5-8-2020

The effects of undesirable distractors on estimates of ability

Kathryn N. Thompson

The effects of undesirable distractors on estimates of ability

Kathryn N. Thompson

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2020

FACULTY COMMITTEE:

Committee Chair: Brian C. Leventhal

Committee Members/ Readers:

Deborah Bandalos

Christine DeMars

Acknowledgements

I would first like to thank my advisor, Brian Leventhal. I am so thankful to have had you as my advisor over the past two years. You have made the process of writing my thesis an incredible one. You have provided me with support and encouragement along the way, and I'm not sure I could have accomplished this without your guidance. You inspire me to wonder and think creatively in everything I do. You have taught me so many new things, and I look forward to continuing our research together.

I would also like to thank my committee members, Deborah Bandalos and Christine DeMars. I greatly appreciate all the feedback you have provided me. Debbi, I cannot thank you enough for your comments and feedback throughout my thesis. Not only did you address things I would have never thought about, but I feel that you have made me a better writer. Christine, I appreciate all of your help related to item response theory. A few months ago, these models seemed so complex, but your feedback has allowed me to actually understand both conceptually and mathematically what everything means. I really cannot express my thanks to both of you for everything.

Finally, I would like to thank my family and friends. Mom, dad, and Ian, you were always there to listen to me ramble on about my research. There were many times where you didn't have to entertain my ideas, but you discussed them with me as if you were my colleagues. Zach, there were numerous times we have debated about statistics and measurement. I don't think this process would have been as enjoyable as it was. Bri and Daigo, thank you for being such a great cohort. I am so proud of all three of us! It has been such a pleasure to get to know both you. And of course, I must thank all of the faculty and staff at CARS for these amazing two years and more to come. Thank you!

Table of Contents

List of Tables

List of Figures

Abstract

Distractors, or the incorrect options, are an important part of the multiple-choice item. Previous literature has supported the inclusion of distractors when estimating abilities. While the effects of well-functioning distractors on estimates of ability have been examined, research has neglected to examine the effects of undesirable distractors on estimates of ability. Undesirable distractors are defined as distractors that are opposite of what test-developers expect or want distractors to behave. For instance, an upper lure distractor is one that high ability examinees select rather than selecting the correct answer. A simulation study was employed to determine these effects by varying undesirable distractor type, percentage of items containing undesirable distractors, and test length. Item responses were generated using the Thissen-Steinberg multiple-choice model for simulating undesirable distractors and the three-parameter logistic model for simulating normal items. Following data generation, item responses were analyzed using the three-parameter logistic model in SAS. An analysis of covariance (ANCOVA) was used to examine the effects of undesirable distractors on estimates of ability for bias and standard error. Multiple significant interactions were identified for bias and standard error. One type of undesirable distractor that was especially problematic was the lower lure distractor, where high ability examinees have a slightly lower, but still high, probability of being selected in comparison to the correct answer. Additionally, a longer test resulted in the least amount of bias and standard error. Overall, test-developers should pay attention to the functioning of distractors, as there are effects of these undesirable distractors on estimates of ability.

Chapter 1. Introduction

Cognitive Items: Multiple-Choice & Constructed-Response

Researchers categorize cognitive items as selected response or constructed response (CR), with a multiple-choice (MC) item an example of the former. A MC item provides the examinee with one correct answer and multiple incorrect options (Briggs, Alonzo, Schwab, & Wilson, 2006). In contrast, a CR item requires an examinee to respond to an open-ended question (Lukhele, Thissen, & Wainer, 1994). Test developers must take into consideration the advantages of using one type of cognitive item over the other to measure examinees' knowledge. For example, test developers should consider the way in which students should exhibit the knowledge they have learned.

MC items have many advantages. To begin, MC items are objectively scored (DiBatitista & Kurzawa, 2011). Because MC items are either correct or incorrect, and therefore require no subjectivity in determining the correct answer, they have the same if not better reliability than CR items (Bacon, 2003; Simkin & Kuechler, 2005; Wainer & Thissen, 1993). CR items and MC items can result in similar scores (Hickson, Reed, & Sander, 2012). They have even been shown to validly measure the same type of cognition (e.g., analyzing) as CR items (Bennett, Rock, Wang, 1991; Traub & Fisher, 1977). MC items are also simpler to grade, and examinees spend less time taking a MC test in comparison to CR items. This advantage allows for a greater number of items on an assessment (Simkin & Kuechler, 2005). The addition of computers in testing has introduced instantaneous scoring, which reduces human error (Xu, Kauer, & Tupy, 2016). While CR items can be scored automatically (e.g., through artificial intelligence), it is a time-consuming process to program computers to read and score essays (Bennet, 1991). For these reasons, MC items are popular in the field of educational testing (Butler, 2018), and they are preferred over CR items (Haladyna & Downing, 1989).

While MC items do have several advantages, they have two major disadvantages: (1) it is difficult to create MC items that measure higher-level cognition and (2) examinees have the possibility of selecting the correct answer due to chance. While there is the possibility of writing MC items that measure high-level cognition, they are challenging to write. Even so, critics believe that MC items are unable to capture an examinee's ability to analyze or synthesize information. Instead, a majority of MC items measure an examinee's memorized information on an exam (Walsh & Seldomridge, 2006). However, some MC item proponent believe that MC items can measure high-level cognitive processes, but they require more time to construct (Aiken, 1982; Palmer & Devitt, 2007; Simkin & Keuchler, 2005; Tractenberg, Gushta, Mulroney, & Weissinger, 2013). Critics also express concern about examinees who possess little knowledge of the material being able to guess the correct answer on MC items (Kurz, 1999). Typically, though, most examinees do not guess without having first thought about their choice (Downing, 2003).

In contrast to MC items, CR items tend to measure more complex abilities, such as analyzing or evaluating learned information (Hancock, 1994; Martinez, 1999; Walsh & Seldomridge, 2006). This is due to the wide variety of ways in which test developers can ask an examinee to exhibit knowledge through a CR item (Livingston, 2009). Proponents also argue that CR items provide greater validity evidence for measurement of a skill since CR items measure complex skills (McClellan, 2010). While CR items have the primary advantage of measuring complex abilities, there are numerous disadvantages that researchers should consider before using them.

In a situation where test developers use raters to measure a skill, CR items run the risk of introducing bias when raters are not trained correctly. A rater can make different interpretations about whether a response is good or not, even when a rubric is provided (Livingston, 2009). Thus, CR items are more complex to grade than MC items (McClellan, 2010). Training raters requires a great deal of time and effort, which can be costly (Livingston, 2009; McClellan, 2010). When raters are not consistent in their grading of CR items, reliability and validity are both diminished. Although these effects are mitigated with training, they still exist and can be problematic. This leads to test scores that cannot be interpreted (McClellan, 2010).

While the selection of a MC or CR item is dependent upon the testing situation, there are advantages and disadvantages of each. MC items tend to be better in large-scale testing situations where items can be scored objectively and without human error (Xu, Kauer, & Tupy, 2016). However, this does not imply that MC items are completely free from issues. For instance, when test developers do not have sufficient time to develop MC items that measure more complex skills (e.g., memorization versus synthesis of information), CR items are a popular alternative (Hancock, 1994; Martinez, 1999; Walsh & Seldomridge, 2006). Although, CR items can be useful in many testing situations, MC items tend to prevail in popularity (Butler, 2018).

The Multiple-Choice Item

The history of standardized testing dates back to ancient Greece (Doyle, 1974), but this far preceded the invention of the MC item. Kelly (1916) was interested in assisting teachers measure students' knowledge in an effective way. This is the first time in literature that anyone had applied the MC item to the field of education. To do so, Kelly developed the Kansas Silent Reading Tests. The test was administered to evaluate students', such as third, fourth, and fifth graders, mastery of various subjects. Figure 1 provides an example of a MC item on this examination.

Three words are given below. One of them has been left out of this sentence: I cannot ____ the girl who has the flag. Draw a line around the word which is needed in the above sentence.		
Red	See	Come

Figure 1. MC item #3 on the Kansas Silent Reading Tests (Kelly, 1916).

Kelly (1916) believed that three criteria had to be met when writing MC items: (1) examinees should have the same interpretation of the item, (2) there should only be one correct option, and (3) only parts of the question that are related to the content being measured (e.g., not wording) should influence an examinee's response. All three criteria expressed by Kelly (1916) continue to be included in item-writing guidelines today (e.g., Haladyna, Downing, & Rodriguez, 2002).

With the development of the MC item, others began using it to create different kinds of tests. Otis developed a group-administered MC test in 1917 (Madaus, 1993). With this test, Otis was determined to create a better version of Binet's intelligence test (Russell, 2006). The United States Army used Otis' MC format when administering their intelligence test during World War I to evaluate recruits based on their test scores, allowing for a more efficient classification of skills (Madaus, 1993).

Following Kelly's (1916) development, the literature has been flooded with research on the properties and applications of MC items in educational testing. Some of the debates in the 1950's are still unsettled today, such as the format of the item (e.g., Dressel & Schmid, 1953; Ebel, 1971) or the optimal number of options to have for an item (e.g., Haladyna & Downing, 1993). A common theme in this research is to improve the MC item in order to more accurately determine an examinee's knowledge of the content.

It is important to understand each aspect of a MC item. A MC item has two components: (1) the stem and (2) the options. The stem is a question or phrase that prompts examinees to select one of the options. There are two different types of options, one correct answer and at least one incorrect answer commonly known as a distractor (Gierl, Bulut, Guo, & Zhang, 2017). To answer a MC item, the examinees select the best answer based on their interpretation of the stem. Ideally, only the examinees' knowledge of the construct being assessed should influence this selection. Unfortunately, there is documented research that irrelevant details in the stem, such as poor wording, have undue influence on the examinee's selection (Haladyna & Downing, 1989).

MC items can be administered to examinees in a variety of ways. Dressel and Schmid (1953) categorized MC items into those which had examinees recall information and those which had examinees apply recalled information. The former does not require the examinee to reason through the options. Instead, the examinee recognizes the information and selects an option based on memorization. Figure 2 provides an example of this, where an examinee needs to have memorized the formula for the area of a circle (Dressel & Schmid, 1953).

The area of a circle with diameter 1" is:

a) 2πsq. in. b) πsq. in. c) $\frac{\pi}{2}$sq. in. d) $\frac{\pi}{4}$sq. in.

Figure 2. Item from Dressel & Schmid (1953) where examinee must recall the formula.

MC items that ask examinees to apply recalled information are more complex since examinees must apply what they have learned to select a correct option. Figure 3 shows an example where an examinee applies learned information rather than simply recalling the formula (Dressel & Schmid, 1953). Examinees need to recall the formula for the radius of a circle, then apply that in finding the difference of radii of two concentric circles (Dressel & Schmid, 1953).

The area between concentric circles of 23-inch and 25-inch radius respectively is most easily found by the formula:

a) $\pi r_2^2 - \pi r_1^2$ b) $\pi(r_2 - r_1)^2$ c) $\pi(r_2 - r_1)(r_2 + r_1)$ d) $\pi(r_2^2 - r_1^2)$

Figure 3. Item from Dressel & Schmid (1953) where examinee must use reasoning.

Other Multiple-Choice Item Types

In addition to the formats Dressel and Schmid (1953) use, Haladyna (1992) synthesized five other types of item formats: true false (TF), multiple true false (MTF), alternate choice (AC), complex multiple choice (Type K or CMC), and context dependent item sets (CDIS). Each type of MC item format has advantages and disadvantages in comparison to one another (Haladyna, 1992).

The TF format is more frequently used in classroom assessment than in standardized testing situations because TF items do not require the development of options (Haladyna, 1999). TF items are statements in which an examinee must determine whether the statement is true or false (Downing, 1992). Figure 4 provides an example of a TF statement, where examinees are tested on their astronomy knowledge (Ebel, 1982).

Is the following statement true or false?
1) An eclipse of the sun can only occur when the moon is new.

Figure 4. Example of a TF item, where the answer is true (Ebel, 1982).

Opinions pertaining to the usefulness of TF items are mixed. Some research suggests that TF items test trivial content (i.e., memorized information), are highly influenced by guessing, and have low psychometric properties in comparison to other item formats (Downing, 1992; Haladyna, 1999). Grosse and Wright (1985) discuss guessing as a large error component in examinees' scores resulting in low score reliability (Pinglia, 1994). On the other hand, Frisbie and Becker (1991) suggest that TF items can evaluate higher-level thinking when test developers incorporate the guidelines for writing good items.

The MTF item is a way to test an examinee's basic knowledge of material (Downing, Baranowski, Grosso, & Norcini, 1995; Haladyna, 1999). MTF items are similar to traditional MC items. However, the stem in a traditional MC item is made into an option, and examinees answer each option as true or false (Haladyna, 1992). There is no set amount of true or false answers (Frisbie, 1992). Figure 5 shows a MTF item, where the item from Figure 3 is transformed into an item of MTF format.

