosynthesis, transpiration, and respiration to plant production.
C.	Describe processes and techniques of plant environmental management.
1.	Comprehend the effect of the plant environment on growth and development, including water, air, light, temperature, and nutrients.
2.	Implement an integrated pest management plan.

Comprehensive Agriculture Test Specification
3. Identify safety practices and chemical control methods.
D. Identify the principles of field crop production.
1. Identify principles of crop management (e.g., planting, harvesting, and storage).
2. Identify basic irrigation systems.
E. Understand principles necessary to effectively manage range sites.
1. Define key terms associated with range and pasture management.
F. Understand and apply principles of greenhouse management.
1. Identify greenhouse function, design, and structure.
2. Identify and compare greenhouse-glazing materials for various applications.
G. Comprehend practices for establishing and maintaining turf and landscape areas.
1. Identify key components of landscape industry (e.g., design, installation, maintenance, and irrigation).
H. Apply management practices for soils.
1. Describe the factors of soil formation.
2. Identify physical characteristics of soil and relate them to soil management.
3. Analyze soil surveys and soil test analysis.
4. Identify causes and control methods of soil erosion.