Mark A if true, B if false.
Which of these is the formula(s) for the area between concentric circles of 23-inch and 25-inch radius, respectively?

1) $\pi r_2^2 - \pi r_1^2$ 2) $\pi(r_2 - r_1)^2$ 3) $\pi(r_2 - r_1)(r_2 + r_1)$ 4) $\pi(r_2^2 - r_1^2)$

Figure 5. Item from Figure 3 (Dressel & Schmid, 1953) transformed to an MTF format.

MTF items have been critiqued for the limited cognitive complexity that can be measured in items (Haladyna, 1999). MTF items are often not used due the association with the TF item format (Frisbie, 1990). In comparison to both TF and traditional MC item formats, MTF items do have multiple advantages. MTF items produce higher reliability estimates (Downing et al., 1999; Dudley, 2006; Frisbie, 1992; Frisbie & Druva, 1986; Frisbie & Sweeney, 1982; Kreiter & Frisbie, 1989), have a decreased length of time examinees take to answer the items, and allow test developers to include more items on the test in comparison to other item types (Dudley, 2006; Kreiter & Frisbie, 1989).

Although a strong proponent of the TF item, Ebel (1982) developed the AC item as a variant of the TF item. The item contains two options (i.e., one correct and one incorrect). Instead of presenting examinees with one statement, the statement is separated into a true version of the statement and a false version of the statement (Ebel, 1971). The word or phrase for each true and false statement is then listed as an option (Ebel, 1982). In Figure 6, the TF statement from Figure 4 follows the process of the TF item being broken into two statements. Figure 7 provides the AC item, where the differing words (i.e., new and full) in the two statements are used as the options.

1) An eclipse of the sun can only occur when the moon is new.
2) An eclipse of the sun can only occur when the moon is full.

Figure 6. The TF statement in Figure 4 divided into a true and a false statement. 1) true. 2) false. Item from Ebel (1982).

An eclipse of the sun can only occur when the moon is:
1) full 2) new

Figure 7. The AC item based on the two statements in Figure 4. Option 2 is correct. Item from Ebel (1982).

The AC item has multiple advantages in comparison to other MC item formats. To begin, examinees are able to reason through the options, where most MC items can be condensed from having four to two options (Ebel, 1982; Haladyna & Downing, 1993). The AC item format allows for the item writing process to be simpler for test developers, a larger number of items to be administered to examinees (Ebel, 1982), higher reliability estimates (Ebel, 1981), and a decrease in the cost of producing the exam (Haladyna, 2002).

The CMC item, or Type K item, is an item format where multiple options can be plausible (Haladyna, 1992). This item type can be beneficial when an examinee should know multiple pieces of information (Harasym, Leong, Violato, Brant, and Lorscheider, 1998; Hughes & Trimble, 1965). Options in this item type, such as all of the above or none of the above, are known as complex alternatives (Hughes & Trimble, 1965). This is because examinees must reason through each option, deciding whether it is correct or incorrect, before selecting their choice. In Figure 8, an item from a medical examination is presented. The examinee is asked to select all of the correct options (Burton, Sudweeks, Merrill, & Wood, 1991).

The fluid imbalance known as edema is commonly associated with:
1. Allergic reactions.
2. Congestive heart failure.
3. Extensive burns.
4. Protein deficiency.

The correct answer is:
A. 1, 2, and 3.
B. 1 and 3.
C. 2 and 4.
D. 4 only.
E. 1, 2, 3, and 4.

Figure 8. Example of a CMC item (Burton, Sudweeks, Merrill, & Wood, 1991). Option E is correct.

While administering this type of item to examinees has some merit, research suggests against using it. Complex options allow examinees with a partial knowledge of the subject to select the correct answer (Burton et al., 1991; Tarrant & Ware, 2008); examinees are still able to reason to the correct answer without knowing the information. In Figure 8, an examinee can eliminate all other options besides E if the examinee knows that the first and fourth options are correct, but unsure of second and third options (Burton et al., 1991).

CDIS items, or testlets (Wainer & Kiely, 1987), are popular in standardized testing due to their ability to measure higher level thinking (Haladyna, 1999). The most frequently used CDIS relate to reading comprehension items (Haladyna, 1999). Examinees are given a reading passage and asked to answer multiple questions based on their comprehension of the passage (Haladyna, 1992). CDIS items allow for test developers to have creativity in what they are developing (Haladyna, 1992). For example, test developers can use different types of reading passages while measuring similar content. However, the group of items has issues of inter-item dependence. This issue can negatively influence estimates of reliability and examinees' scores (Allen & Sudweeks, 2001; Thissen, Steinberg, & Mooney, 1989).

The traditional MC item is the most popular type of item on standardized assessments (Butler, 2018). Although other MC item formats can be used (e.g., TF or MTF), the most widely studied format is the traditional MC item format. Topics include accurately estimating item properties of MC items, examining the effects of poorly written MC items (Tarrant, Knierim, Hayes & Ware, 2006), or determining the optimal number of options in a MC item (Haladyna & Downing, 1993). The MC item as a whole has been the primary focus of research, but there is evidence that the distractors of a MC item are important to study (Gierl et al., 2017; Sideridis, Tsaousis, & Harbi, 2017; Thissen, Steinberg, & Fitzpatrick, 1989). By including distractors in the analysis of MC items, researchers have better accuracy in determining an examinee's ability in comparison to not including the incorrect answers (Bock, 1972; Levine & Drasgow, 1983; Thissen, 1976). Because distractors are useful components of MC item analysis, there is a need to understand the effects of the type of distractor on estimating examinee ability. In these studies, distractors were informative about ability, but no literature specifies the effects of distractors that are uninformative on or even detrimental to estimates of ability. Researchers should be concerned about the potential consequences that these types of undesirable distractors have on estimates of ability.

In the current study, I extend the process of evaluating distractor functioning by examining the effects of undesirable distractors on estimates of ability. In order to understand how undesirable distractors (i.e., distractors that do not provide any pertinent information) may come about, MC item writing guidelines are presented. An inattention to the guidelines can result in undesirable distractors. Once items are written, researchers are only able to identify undesirable distractors through the use of a distractor analysis. I believe that undesirable distractors will have negative impacts on estimates of ability.

Chapter 2. Literature Review

Writing and Developing Multiple-Choice Items

Flaws in MC items may negatively influence examinees’ responses rather than examinees’ ability of the construct of interest. For example, in a MC item, examinees’ ability levels should influence their choice instead of irrelevant details in the stem (Kelly, 1916). Estimates of examinees’ ability levels have the possibility of being biased if parts of a MC item are poorly worded (Haladyna et al., 2002). To combat this issue, item-writing guidelines have been created that suggest how to write good MC items.

Haladyna and Downing (1989) synthesized MC item-writing guidelines from literature in the educational measurement field to create a comprehensive review. Haladyna et al. (2002) later validated 31 of the original 43 MC item-writing guidelines by studying their use in literature. To do so, they recorded the frequency of authors who had ruled for, made no mention of, or ruled against the use of the 31 guidelines. The guidelines are organized into five sections about content, formatting, style, writing of the stem, and writing of the options.

Similar to Haladyna and Downing (1989), Haladyna et al. (2002) stress the development of the content, formatting, and style of the overall item including both the stem and all options. As shown in Table 1, the first 13 guidelines refer to general item writing. These aspects pertain to all parts of a MC item.

Table 1

Content, Formatting, and Style Item-Writing Guidelines

Content	Formatting	Style
1. Single content 2. Non-trivial content 3. Novel material 4. Keep items independent 5. Avoid general/specific 6. Avoid opinions 7. Avoid trick items 8. Simple vocabulary	9. Use correct item format 10. Format vertically	11. Edit and proof items 12. Use correct grammar, punctuation, capitalization, and spelling 13. Minimize amount of reading for items

Note. Guidelines are from Haladyna et al. (2002).

Of the thirteen guidelines, only the tenth guideline has disagreement amongst researchers. While most researchers believe MC items should be vertically formatted, 11% disagree that this should be a requirement. Haladyna et al. (2002) suggest that the only exception to this rule be when trying to save space on paper. Research on the effect of item formatting on examinees' abilities has yet to be done (Haladyna et al., 2002).

Inconsistencies with the style of an item can lead examinees to use test-wiseness to answer items correctly instead of answering items based on their actual abilities (Dolly & Williams, 1986; Millman, Bishop, & Ebel, 1965). Test-wiseness is defined as the ability to use characteristics of an exam to gain a higher score. As an example, when the item and options are not grammatically consistent examinees can easily eliminate distractors, as it is typical that the inconsistencies are with the distractors, not the correct answer (Frary, 1995). Because of this, items become easier for examinees (Dunn & Goldstein, 1959; McMorris, Brown, Snyder, & Pruzek, 1972).

There are several item-writing guidelines that refer to only the writing of the stem. Violating the stem-writing guidelines in Table 2 may lead to unnecessarily difficult items, which decreases the accuracy in the interpretations made about the scores from the test (Downing, 2002).

Table 2

Stem Item-Writing Guidelines

14. Directions in the stem are clear
15. The stem should contain the central idea, not the options
16. Avoid window dressing/excessive verbiage
17. Word the stem positively and avoid negatives, such as NOT

Note. Guidelines are from Haladyna et al. (2002).

The first stem item-writing guideline explains that test developers should devote time to writing the stem with as much clarity as possible. If the directions are unclear to examinees, they may have a difficult time answering the item. A majority of researchers cite that they are in favor of the use of this item-writing rule (82%), or do not cite the rule at all (Haladyna et al., 2002).

The second item-writing guideline states that the stem of a MC item, rather than the options, should contain the central idea. Researchers unanimously support this item-writing guideline when creating MC items (Haladyna et al., 2002). Consider the item Downing (2005) provides from a medical licensure exam in Figure 9. The stem is unfocused since the options contain the content that the test developers are trying to evaluate. Instead, the stem should pose a question to examinees, so they know what content is being asked. This would result in a more focused stem.

It is correct that:
A. Growth hormone induces production of IGFBP3
B. The predominant insulin-like growth factor binding protein (IGFBP) in human serum is
IGFBP3
C. Multiple forms of IGFBP are derived from a single gene
D. All of the above
E. Only A and B are correct

Figure 9. A flawed item containing an unfocused stem (Downing, 2005).

Although the second stem guideline emphasizes that the stem should not contain too little information, it is also possible to contain too much information in the stem. Specifically, the third item-writing guideline states that the stem of a MC item should not contain language that is irrelevant to an examinee when answering an item. Excessive language, such as a stem that is very long or material that does not match the content of the item, can make an item unnecessarily tricky (Haldayna et al, 2002). In Figure 10, the stem contains information that an examinee does not need to answer the item correctly. If the test developer is interested in determining whether an examinee can determine the sum of ‘X’ by knowing the mean and sample size, then being told the standard deviation and variance over complicates the item (Roberts, 1993). On the other hand, if the goal of the item was to test whether examinees can parse out which information is relevant (e.g., measure of central tendency versus variability), then the item stem would not contain irrelevant details. While there is no apparent disagreement toward this guideline, only half of test developers use this guideline when writing items (Haladyna et al., 2002).

A researcher collected some data on 15 students and the mean value was 30 and the standard deviation was 3. What is the sum of X if the variance is 9 and the median is 29?
A. 3 B. 30 C. 15 D. 450

Figure 10. A flawed MC item containing window dressing (Roberts, 1993).

Finally, the stem should be positively worded. Negatively worded items confuse examinees, and examinees become unsure of what the item is asking. This item-writing guideline is the most controversial, with 18% of researchers stating that it is not an issue to use negatively worded stems on exams (Haladyna et al., 2002). Figure 11 is an example of a negatively worded item (Downing, 2005). It has not been determined whether the incorrect response is a direct consequence of the negatively worded item, or if examinees answer the item incorrectly due to a lack of ability (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005).

Which of the following will NOT occur after therapeutic administration of chlorpheniramine?
A. Dry mouth
B. Sedation
C. Decrease in gastric acid production
D. Drowsiness
E. All of the above

Figure 11. A flawed MC item containing a negative stem (Downing, 2005).

While some might argue against the use of negatively worded stems, there is no conclusive evidence to suggest that psychometric properties of the item, such as item difficulty, are influenced by negatively worded stems (Rachor & Gray, 1996; Tamir, 1993). It may be useful to have a negatively-worded stem if test developers creating the items are interested in examinees having knowledge about when not to do something. However, it is suggested that these items should measure lower cognitive abilities (e.g., memorization of facts) rather than higher-order thinking (e.g., analyzing information; Maher, Barzegar, & Ghasempour, 2016).

Writing and Developing Options and Distractors

While the stem is an important part of the item, multiple item writing guidelines are about developing options. Haladyna et al. (2002) states that test developers should spend a majority of their time developing the correct answer and distractors. Table 3 provides 14 item-writing rules pertaining to both the writing of the correct option and the distractors (Haladyna et al., 2002).

Table 3

Guidelines for Writing Options for MC items

18. Write as many plausible distractors as possible
19. One right answer
20. Vary location of right answer
21. Logical/numerical order
22. Choices should not overlap
23. Choices homogeneous
24. Choice length equal
25. Use carefully none of the above (NOTA)
26. Avoid all of the above (AOTA)
27. Avoid NOT in choices
28. Avoid clues
29. Make distractors plausible
30. Use common errors of students
31. Use humor sparingly

Note. Guidelines are from Haladyna et al. (2002).

The first guideline for writing options is to create as many plausible distractors as possible. While a low percentage (4%) of test developers disagree with this guideline, a number of researchers have examined the impact of the number of distractors on item quality (e.g., Andres & del Castillo, 1990; Crehan & Haladyna, 1993; Haladyna & Downing, 1993; Kilgour & Tayyaba, 2016). Although it would seem best to decrease the chances of guessing by writing as many distractors as possible, the distractors must be plausible and not contain irrelevant information. This is because most examinees have enough test-wiseness to not select options that seem random (Frary, 1995). It can also be a challenge to develop plausible distractors. Rather than developing as many plausible distractors as possible, test developers argue against this technique because qualities of the item do not change when there are only three options (Haladyna & Downing, 1993; Rodriguez, 2005; Vyas & Supe, 2008).

It is also important to vary the location of the correct option. Attali and Bar-Hillel (2003) note that test developers typically place correct answers in the middle position of the options or show middle bias when creating an answer key. Doing so may cause items to become easier and less discriminating amongst examinees. Key balancing calls for test-makers to vary the position of correct options so that each choice is the correct option an equal number of times throughout the test (Bar-Hille, Budescu, & Attali, 2005). Test developers can also randomize the location of correct answers instead of using key balancing, as patterns in key balancing are decided beforehand. At a minimum, the correct option should not be placed as the same choice multiple times in a row. When this occurs, examinees are able to pick up on patterns. Randomization corrects this issue since patterns cannot be created beforehand (Bar-Hillel & Attali, 2002).

One option that cannot vary in location is all of the above (AOTA). AOTA is one of the two most controversial guidelines pertaining to option writing. While 22% of researchers say using AOTA is acceptable, 70% believe this option should never be used (Haladyna et al., 2002). Tarrant and Ware (2008) argue that using AOTA as a distractor can lead to examinees guessing the correct option more easily. For example, if an examinee is able to eliminate one option, then they are able to also able to eliminate AOTA. Additionally, having AOTA as an option only when it is the correct option cues examinees to pick the AOTA option, decreasing reliability of scores. As an alternative, using a constructed response format facilitates testing examinees’ knowledge on multiple details (Harasym, Leong, Violato, Brant, & Lorscheider, 1998).

The option none of the above (NOTA) is just as controversial as AOTA. Nearly half of researchers believe that NOTA should not be used when writing options for MC items (Frary, 1991). While there are some instances where NOTA can be used, such as when learning objectives call for examinees to distinguish when or when not to do something, items containing NOTA as an option can cause an item’s effectiveness to decrease (Frary, 1991). In particular, there are mixed results about how well using NOTA discriminates between examinees. Rich and Johanson (1990) found that the use of NOTA allows for items to discriminate among examinees better (Rich & Johanson, 1990), but others find that NOTA has no effect Crehan & Haladyna, 1991; Frary, 1991). After conducting a meta-analysis, Knowles and Welch (1992) concluded that using NOTA as an option does not significantly affect the quality of the item.

In comparison to NOTA, humorous options are not as controversial. When examining the use of humor in option development, most test developers (85%) do not mention using humor (Haladyna et al., 2002). Since humor does not have an effect on how well an examinee does, it can be implemented in certain situations (McMorris, Boothroyd, & Pietrangelo, 1997). If one of the options for the items in Figure 9 and Figure 10 included a pun or a nonsensical option, it might be acceptable in low-stakes, but not in high-stakes, standardized testing situations as these tests are associated with important decisions (McMorris, Boothroyd, & Pietrangelo, 1997; Brown & Itzig, 1976).

What is 2(3+6)?
a) 12 b) 18
c) 100 d) 15

Figure 12. Hypothetical MC item on a math test. Option B is the correct answer.

Writing good distractors lies in common misconceptions examinees have of the material (Haladyna et al., 2002). These can be typical errors an examinee makes when problem solving (Gierl et al., 2017). For example, consider the plausibility of the options for the item in Figure 12. The examinee who chooses option A would be making an error since they likely multiplied two and three and then added six. Examinees who have limited knowledge of the rules for order of operations might pick this option. While there was no disagreement for the use of student common errors as distractors, only 70% of authors specifically wrote about being in favor of using this technique to develop their own assessments (Haladyna et al., 2002).

There are two ways to obtain information about examinees' misconceptions: analyze responses to open-ended items or use options that are similar in content to the correct answer. By examining responses from open-ended items, where the examinee is only given the stem, test developers are able to determine where examinees make mistakes that lead to incorrect answers. In turn, the incorrect answers can be used to craft the distractors (Halloun & Hastenes, 1985). If obtaining information from open-ended items is not possible, distractors can be created using similar content to the correct answer. This leads examinees who are not knowledgeable on the topic to have a lower probability of selecting the correct answer (Ascalon, Meyers, Davis, & Smits, 2007). In Figure 13, options A, B, and C are steps that examinees would take to simplify the expression. Options B and C are examples of answers that examinees supplied on open-ended items. In comparison, option D is not similar in content to the other options as it has nothing to do with the item nor the other options. This is a violation of the guideline suggesting to write plausible distractors.

What would be the first step to simplify the following expression?

$$5(7-2)+10$$

a) Subtract what is in the parentheses b) The expression is already simplified
c) Multiply 5 and what is in the parentheses d) Find the derivative of the expression

Figure 13. Hypothetical MC item on a math test. Option A is the correct answer.

There is a significant decrease in the percentage of examinees who select the correct option when using technical phrases in distractors and nontechnical phrases in the correct option (Haladyna et al., 2002). This is likely due to students associating technical phrases with being the correct option, which is not related to an examinee's knowledge of the subject (Strang, 1980). When examining the item in Figure 13, option D is not only illogical as an answer for this item, but examinees who are being tested on their knowledge of order of operations would not be likely to know what a 'derivative' is. Even though option D is illogical, because it uses the technical term 'derivative', examinees may be erroneously drawn to it.

Measurement Paradigms: Classical Test Theory and Item Response Theory

The proper development of each part of the item is imperative in producing good MC items. However, simply using the guidelines stated above does not ensure a well-functioning item. Once items are given to examinees and data are collected, researchers are able to empirically investigate psychometric properties of the items. There are two measurement paradigms primarily used to analyze the functioning of items: classical test theory (CTT) and item response theory (IRT).

CTT uses an examinee's total score on a test as an estimate of ability, which is represented with the formula:

$$X = T + E. \tag{1}$$

Equation 1 is known as the true score model, where the observed score, X, is the sum of true score, T, and error score, E (De Ayala, 2013). The observed score is the score an examinee earns on a test. For example, if all items were scored as correct (1 point) or incorrect (0 points), then an examinee who answers 30 out of 40 items correct has an observed score of 30.

The true score is not a measure of truth. Instead, it can be thought of as an examinee's trait score. Theoretically, if an examinee were tested an infinite number of times, measurement error would average zero, and the average would be the examinee's true score of the ability being measured (Cronbach, 1990). Of course, this is not necessarily practical and only theoretical, which is why we never know an examinee's true score. It can only be estimated by the observed score (De Ayala, 2013).

Unfortunately, the observed score is also influenced by measurement error (De Ayala, 2013). When discussing measurement error in this context, researchers examine whether observed scores are reliable or consistent. To understand the reliability of scores, we need to first understand the decomposition of variance observed. The variance of the observed score is decomposed as:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \qquad (2)$$

where the observed score variance, σ_X^2, is the sum of the true score variance, σ_T^2, and the error score variance, σ_E^2. Observed scores from the instrument are considered more consistent as the amount of measurement error variance becomes smaller (σ_E^2). As measurement error variance decreases, the variance of the observed scores becomes closer to the variance of the true scores. The observed scores are then considered to be more reliable (Traub & Rowley, 1991).

There are certain advantages to using CTT when developing tests. For example, test developers can use CTT to estimate difficulty, discrimination, and ability when the sample size is small (Hambleton & Jones, 1991). In addition, CTT is a good paradigm for estimating scores when test developers would like to rank-order examinees and not generalize observed score beyond the classroom (De Champlain, 2010).

While CTT is appropriate in certain situations, it does have disadvantages. Item properties in CTT are sample dependent, meaning that a group of examinees taking a test may produce different item statistics than another sample taking the same test (De Champlain, 2010). Because of this, it is difficult to compare estimates across samples (Hambleton & Jones, 1991). Ideally, test characteristics would not depend on the sample (i.e., person-free), and the sample characteristics would not depend on the test (i.e., item-free measurement). Unfortunately, in CTT, test characteristics cannot be separated from sample characteristics. If scores are low, it is not always clear whether the test was hard or the sample had low ability. In comparison to CTT, IRT allows parameter estimates to be person-free and item-free (Embretson & Reise, 2000).

When using IRT, the goal is to use examinees' observed item responses to model the relationship between examinees' abilities and the probability of an examinee correctly answering an item (Harris, 1989). Ability, θ, is on a scale of -∞ to ∞, but typical values range from -3 to 3. Examinees with a high ability level, or high θ value, should have a higher probability (i.e., close to 1) of selecting the correct option (Hambleton & Jones, 1991). To model the probability of selecting a correct response on a MC item, test developers have the option of using the one-parameter logistic model (1-PL), the two-parameter logistic model (2-PL), or the three-parameter logistic model (3-PL).

The 1-PL model allows test developers to estimate a difficulty, b_i, for each item and it can be expressed with the following equation:

$$P_j(\theta) = \frac{e^{\left(\theta - b_j\right)}}{1 - e^{\left(\theta - b_j\right)}} \quad j = 1, 2, \ldots, n \tag{3}$$

where $P_j(\theta)$ is the probability of an examinee with an ability θ answering item j correctly on a test with n items. The difficulty parameter is on the same scale as ability, where typical values range from -3 to 3. A higher value indicates that the item is more difficult to examinees. Examinees with an ability level higher than the item difficulty have a greater than 50% chance of getting the item correct. Those with a lower ability level have a smaller chance of getting the item correct (Hambleton, Swaminathan, & Rogers, 1991).

We can graphically investigate the logistic function shown in Equation 3, which produces an S-shaped curve with the ability scale on the horizontal axis. The location of b_j on the ability scale is where an examinee at certain ability of θ has a 50% probability of selecting the correct response. As the b_j parameter increases, the item difficulty increases. Values of b_j that are below -2 are easy, and values greater than 2 are difficult (Harris, 1989). The location of b_j tells us where the item information is maximized (Hambleton et al., 1991).

Figure 14, provided by Hambleton et al. (1991), displays the item characteristic curves (ICCs) for four items from a 1-PL model. Item three has the lowest difficulty (b_3= -1.0) while item two has the highest difficulty (b_2=2.0). So, examinees with an ability level of -1.0 have a probability of 0.5 of selecting the correct answer on item three, but significantly less than a 50% chance of selecting the correct answer on item two. Harris (1989) explains that the 1-PL is advantageous compared to the 2-PL and 3-PL models because a total score can be used to estimate θ, and the number of examinees selecting the correct option can be used to estimate b_j. The total score estimated in the 1-PL model is a sufficient statistic for ability.

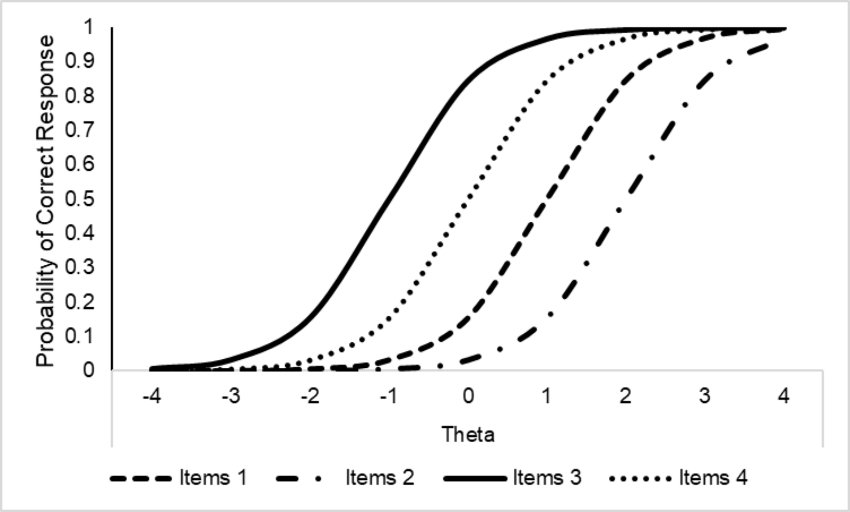

Figure 14. Four item characteristic curves using the 1-PL model (Hambleton, Swaminathan, & Rogers, 1991, p. 14).

The two-parameter logistic model (2-PL) is similar to the 1-PL model, except for the addition of a parameter for discrimination (a_j). The 2-PL model formula is expressed as:

$$P_j(\theta) = \frac{e^{Da_j(\theta - b_j)}}{1 - e^{Da_j(\theta - b_j)}} \quad j = 1, 2, \ldots, n \qquad (4)$$

where the a_j parameter is the slope for each of the ICCs. An ICC with a steep slope indicates a more discriminating item than an ICC with a flat slope. The value of a_j ranges from 0 to ∞, but is typically between 0 and 2. Compared to Equation 3, in Equation 4 we see the addition of a_j and *D*. The *D* is a scaling factor that allows for the logistic function to be similar to a normal ogive function, fixed to 1.7.

Item discrimination can be examined graphically by looking at the ICC’s slope. In Figure 15, the item that is most discriminating is item three because it has the steepest slope (a_3=1.5). Item two has the lowest slope (i.e., closest to a horizontal line; $a_i = 0$), and thus the lowest discrimination (a_2=0.5). The discrimination parameter for item three would be interpreted as having the highest utility for separating examinees at an ability level of $b_j = -1$ with the inclusion of the discrimination parameter (Hambleton & Jones, 1993).

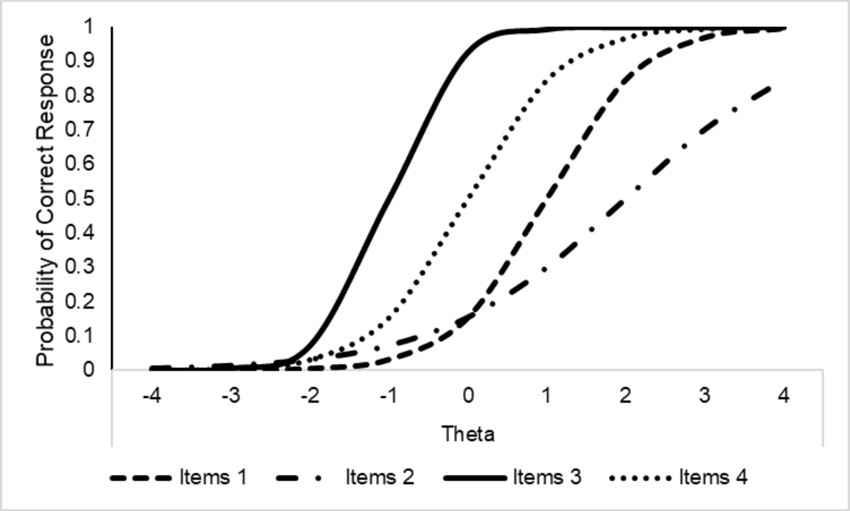

Figure 15. Four item characteristic curves using the 2-PL model (Hambleton, Swaminathan, & Rogers, 1991, p. 16).

The three-parameter logistic model (3-PL) introduces a third item parameter, the pseudo-guessing parameter. It is expressed as:

$$P_j(\theta) = c_j + (1 - c_j)\frac{e^{Da_j(\theta - b_j)}}{1 - e^{Da_j(\theta - b_j)}} \quad j = 1, 2, \dots, n \qquad (5)$$

The only addition is the c_j parameter, or the pseudo-guessing parameter, which is a lower asymptote for the ICCs. This lower asymptote is the probability of low ability examinees answering an item correctly. Because the lower asymptote is no longer 0, the interpretation of the difficulty parameter is adjusted for guessing (Hambleton et al., 1991).

In Figure 16, there is a different lower asymptote for each item (c_1=0.19, c_2=0.17, c_3=0.07, c_4=0.04; Hambleton & Jones, 1993, p. 41). 19% of low ability examinees select the correct answer, which could be due to guessing. We do not necessarily know whether examinees are truly guessing or guessing after eliminating options. The smallest chance of guessing is for low ability examinees on item four.

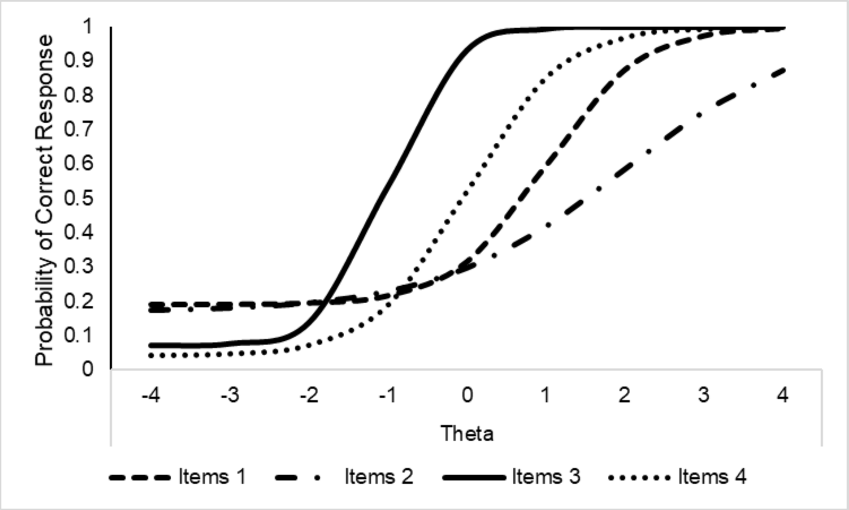

Figure 16. Four item characteristic curves using the 3-PL model (Hambleton & Jones, 1993, p. 41).

To use the IRT paradigm, the data should meet certain assumptions. First, there should only be one ability measured by the items. This is known as unidimensionality. Responses may be influenced by other traits (e.g., test anxiety or motivation), but the primary trait is what influences examinees to respond to items on the test (Hambleton, Swaminathan, & Rogers, 1991).

A second assumption for unidimensional IRT models is local independence. To satisfy this assumption, examinees' responses to each item must be statistically independent of responses to other items, once ability is accounted for. Local independence is that the "probability of a response pattern on a set of items is equal to the product of the probabilities associated with the examinee's responses to the individual items" (Hambleton et al., 1991, p. 11). Once the ability is accounted for, responses to items should be uncorrelated since the responses do not have any other shared variance (Hambleton et al., 1991).

An advantage to using IRT is that items possess the property of invariance. This means that item properties are not sample dependent. In other words, up to a linear transformation, item parameters are not dependent on the ability distribution for the examinees used when collecting responses. After scaling to a common metric, the ICCs will be the same for any group of examinees as long as the model fits the data (Hambleton et al., 1991). Because of this, IRT is especially useful for test developers estimating ability levels of examinees and tracking the scores across time (De Champlain, 2010).

A disadvantage of IRT is that it requires the use of large sample sizes, which is not always feasible in a classroom environment. Although sample size is dependent on the type of model (e.g., 1-PL, 2-PL, 3-PL) a test developer uses, a sample size of 500 or more is appropriate (Hambleton & Jones, 1991). IRT can also be challenging for those not in the measurement field. The theoretical concepts behind IRT, and the difficulty in using some programs associated with IRT can hinder test developers use of this paradigm (Thorpe & Favia, 2012).

CTT and IRT both have advantages and disadvantages. In comparison to each other, CTT and IRT also have major similarities and differences that need to be addressed to understand which is better for test developers to use.

Comparison of Measurement Paradigms

Table 4 contains two examinees' responses to ten items on a test. For simplicity, the items become increasingly difficult from one to ten. Both examinees have the same total score, but they answered different items correctly. Examinee A answered the 6 easiest items correctly while Examinee B answered a mix of easy and hard items correctly.

Table 4

Two Examinees' Item Responses

Item	1	2	3	4	5	6	7	8	9	10
Examinee A	1	1	1	1	1	1	0	0	0	0
Examinee B	1	1	1	1	1	0	0	1	0	0

Note. 0 – incorrect; 1 – correct

Under the CTT framework, Examinee A and Examinee B have the same estimated true score (i.e., their total score). The CTT total scores are equivalent because both examinees obtain a total score of six. The ability estimates in the 1-PL model will also be equivalent since the total score is a sufficient statistic for ability in the 1-PL model. Under the 2-PL and 3-PL IRT models, however, Examinee A and Examinee B will have different estimated ability levels for the latent construct that this test is measuring. CTT and IRT may produce differing interpretations of ability levels, and both paradigms have differences that explain why this occurs.

Table 5 lists four differences between CTT and IRT pertaining to the levels of analysis, assumptions, item-ability relationship, and the way ability is defined. In CTT, the model focuses on true scores at the test-level while IRT models focus at the item-level (Hambleton & Jones, 1991). In contrast to the clear specification of the item-ability relationship in IRT, CTT does not specify this relationship (Hambleton & Jones, 1991). While CTT only looks at the test level information, such as total scores, IRT looks at information at the item level. When developing tests, those using IRT rely on an ICC to see the relationship between ability levels and the probability of selecting the correct option for an item. This relationship is not specified in CTT (Hambleton et al., 1991).

Table 5

Major Differences Between CTT and IRT

Area	CTT	IRT
Levels of analysis	Test	Item
Item-ability relationships	Not specified	Item characteristic functions
Assumptions	Weak	Strong
Ability	Estimated true scores reported on test-score scale	Ability scores reported on scale -∞ to ∞

Note. From Hambleton and Jones (1991, p. 43).

Another difference is that assumptions in CTT are considered weak compared to strong assumptions in IRT. Assumptions that are weak tend to be easier to meet while strong assumptions are more challenging to meet (Hambleton & Jones, 1991). The assumptions of unidimensionality and local independence in IRT are more complex, but test developers who use IRT also focus on assumptions more than those using CTT (Osterlind & Wang, 2018).

Another difference is that ability estimates are in different metrics when comparing CTT and IRT. In CTT, an estimate of an examinee's true score is the total score and is based on the metric that the test is on. In comparison, test developers using IRT report ability on a scale of -∞ to ∞ rather than a total score (Hambleton et al., 1991).

Due to advances in the measurement field, a distinction between the old (i.e., CTT) and new (i.e., IRT) rules is made. Embretson and Reise (2000) compare CTT and IRT by explaining the differences between ten rules of measurement. Table 6 presents the ten rules. Rules one, two, and four are explained and contrasted in detail for each paradigm due to their relevancy in this study.

Table 6

Rules of Measurement

1. Standard error of measurement
2. Test length and reliability
3. Interchangeable test forms

Table 6 (continued)
4. Unbiased assessment of item properties
5. Establishing meaningful scale scores
6. Establishing scale properties
7. Mixing item formats
8. Meaning of change scores
9. Factor analysis of binary items
10. Importance of item stimulus features

Note. Rules of measurement from Embretson & Reise (2000).

The first rule pertains to the standard error of measurement (SEM). The SEM is defined as variation in scores that is due to error on a test (Crocker & Algina, 1986). As the SEM increases, test developers become less certain about whether the test is measuring the construct (Magno, 2009). In CTT, the SEM applies to all scores for a population (Embretson & Reise, 200). In other words, the SEM is constant across all examinee test scores (Hambleton & Jones, 1993). In contrast, the SEM differs across scores but is averaged across populations in IRT (Embreston & Reise, 2000). Each ability level has an estimated SEM value (Embretson, 1996).

Another rule states that the relationship between test length and reliability differs for CTT and IRT. It is important to note that SEM estimates are related to reliability estimates. As reliability increases, the SEM decreases (Crocker & Algina, 1986). In CTT, a test that is longer is typically more reliable (Embretson & Reise, 2000). In comparison to CTT, a test with fewer items can sometimes produce more reliable estimates under the IRT paradigm. Whereas in CTT there is one SEM for scores on the entire test, in IRT we estimate reliabilities that are not constant across ability. We refer to this as conditional reliability because reliability is a function of both ability and item parameters. Recall that ability and difficulty have the same metric. When item difficulties are around the same value, and discrimination is high, marginal reliability is large because we have the most information about a specific ability. The inverse of information is SEM, meaning that the smallest SEM is where we have the most information and largest reliability.

The unbiased assessment of item properties differs in CTT and IRT, depending upon the type of sampling. In CTT, to determine if items properties are unbiased, researchers must use a representative sample (Embretson & Reise, 2000). If a researcher sampled examinees with low abilities the first time but high abilities the second time, the second group of examinees would answer the items correctly more often than the low ability examinees. Because of the biased sample, researchers cannot compare the item properties since the groups are very different (Embretson, 1996). When examining item properties in IRT, samples that are unrepresentative, to some degree, can be used (Embretson & Reise, 2000). This is because IRT has the assumption of invariance. When the model fits the data, researchers can infer that different samples will still produce unbiased item properties (Nguyen, Han, Kim, & Chan, 2014).

While CTT and IRT have differences, there are similarities. CTT and IRT paradigms address issues in measurement (Hambleton & Jones, 1993) and both are used to create and evaluate examinations (De Champlain, 2010). Both CTT and IRT assume that the underlying, latent ability of the test is continuous (De Ayala, 2013). While each can help evaluate the psychometric properties of the scores, they do so in different ways. Despite limitations already spoken to (i.e., sample size considerations), test developers may use either paradigm to analyze how items perform, known as an item analysis.

Item Analysis

Once a test is created, psychometric properties of the scores and items can be calculated to see if the test is functioning well. In an item analysis under the CTT paradigm, researchers can investigate item difficulty. Item difficulty, P, is the proportion of examinees who answer the item correctly, which is calculated as:

$$P = \frac{\#\ correct\ responses}{Total\ \#\ of\ responses} \tag{7}$$

If, for example, there is a total of 50 examinees and 35 examinees selected the correct response, the item difficulty would be .7 ($P = \frac{35}{50} = .7$). Item difficulty in this example can also be thought of as 70% of examinees selecting the correct option.

Item difficulty ranges from 0 to 1, where 0 means that no examinees answered the item correctly, and 1 means that all examinees answered the item correctly. Difficulties less than .30 indicate a very hard item and difficulties above .7 indicate an easy item (Bandalos, 2018). Items that do not fall into this range require revision (Ding & Beichner, 2009). In cases where p < .3, examinees who actually have high ability may answer the item incorrectly due to confusion in the wording of the item, and not necessarily as a result of the item content being difficult (Boland, Lester, & Williams, 2010).

As an alternative to CTT, test developers can estimate item difficulty using IRT models. The difficulty parameter, b_j, is on the same metric as ability, which can range from -∞ to ∞. However, values are typically between -3 and 3. Higher values of difficulty indicate a more difficult item (Harris, 1989). The interpretation of item difficulty changes slightly in the 3-PL model to include the pseudo-guessing parameter.

Test developers can also examine item discrimination (*D*) in CTT. Item discrimination refers to how well an item distinguishes between low and high ability examinees. Since it is inferred that high ability examinees will answer more items correctly, and low ability examinees will answer fewer items correctly, item discrimination is an index for quantifying the discrimination between examinees with specific ability levels (Bandalos, 2018).

In CTT, item discrimination values range from -1 to 1. A value of -1 indicates that all low ability examinees answered the item correctly and no high ability examinees did, whereas a value of 1 means all high ability examinees answered the item correctly and no low ability examinees did. Towns (2014) explains that item discrimination is preferred to be above .4, meaning that the item can discriminate between low and high ability examinees. Item with values between 0 to .2 should either be revised or discarded. Negative discrimination values are indicative of items that are either flawed or keyed incorrectly.

One index of discrimination is equal to the proportion of high ability examinees who select the correct option minus the proportion of examinees with low ability (i.e., lower 27th percentile of total score) who select the correct option (De Ayala, 2013).

$$D = P_{high} - P_{low} \tag{8}$$

The discrimination would be positive if a higher proportion of high ability examinees select the correct answer (Bandalos, 2018; Towns, 2014). Table 7 displays examinees from both the low and high ability groups and number of examinees who chose each option for this example item.

Table 7

Table of responses for each option from low and high ability examinees

	A*	B	C	D	Total
High Ability	41	2	6	1	50
Low Ability	20	7	17	6	50

Note. Option A is correct (*).

To calculate item discrimination, we take the proportion of high ability examinees who selected the correct option and subtract it from the proportion of low ability examinees:

$$D = \frac{41}{50} - \frac{20}{50} = .82 - .40 = .42$$

This item is considered to have high discrimination amongst examinees because the value is positive and above .4.

In CTT, item discrimination can also be examined by calculating the point biserial coefficient. This coefficient is the correlation between item and total scores and produces the reliability of scores for each item (Ding & Beichner, 2009). The point biserial coefficient is calculated as:

$$r_{pbi} = \frac{(\frac{M_1 - M_0}{s})}{[((n_1 * n_0) * .5)/n^2]} \tag{9}$$

where, M_1 is the mean of examinees who answer the item correctly, M_0 is mean of examinees who answer the item incorrectly, s is the standard deviation of all scores, n_1 is the number of examinees who answer the item correctly, n_0 is the number of examinees who answer the item incorrectly, and n is the total number of examinees. Similar to *D* values, point biserial coefficients range from -1 to 1, where a negative value indicates that examinees possessing low total scores tend to answer the item correctly and those with high scores answer incorrectly. Alternatively, a positive value indicates that examinees with high total scores tend to answer the item correctly and examinees with low total scores answering the item incorrectly, the ideal situation. The point-biserial correlation is equivalent to the Pearson correlation, where an examinees' total score is correlated with selection of the correct option.

IRT provides a way to estimate item discrimination through the use of the a_j parameter. In the 1-PL, 2-PL, and 3-PL models, item discrimination is the slope of the ICC for each item. In the 1-PL model, a_j is fixed at a specific value. In most cases, the discrimination parameter is fixed to have a slope of one, which is known as the Rasch model (Rasch, 1960), as presented in Equation 3 (Harris, 1989). Test developers can also use the 2-PL model to estimate the discrimination parameter for each item. Item discrimination values typically fall between 0 and 2, but values can range from 0 to ∞. As an ICCs slope becomes steeper, discrimination values increase (Hambleton et al., 1991). The addition of the pseudo-guessing parameter in the 3-PL model does not alter the interpretation of the discrimination parameter (Hambleton & Jones, 1993).

While test developers can gain important information about tests by examining item difficulty and discrimination, analyzing the functioning of distractors provides test developers with a more complete understanding of item functioning. Instead of solely looking at dichotomous responses (i.e., correct or incorrect), test developers can investigate the relationship between ability groups and their propensity to select a certain option (Gierl et al., 2017). Polytomous IRT models allow for this observation of properties of item responses.

Distractor analysis

Regardless of the paradigm, test developers examine multiple sources of evidence to conclude whether distractors are functioning as intended. In CTT, a simple way to accomplish this is to calculate the frequency of examinees picking each option. A distractor is considered low-functioning when it is chosen by less than 5% of examinees (Gierl et al., 2017; Haladyna & Downing, 1993). In Table 8, the percentage of examinees who selected each option for an item is displayed. Based on Haladyna and Downing’s (1993) rule, option D is considered a low-functioning distractor because less than 5% of examinees selected it. Options B and C, however, are well-functioning distractors as they fall above this cut-off. Although it is useful to see how many respondents are selecting each distractor, it may be more useful to see the type of examinee who is selecting each distractor.

Table 8
Percentage of examinees selecting each option

Item 6				
	A*	B	C	D
Total	82.56	8.36	7.05	1.67

Note. Option A is the correct option (*).

To gain a better insight of who is selecting distractors, test developers can separate examinees into low, medium, and high groups based on their total scores. The proportion of examinees who select each option for the item is examined (Gierl et al., 2017). Table 9 shows a frequency table with the percentages of examinees selecting certain options based on their ability levels (low, medium, and high). This is a version of Table 8 that is broken down by ability level. Less than 5% of examinees selected option D. Because of this, we would consider option D to be a poorly functioning distractor. The low percentage of examinees who are choosing the distractor for this item are likely to be guessing (Haladyna & Downing, 2013). Although options B and C contain percentages in the high group that are below 5%, these distractors are still functioning well since those with low and medium abilities are selecting them somewhat frequently. By breaking down the percentages of examinees who select each option by their ability level, it is easy to see which examinees are selecting certain options more often.

Table 9

Percentage of low, medium, and high ability levels selecting certain options

Item 6				
	A*	B	C	D
Low	65.59	18.74	13.15	1.98
Medium	84.52	6.73	6.13	2.21
High	93.82	2.11	3.32	0.6

Note. Option A is the correct option (*). Table 8 is a frequency table for Figure 9.

Another way to examine distractor functioning in the CTT paradigm is to create trace plots for each item. Trace plots provide a visual representation of how the options are functioning in relation to total score (Wainer, 1989). Examinees are grouped by total score on the x-axis. The frequency of examinees selecting each option is displayed on the y-axis. A unique line is shown for each distractor (see Figure 17).

Based on the appearance of distractor trends in the trace plots and frequency tables, Haladyna et al. (1993) categorize four types of distractors as undesirable. They use the criteria that the lines of the trace plots should be monotonically decreasing for distractor options. Four types of undesirable distractors are: (1) a distractor where less than 5% of examinees select it, (2) a distractor with a completely flat trace line, (3) a distractor with a non-monotonic trace line, and (4) a distractor that is acting as the correct answer. First, as stated previously, the frequency of selecting a distractor should be greater than 5% to be considered functioning. Second, when a trace line is completely flat (i.e., slope is zero) for all levels of ability, the distractor is considered undesirable. This is because it is not discriminating high and low ability examinees.

Third, a trace line that is non-monotonic is not desirable. When examining a trace plot, this type of line will have a slope that is not consistent across the ability continuum. A non-monotonic slope is undesirable since it tells us little about the examinees. Instead, it is contradictory to the way we want a distractor to function (Haladyna et al., 1993).

Finally, a distractor that has a monotonically increasing trace line indicates a pattern similar to the correct option. If examinees with high ability select a distractor more frequently than those with low ability, there is an issue with the options or instructions. Examinees with a high ability should select the correct option and not a distractor (Haladyna et al., 1993).

Figure 17 is a trace plot for Table 8. On the horizontal axis of the graph, ability is quantified as total score, where examinees are placed into low, medium, and high ability groups. The vertical axis displays the percentage of examinees in each group who chose a specific option. Option A is the correct option, and it follows a trend we would expect. From low to medium to high total scores, the percentage of examinees selecting the correct option increases. Option C in Figure 9 is especially problematic because the trace line is flat (i.e., nondiscriminating) and less than 5% of examinees select it (i.e., nonfunctioning).

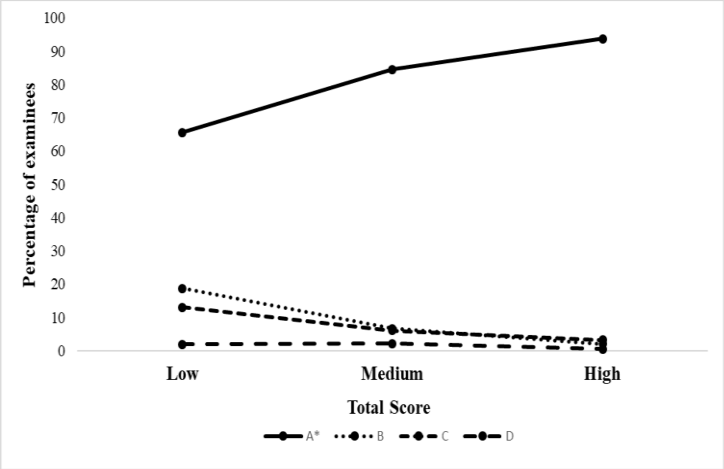

Figure 17. Trace plot of the frequencies in Table 8. Option C is the undesirable distractor. Option A is the correct option.

Often times, the flatness of the lines is not clear. To test the slope objectively, a chi-square goodness-of-fit test can be used to see whether the slope of the trace line for a suspected nonfunctioning distractor is significantly different from zero (Haladyna et al., 1993). This statistic tests a null hypothesis of equal proportions of selection among the ability groups in the population. When a distractor displays a flat trace line, the same percentages of examinees in each ability group are selecting the distractor.

Once distractors are identified as problematic, test developers may delete the distractor for future revisions of the test if they want to increase discrimination power between examinees with low and high abilities (Haladyna & Rodriguez, 2013). However, the distractor should not simply be removed based upon poor functioning. First, the distractor should be reviewed by subject-matter experts and examined for content before removing it (Gierl et al., 2017; Haladyna et al., 1993).

Test developers may alternatively use IRT to analyze distractor functioning. IRT allows for simultaneous analysis of distractor functioning and estimating ability (Gierl et al., 2017). However, using the 1-PL, 2-PL, or 3-PL IRT model to perform a distractor analysis is not possible. These models only provide information for examinees who select the correct versus the incorrect options. Examinees who select any of the distractors are collapsed into a single, “incorrect” category. Instead of modeling dichotomous item responses, polytomous IRT models consider all response options as unique categories.

The nominal response model (NRM; Bock, 1972) is amongst the most commonly used polytomous IRT models to examine distractor functioning. The NRM was originally created to examine information about distractors by modeling responses from MC tests (Bock, 1972; Stone & Zhu, 2015). This polytomous model allows test developers to examine the probability of examinees selecting each of the options in a MC item (Gierl et al., 2017). The NRM can be expressed as the probability of an examinee with ability θ selecting option *k* of item *j*:

$$P_{jk}(\theta) = \frac{e^{zjk}}{\sum_{x=0}^{m_j} e^{zjx}} \tag{10}$$

where $z_{jk} = a_{jk}\theta + c_{jk}$. The option discrimination parameter, a_{jk}, represents the strength and direction of the relationship between the propensity to select an option and θ. The option extremity parameter, c_{jk}, represents the probability for an examinee to select option *k* when θ = 0. The intersection of adjacent categories is notated as b_{jk}, where $b_{jk} = \frac{c_{jk}-c_{jk+1}}{a_{jk+1}-a_{jk}}$ (Stone & Zhu, 2015; Thissen, Cai, Bock, Nering, & Ostini, 2010).

Thissen and Steinberg (1988) provide the item characteristics choice curves (ICCCs) for a MC item with four options in Figure 18. Option A corresponds with the ICCC labeled 0, option B corresponds with option 1, option C corresponds with option 2, and option D corresponds with option 3. Option 3 is the correct answer for this item. For option 3, $a_3 = 2.7$ and $c_3 = 0.7$. The large a_3 parameter estimate indicates that option 3 discriminates well between low and high ability examinees, where high ability examinees have a higher probability of selecting it.

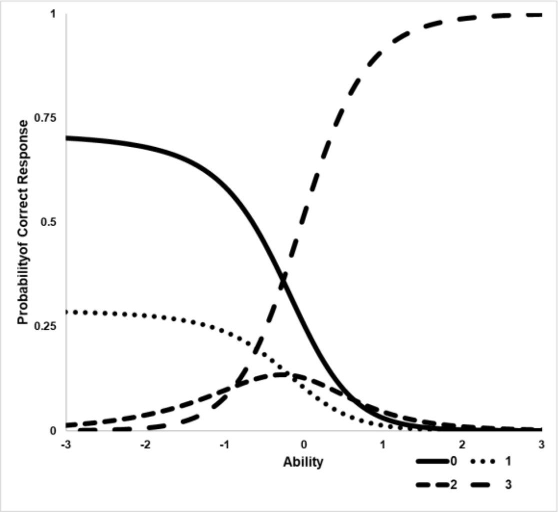

Figure 18. ICCCs for an item with four options (Thissen & Steinberg, 1988).

Option 0 is a well-functioning distractor since those with a low ability have a high probability of selecting it, and the probability decreases as ability (θ) increases. Option 0 and option 1 have the same slope, meaning they have the same a_{jk} parameter estimate (i.e., the slopes of both lines are proportional). Low ability examinees tend to select either distractor, but the probability for selecting either option is different. We can examine the c_{jk} to examine the difference in probabilities. Option 0 has a c_0 parameter estimate of 0 while option 1 has a c_1 has a parameter estimate of -.9. Because the c_1 parameter estimate is lower than the c_0 parameter estimate, option 1 has a lower probability of being selected. The ICCC of option 0 is always greater than ICCC of option 1 due to unequal c_{jk} values.

The NRM may be appropriate to analyze distractors when the 1-PL or 2-PL are used to score responses. However, it is less appropriate when the 3-PL is used to score the responses as the NRM does not account for guessing. In the NRM, the probability of not selecting the correct answer tends to approach one as ability decreases. This is likely not true, however, as examinees with low ability are able to guess the correct answer (Gierl et al., 2017). Due to this limitation, Samejima (1979) extended the NRM to account for guessing – conceptually similar to how the 3-PL IRT model extends the 2-PL to account for guessing. Under Samejima's (1979) multiple-choice model (SMCM) , the probability of selecting option k, $k = 1, \ldots, m_j$ on item j is,

$$P_j(k) = \frac{\exp(z_k) + d_k \exp(z_0)}{\sum_{h=0}^{m_j} \exp(z_h)} \tag{11}$$

where parameters a_{jk} and c_{jk} are interpreted in a similar way as the NRM, and d_{jk} is a fixed guessing parameter across items. Under the SMCM, A, B, C, and D are treated as categories 1-4. Now, a latent category, 0, is added to incorporate guessing. The d_{jk} is a "don't know" parameter for each category, 1-4. This parameter is defined as the proportion of examinees who select an option (k) because they, theoretically, don't know the correct answer. A DK ICCC expresses the probability of guessing by an examinee with given ability. In Samejima's (1979) model, the d_{jk} parameter is fixed to be $1/m_j$, where m_j is the number of response categories m, for item j.

The guessing parameter, however, is fixed for each item when using this model. The fixed d_{jk} parameter in Samejima's (1979) model has been found to be more restrictive for all applications when modeling MC response data (Thissen & Steinberg, 1984; Thissen, Steinberg, & Fitzpatrick, 1989). Thissen and Steinberg (1984) developed a multiple-choice model (MCM) where the parameter for guessing can be estimated for each item. Estimates of parameters of a_k and c_k undergo reparameterization because guessing is likely not fixed for each item.

The NRM and SMCM both have strict assumptions that are typically violated when guessing occurs. First, the NRM assumes that as ability decreases, the probability of selecting the correct answer tends to approach zero. Because examinees are able to guess, this is likely untrue as there is still a likelihood of guessing the correct answer. While the SMCM model includes guessing, the guessing parameter is fixed to $^1/_{m_j}$, where m_j is the number of choices to guess from. Thissen and Steinberg (1984) extended both models to create the Thissen-Steinberg Multiple-Choice Model (TSMCM), allowing the guessing parameter for each option to vary.

The TSMCM is mathematically defined in the same way as Samejima's (1979) extension of the NRM (Equation 10) except that the d_{jk} can be freely estimated under specific constraints. Since d_k represents the proportion of guessing each category, d_{jk}

falls in the interval $[0,1]$ and $\Sigma_k d_k = 1$. The d_{jk} parameter has the same interpretation as in Samejima's (1979) model. The TSMCM is a generalized model in the sense that when d_{jk} is fixed to $1/m_j$ for all options, the model simplifies to Samejima's (1979) model. Furthermore, when a_o and c_o are fixed to 0, the d_{jk} becomes trivial and the TSMCM essentially simplifies to the NRM (Thissen & Steinberg, 1984).

The addition of the d_{jk} affects the interpretation of the parameter values a_{jk} and c_{jk}. Parameter a_{jk} can have negative and positive values, which reflect the propensity to select that option k for a given ability θ. For example, consider a positive a_{jk} value. Examinees with a higher ability ($\theta > 0$) tend to have a higher propensity of selecting option k. This means that the higher the a_{jk} parameter estimate, the steeper the slope of an ICCC for option k. The c_{jk} parameter is known as the intercept parameter. The intersection used for the NRM is no longer applicable because of the addition of the d_{jk} parameter. The inclusion of d_{jk} parameter changes the location where adjacent ICCCs intersect because the d_{jk} parameter has its own ICCC for the DK category (Thissen et al., 1989).

The ICCCs for a four option MC item are shown in Figure 19. Option D is the correct answer, and it has an a_D estimate of 1.04, c_D estimate of 1.69, and d_D estimate of 0.41. The high, positive value of the slope indicates that as ability increases, examinees are more likely to select this option. However, option A also appears to be attracting high ability examinees. Option A has an a_A estimate of 1.03. The similarity in the values of the slopes indicate that high ability examinees have a proportional propensity for selecting either option. Although both options have similar slope values, the parameter estimate of c_A is equal to .02. This value is much smaller than that of option D, indicating that high ability examinees have a higher probability of selecting option D than option A. Option D has a guessing parameter estimate of $d_D = 0.41$ while option A has an estimate of $d_A = 0.17$. Option D has a higher value than option A, which is seen by comparing the asymptotic tendencies of the two ICCCs as ability decreases in Figure 19.

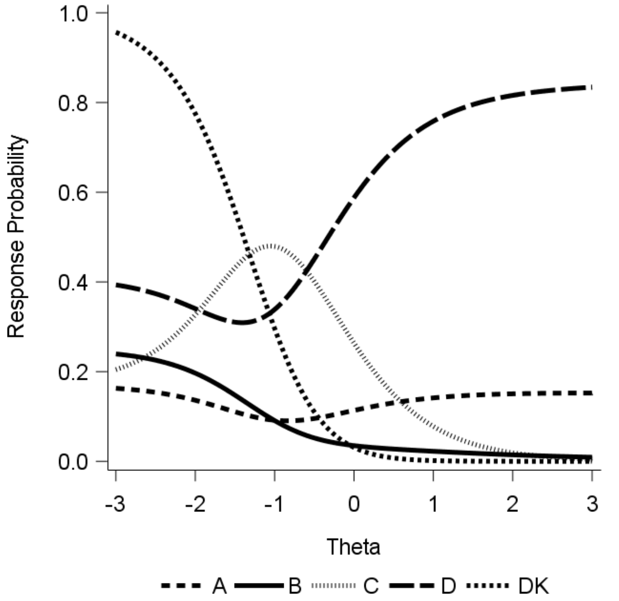

Figure 19. ICCCs for an item with four options. The correct option is D. DK=don't know (Thissen, Steinberg, & Fitzpatrick, 1989).

The modeling of examinee item responses with polytomous IRT models provides greater information about the functioning of items in comparison to dichotomous IRT models (Thissen et al., 1989). In addition, the NRM has been found to improve examinee ability estimates in comparison to the 2-PL model, especially for low ability examinees (Bock, 1972; Thissen, 1976). Levine and Drasgow (1983) showed that distractors can provide accurate information about an examinee's ability since distractors provide diagnostic information. For example, the distractor an examinee selects might be a common misconception (Briggs et al., 2006). Because dichotomous IRT models may only provide partial information about examinee knowledge (DeAyala, 1989), there is evidence that polytomous IRT models should be used more often when estimating an examinee’s ability because of the information that distractors can provide.

Because there is evidence to support that distractors are an important part of the MC item (Thissen et al., 1989), they should be examined more closely. Samejima (1988) named types of desirable distractors, which had positive influences on item properties. For example, informative distractors are able to discriminate examinees based on their ability level. The number of examinees who select the correct answer tends to be low (i.e., high difficulty) because the distractor draws in examinees who are not as high ability. Although, for example, the informative distractor can cause high discrimination, there is no evidence of how it affects ability estimates.

Because properly functioning distractors are useful when analyzing items by providing more accurate estimates of ability, the effects of distractors that are not functioning well should be examined to see how they impact ability estimates. It is unknown whether ability estimates may be biased if multiple items on a test contain undesirable distractor properties. In all testing situations, scores are important to examinees. If distractors are not functioning as they are supposed to, researchers should be aware of the consequences they have on ability estimates.

Chapter 3. Methods

Research Questions

In the current study, I begin by examining the capabilities of the IRT procedure in SAS 9.4 (TS1M4) before examining the effects of undesirable distractors on estimates of ability. Without the accurate recovery of parameters and neglecting to report possible issues, interpretations of the second, third, and fourth research questions are not valid:

1. How well does the SAS 9.4 IRT procedure recover estimates of ability using the 3-PL model?
2. What are the effects of the percentage of items containing undesirable distractors on estimates of ability?
3. Does the length of the test have an effect on an accuracy of ability estimates when items contain undesirable distractors?
4. Is one type of undesirable distractor more prone to producing inaccurate estimates of ability?

Simulation Studies

The interest of this study is to determine the effects that undesirable distractors have on estimates of ability. This main research question is challenging to answer in an empirical study because obtaining test-takers' true ability is not possible. Additionally, it is unknown what the effects of undesirable distractors are on estimates of ability. There is the possibility to negatively impact test-takers in a high-stakes testing situation, which can lead to detrimental consequences to their futures. It is also not possible to manipulate distractors to be poorly functioning. Distractors can only be determined as poorly functioning after data collection and analysis. For these reasons, a simulation study is appropriate to answer the research questions (Beaujean, 2018; Harwell, Stone, Hsu, & Kirisci, 1996; Luecht & Ackerman, 2018).

In a simulation study, item responses are generated using true ability values instead of collecting responses (Feinberg & Rubright, 2016). An important consideration when performing a simulation study is the generalizability of findings to real data (Feinberg & Rubright, 2016; Luecht & Ackerman, 2018). Specific reasoning is imperative when justifying the use of certain decisions (e.g., sample size, test length, etc.). The following sections explain reasoning for the use of conditions in the current study.

Varying Conditions. Table 10 displays the simulation conditions for three factors: type of undesirable distractor, length of test, and percentage of items containing undesirable distractors. Each simulated test contained either 30, 50, or 100 items with a specific type of undesirable distractor in either 10%, 30%, or 50% of the items. When deciding on the number of items to use, it is appropriate to use at least 25 test items to obtain good correlations with values of true ability and estimated ability (Reise & Yu, 1990). I selected other test lengths (i.e., 50 and 100 items) due to the interest of generalizing results to tests in both the classroom and standardized testing settings. I calculated the percentage of items containing undesirable distractors for each condition of test length, so each condition had a whole number of items. Crossing each factor (4x3x3) produced 36 conditions.

Table 10
Simulation conditions

Undesirable distractor	**Test length**	**Frequency**
Implausible	30	10% / 30% / 50%
	50	10% / 30% / 50%
	100	10% / 30% / 50%
Equivalent	30	10% / 30% / 50%
	50	10% / 30% / 50%
	100	10% / 30% / 50%
Upper lure	30	10% / 30% / 50%
	50	10% / 30% / 50%
	100	10% / 30% / 50%
Lower lure	30	10% / 30% / 50%
	50	10% / 30% / 50%
	100	10% / 30% / 50%

I generalized the types of undesirable distractors to the TSMCM based upon Haladyna and Downing's (1993) undesirable distractor trace lines. There are three types of undesirable distractors: *implausible distractors*, *equivalent distractors*, and *lure distractors*. There are two distinct types of lure distractors: upper lure and lower lure. The non-monotonic trace lines discussed by Haladyna and Downing (1993) were not simulated for this study because there is not complete agreement among researchers that this type of distractor is undesirable. For example, Thissen et al. (1989) discuss that non-monotonic trace lines can provide useful information about examinees in the middle of the ability continuum when using polytomous models to analyze data.

Implausible distractors occur when there is a low frequency of responses, no matter the ability level (Haladyna & Downing, 1993). Because I use the TSMCM in the current study, which incorporates guessing, the implausible distractor has a higher probability of selection for low ability examinees, but the item characteristic choice curve (ICCC) returns to less than 5% as ability increases. Table 11 provides the example item parameter values and Figure 20 provides the ICCCs that lead to an implausible distractor (choice B).

Table 11
Example of item parameters for the implausible distractor condition

a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
2.60	1.33	0.33	1.22	-1.68	0.33	-0.47	3.56	0.34

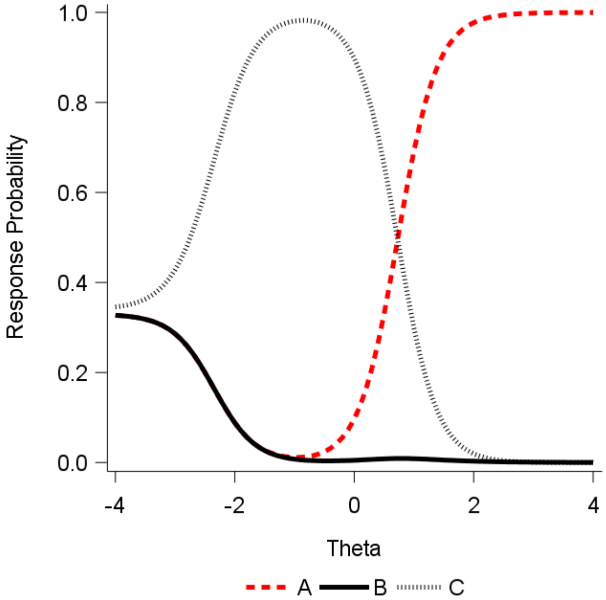

Figure 20. Implausible distractor condition item. A is the correct answer. B is the distractor of interest.

The second type of undesirable distractor is one with a flat trace line (Haladyna & Downing, 1993). Samejima (1988) referred to these distractors as *equivalent distractors* because examinees across all ability levels have an equal chance of selecting this option. *Equivalent distractors* in an item do not provide any pertinent information to estimates of ability due to the low discrimination they produce (Thissen et al., 1989). Figure 21 and Table 12 provide the ICCCs and parameter values for an example item containing an equivalent distractor (choice B).

Table 12

Example of item parameters for the equivalent distractor condition

a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
1.85	1.73	0.33	1.68	1.46	0.33	0.43	0.37	0.34

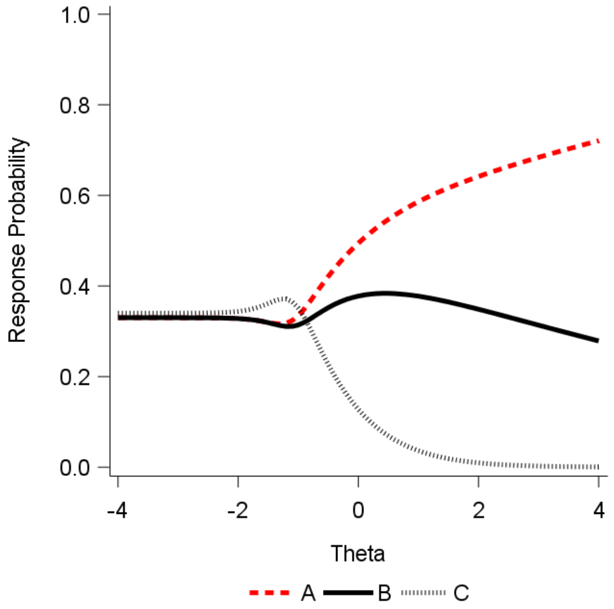

Figure 21. Equivalent distractor condition item. A is the correct answer. B is the distractor of interest.

The final type of undesirable distractor is a distractor that has a monotonically increasing trace line (Haladyna & Downing, 2013). The lure distractor can either be (1) higher in probability of selection than the correct answer, or (2) slightly lower in probability of selection of the correct answer for high ability examinees. If the ICCC for this distractor mirrors the pattern expected of the correct answer, then this type of distractor is known as an *upper lure distractor*. Although unlikely, *upper lure distractors* are possibly due to an issue with the test key or confusion among examinees about what the actual correct response is. Table 13 and Figure 22 provide ICCCs and parameter estimates for an item containing an *upper lure distractor* (choice B).

Table 13

Example of item parameters for the high lure distractor condition

a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
2.67	1.46	0.33	2.67	1.56	0.33	-1.42	0.10	0.34

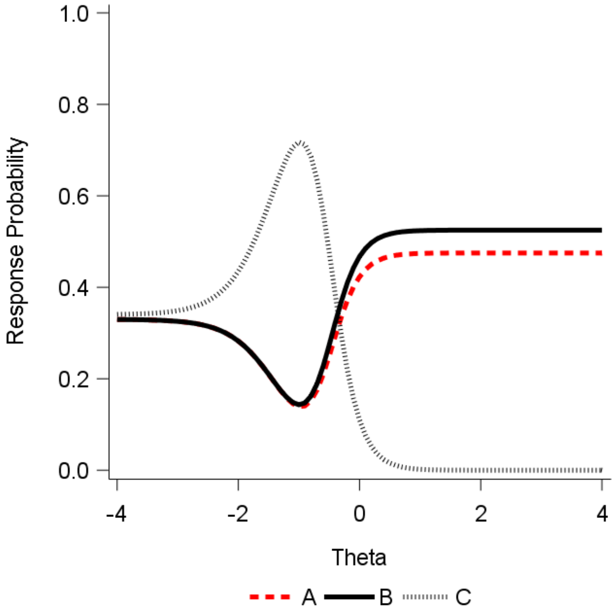

Figure 22. Upper lure distractor condition item. A is the correct answer. B is the distractor of interest.

The parameter values for the lower lure distractor condition were similar to the upper lure distractor condition parameters, but the lower lure distractor ICCC is below the correct option ICCC. Table 14 and Figure 23 present the ICCCs and item parameters for an item containing a low lure distractor (choice B).

Table 14

Example of item parameters for the low lure distractor condition

a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
2.46	1.47	0.33	2.46	0.98	0.33	-1.08	0.76	0.34

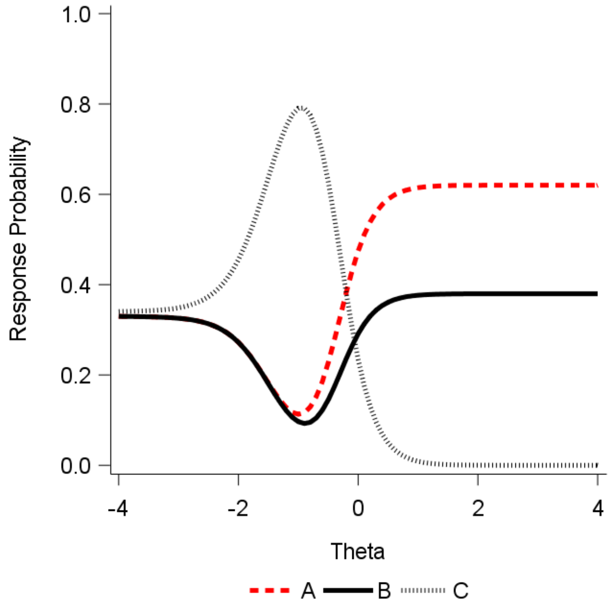

Figure 23. Lower lure distractor condition item. A is the correct answer. B is the distractor of interest.

Fixed Conditions. I fixed three conditions: sample size (*N*=2000), number of options per item (3), and guessing. I chose these conditions based on their relevance in high-stakes educational testing situations. I used a sample size of 2000 because a sufficient sample size is needed in order to produce accurate parameter estimates when using polytomous data (Reise & Yu, 1990). I selected the number of options based on evidence that suggests the use of three options, since four options do not provide more psychometric information in comparison to three options (Baghaei & Amrahi, 2011; Crehan, Haladyna, & Brewer, 1993; Haladyna & Downing, 2003; Haladyna, Rodriguez, & Stevens, 2019; Rodriguez, 2005; Vyas & Supe, 2008). The three options are (1) the correct answer, (2) the distractor of interest, and (3) all other distractor information that examinees are randomly choosing between. The third category can be generalized to two, three, or more total distractors for a test item with unspecified distractor patterns.

Data Generating Models

I systematically sampled true abilities from a standard normal distribution with a sample of *N*=2000 persons. I took 2000 theta values at equally spaced percentiles between .02275 for an ability of -2 and .97725 for an ability of 2. The probability of an examinee selecting the correct response for each item was computed as a function of the examinee's ability and the item parameters. I used the 3-PL and the TSMCM models to generate item responses. Specifically, I generated responses to items containing undesirable distractors using the TSMCM and responses to items not containing manipulated distractors using the 3-PL. The TSMCM is used to generate the responses to item containing undesirable distractors because the 3-PL only analyzes the correct answer and groups the distractors together. If the distractors are grouped together as simply incorrect, then I cannot model the undesirable distractors.

Under the TSMCM, true item parameters reflected characteristics of the four types of undesirable distractors. I auditioned true item parameters in SAS by specifying a range of values and randomly selecting values in this range. This allowed for me to produce items with trace lines that reflected the undesirable distractor presented with some variability that would be present in a typical testing situation. Tables 15, 16, 17, and 18 provide the ranges that true item parameters were selected from for each type of undesirable distractor.

I specified option A to be the correct answer for all items. This was arbitrarily chosen for simplicity. I set option B as the distractor of interest (i.e., undesirable distractor). Finally, I fixed item parameters associated with option C depending upon the true parameters of the correct option and undesirable distractor. I fixed the guessing parameter, d_k, at 0.33 for option A, 0.33 for option B, and 0.34 for option C. By fixing d_k to these values, the “don’t know” (DK) parameters converge to a probability of one. Fixing the d_k parameters to be equal in the TSMCM is equivalent to Samejima’s (1979) model.

Table 15

Range of true item parameter estimates for an implausible distractor

	Don’t know $(k = 0)$		Correct Answer $(k = 1)$		Implausible Distractor $(k = 2)$		Other option $(k = 3)$
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
a_k	-4	-3	2	3	1	2	$= 0 - \sum_{k=0}^{2} a_k$
c_k	-4	-3	0	2	-2	-1	$= 0 - \sum_{k=0}^{2} c_k$
d_k			.33	.33	.33	.33	$= 1 - \sum_{k=1}^{2} d_k$

Table 16

Range of true item parameter estimates for an equivalent distractor

	Don’t know $(k = 0)$		Correct Answer $(k = 1)$		Equivalent Distractor $(k = 2)$		Other option $(k = 3)$
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
a_k	-4	-3.9	1.8	2	1.5	1.75	$= 0 - \sum_{k=0}^{2} a_k$
c_k	-4	-3.5	1.6	2	1.3	1.5	$= 0 - \sum_{k=0}^{2} c_k$
d_k			.33	.33	.33	.33	$= 1 - \sum_{k=1}^{2} d_k$

Table 17

Range of true item parameter estimates for an upper lure distractor

	Don't know $(k = 0)$		Correct Answer $(k = 1)$		Lure (1) Distractor $(k = 2)$		Other option $(k = 3)$
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
a_k	-4	-3	2	3	a1	a1	$= 0 - \sum_{k=0}^{2} a_k$
c_k	-4	-3	1	2	c1	3	$= 0 - \sum_{k=0}^{2} c_k$
d_k	[redacted]	[redacted]	.33	.33	.33	.33	$= 1 - \sum_{k=1}^{2} d_k$

Table 18

Range of true item parameter estimates for a lower lure distractor

	Don't know $(k = 0)$		Correct Answer $(k = 1)$		Lure (2) Distractor $(k = 2)$		Other option $(k = 3)$
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	
a_k	-4	-3	2	2.5	a1	a1	$= 0 - \sum_{k=0}^{2} a_k$
c_k	-4	-3	1	1.5	0.5	c1	$= 0 - \sum_{k=0}^{2} c_k$
d_k	[redacted]	[redacted]	.33	.33	.33	.33	$= 1 - \sum_{k=1}^{2} d_k$

Using SAS, I randomly selected true item parameters. The item parameters are listed in Tables 19, 20, 21, and 22. The tables each contain 50 sets of true item parameters for each type of undesirable distractor. The condition with the largest number of items containing undesirable distractors (50%) and the longest test (100 items) results in a total of 50 true item parameter sets. The true parameters for conditions that contained less than 50% of undesirable distractors or a shorter test were randomly selected from the tables.

Table 19
Implausible distractor true item parameters

Multiple-Choice Model									
Item	a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
1	2.73	0.11	0.33	1.32	-1.83	0.33	-0.75	5.23	0.34
2	2.97	1.70	0.33	1.93	-1.10	0.33	-1.62	3.04	0.34
3	2.54	0.45	0.33	1.29	-1.19	0.33	-0.02	4.07	0.34
4	2.96	0.29	0.33	1.92	-1.49	0.33	-0.95	4.66	0.34
5	2.03	1.74	0.33	1.47	-1.91	0.33	-0.48	3.53	0.34
6	2.34	1.40	0.33	1.43	-1.37	0.33	0.15	3.89	0.34
7	2.60	0.17	0.33	1.11	-1.39	0.33	-0.44	4.87	0.34
8	2.51	1.26	0.33	1.14	-1.78	0.33	0.26	4.08	0.34
9	2.72	1.63	0.33	1.17	-1.57	0.33	-0.30	3.60	0.34
10	2.77	0.86	0.33	1.78	-1.24	0.33	-0.98	3.62	0.34
11	2.75	1.89	0.33	1.29	-1.23	0.33	-0.65	3.05	0.34
12	2.62	0.96	0.33	1.25	-1.18	0.33	0.07	3.46	0.34
13	2.11	0.88	0.33	1.46	-1.00	0.33	0.08	3.74	0.34
14	2.68	1.39	0.33	1.41	-1.43	0.33	-0.95	3.58	0.34
15	2.10	1.30	0.33	1.60	-1.13	0.33	-0.04	3.00	0.34
16	2.10	0.68	0.33	1.78	-1.82	0.33	-0.43	4.87	0.34
17	2.20	1.13	0.33	1.11	-1.34	0.33	0.35	3.45	0.34
18	2.58	1.07	0.33	1.84	-1.97	0.33	-1.31	4.59	0.34
19	2.63	1.40	0.33	1.23	-1.12	0.33	0.14	2.78	0.34
20	2.07	1.30	0.33	1.45	-1.66	0.33	-0.18	3.60	0.34
21	2.17	1.46	0.33	1.67	-1.20	0.33	0.02	3.65	0.34
22	2.15	1.09	0.33	1.84	-1.38	0.33	-0.59	3.73	0.34
23	2.53	1.84	0.33	1.35	-1.39	0.33	-0.08	3.33	0.34
24	2.97	0.33	0.33	1.41	-1.59	0.33	-0.87	4.73	0.34
25	2.53	0.79	0.33	1.18	-1.73	0.33	-0.36	4.19	0.34
26	2.19	0.22	0.33	1.61	-1.50	0.33	-0.35	4.59	0.34
27	2.91	0.47	0.33	1.56	-1.03	0.33	-1.03	3.75	0.34
28	2.23	1.07	0.33	1.46	-1.33	0.33	-0.48	3.57	0.34
29	2.01	1.94	0.33	1.24	-1.87	0.33	0.73	3.85	0.34
30	2.77	1.34	0.33	1.78	-1.74	0.33	-1.40	4.05	0.34
31	2.84	1.35	0.33	1.84	-1.40	0.33	-0.76	4.01	0.34
32	2.32	0.56	0.33	1.19	-1.72	0.33	-0.08	4.39	0.34
33	2.86	1.07	0.33	1.47	-1.99	0.33	-0.42	4.07	0.34
34	2.87	0.94	0.33	1.09	-1.80	0.33	-0.29	3.92	0.34
35	2.03	0.02	0.33	1.39	-1.96	0.33	0.27	5.18	0.34
36	2.18	0.12	0.33	1.68	-1.68	0.33	-0.61	4.59	0.34

Table 19 (contined)

37	2.81	1.07	0.33	1.25	-2.00	0.33	-0.74	4.91	0.34
38	2.91	1.79	0.33	1.91	-1.44	0.33	-1.47	3.16	0.34
39	2.91	1.17	0.33	1.05	-1.82	0.33	-0.26	4.23	0.34
40	2.42	0.63	0.33	1.80	-1.81	0.33	-0.30	4.34	0.34
41	2.59	0.13	0.33	1.66	-1.97	0.33	-0.56	5.34	0.34
42	2.79	1.97	0.33	1.71	-1.54	0.33	-0.80	2.70	0.34
43	2.05	0.77	0.33	1.93	-1.49	0.33	-0.39	3.83	0.34
44	2.10	0.35	0.33	1.14	-1.53	0.33	0.67	4.29	0.34
45	2.04	0.90	0.33	1.39	-1.65	0.33	0.24	4.23	0.34
46	2.91	0.72	0.33	1.43	-1.93	0.33	-0.65	5.14	0.34
47	2.18	0.50	0.33	1.79	-1.68	0.33	-0.19	4.98	0.34
48	2.68	0.60	0.33	1.56	-1.34	0.33	-0.25	3.86	0.34
49	2.14	0.15	0.33	1.89	-1.20	0.33	-0.37	4.41	0.34
50	2.29	0.62	0.33	1.01	-1.03	0.33	0.52	4.30	0.34

Table 20

Equivalent distractor true item parameters

Multiple-Choice Model									
Item	a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
1	1.98	1.74	0.33	1.74	1.40	0.33	0.25	0.38	0.34
2	1.82	1.68	0.33	1.61	1.35	0.33	0.49	0.85	0.34
3	1.93	1.61	0.33	1.69	1.46	0.33	0.30	0.80	0.34
4	1.96	1.94	0.33	1.55	1.50	0.33	0.47	0.42	0.34
5	2.00	1.64	0.33	1.56	1.40	0.33	0.41	0.66	0.34
6	1.99	1.79	0.33	1.73	1.37	0.33	0.23	0.47	0.34
7	1.89	1.64	0.33	1.74	1.35	0.33	0.34	0.90	0.34
8	1.92	1.80	0.33	1.71	1.44	0.33	0.33	0.30	0.34
9	1.86	1.73	0.33	1.60	1.48	0.33	0.45	0.76	0.34
10	1.84	1.93	0.33	1.53	1.34	0.33	0.58	0.63	0.34
11	1.93	1.79	0.33	1.74	1.33	0.33	0.33	0.64	0.34
12	1.98	1.63	0.33	1.67	1.44	0.33	0.27	0.72	0.34
13	1.94	1.82	0.33	1.75	1.46	0.33	0.29	0.41	0.34
14	1.87	1.85	0.33	1.72	1.33	0.33	0.38	0.61	0.34
15	1.86	1.73	0.33	1.70	1.47	0.33	0.42	0.71	0.34
16	1.94	1.67	0.33	1.62	1.41	0.33	0.39	0.56	0.34
17	1.87	1.73	0.33	1.50	1.38	0.33	0.57	0.46	0.34
18	1.87	1.73	0.33	1.64	1.37	0.33	0.47	0.56	0.34
19	1.97	1.69	0.33	1.70	1.42	0.33	0.31	0.66	0.34
20	1.81	1.97	0.33	1.55	1.35	0.33	0.61	0.28	0.34
21	1.82	1.84	0.33	1.66	1.44	0.33	0.46	0.63	0.34
22	1.81	1.69	0.33	1.58	1.43	0.33	0.54	0.59	0.34
23	1.98	1.84	0.33	1.58	1.34	0.33	0.37	0.57	0.34
24	1.85	1.75	0.33	1.55	1.38	0.33	0.58	0.71	0.34

Table 20 (continued)

25	1.81	1.97	0.33	1.68	1.45	0.33	0.43	0.46	0.34
26	1.86	2.00	0.33	1.70	1.48	0.33	0.43	0.26	0.34
27	1.84	1.90	0.33	1.59	1.32	0.33	0.55	0.34	0.34
28	1.81	1.68	0.33	1.65	1.44	0.33	0.45	0.67	0.34
29	1.94	1.70	0.33	1.62	1.46	0.33	0.39	0.39	0.34
30	1.91	1.85	0.33	1.61	1.42	0.33	0.43	0.50	0.34
31	1.91	1.89	0.33	1.69	1.35	0.33	0.37	0.45	0.34
32	1.81	1.68	0.33	1.55	1.46	0.33	0.61	0.42	0.34
33	1.91	1.81	0.33	1.66	1.48	0.33	0.36	0.30	0.34
34	1.99	1.84	0.33	1.53	1.33	0.33	0.42	0.74	0.34
35	1.90	1.71	0.33	1.69	1.38	0.33	0.38	0.75	0.34
36	1.98	1.92	0.33	1.66	1.46	0.33	0.28	0.27	0.34
37	1.99	1.65	0.33	1.75	1.34	0.33	0.18	0.93	0.34
38	1.96	1.72	0.33	1.58	1.40	0.33	0.42	0.45	0.34
39	1.85	1.93	0.33	1.59	1.32	0.33	0.51	0.31	0.34
40	1.83	1.84	0.33	1.61	1.32	0.33	0.55	0.83	0.34
41	1.88	1.66	0.33	1.63	1.45	0.33	0.44	0.88	0.34
42	1.95	1.96	0.33	1.67	1.32	0.33	0.37	0.65	0.34
43	1.93	1.76	0.33	1.67	1.43	0.33	0.39	0.59	0.34
44	1.94	1.64	0.33	1.69	1.32	0.33	0.35	0.82	0.34
45	1.86	1.88	0.33	1.69	1.46	0.33	0.37	0.63	0.34
46	1.95	1.66	0.33	1.65	1.48	0.33	0.37	0.85	0.34
47	2.00	1.90	0.33	1.57	1.33	0.33	0.39	0.28	0.34
48	2.00	2.00	0.33	1.62	1.36	0.33	0.29	0.62	0.34
49	1.90	1.80	0.33	1.62	1.38	0.33	0.39	0.67	0.34
50	1.81	1.84	0.33	1.69	1.49	0.33	0.42	0.17	0.34

Table 21

Upper lure distractor true item parameters

Multiple-Choice Model									
Item	a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
1	2.21	1.59	0.33	2.21	1.71	0.33	-0.65	0.00	0.34
2	2.36	1.49	0.33	2.36	1.73	0.33	-1.00	-0.16	0.34
3	2.13	1.99	0.33	2.13	2.00	0.33	-0.66	-0.36	0.34
4	2.14	1.67	0.33	2.14	1.96	0.33	-1.21	-0.08	0.34
5	2.22	1.52	0.33	2.22	1.80	0.33	-1.40	-0.31	0.34
6	2.00	1.63	0.33	2.00	1.65	0.33	-0.87	-0.09	0.34
7	2.30	1.59	0.33	2.30	1.96	0.33	-1.57	0.26	0.34
8	2.25	1.61	0.33	2.25	1.70	0.33	-1.17	-0.09	0.34
9	2.31	1.08	0.33	2.31	1.82	0.33	-1.53	0.65	0.34
10	2.38	1.38	0.33	2.38	1.95	0.33	-0.78	-0.05	0.34
11	2.49	1.52	0.33	2.49	1.70	0.33	-1.14	0.15	0.34
12	2.21	1.24	0.33	2.21	1.48	0.33	-0.87	1.12	0.34

Table 21 (continued)

13	2.15	1.85	0.33	2.15	1.96	0.33	-1.05	-0.46	0.34
14	2.42	1.74	0.33	2.42	1.78	0.33	-1.31	-0.27	0.34
15	2.32	1.31	0.33	2.32	1.42	0.33	-1.26	1.14	0.34
16	2.28	1.08	0.33	2.28	1.86	0.33	-0.83	0.98	0.34
17	2.14	1.33	0.33	2.14	1.96	0.33	-0.77	0.39	0.34
18	2.43	1.96	0.33	2.43	2.00	0.33	-0.87	-0.42	0.34
19	2.32	1.09	0.33	2.32	1.36	0.33	-0.76	1.25	0.34
20	2.49	1.60	0.33	2.49	1.95	0.33	-1.41	0.25	0.34
21	2.34	1.91	0.33	2.34	2.00	0.33	-1.12	0.04	0.34
22	2.31	1.03	0.33	2.31	1.93	0.33	-1.62	0.95	0.34
23	2.40	1.80	0.33	2.40	1.81	0.33	-1.07	-0.59	0.34
24	2.41	1.49	0.33	2.41	1.66	0.33	-0.89	-0.15	0.34
25	2.46	1.50	0.33	2.46	1.87	0.33	-1.69	0.58	0.34
26	2.04	1.28	0.33	2.04	1.91	0.33	-0.91	0.65	0.34
27	2.15	1.29	0.33	2.15	1.36	0.33	-0.56	0.98	0.34
28	2.16	1.88	0.33	2.16	1.90	0.33	-1.00	-0.20	0.34
29	2.47	1.15	0.33	2.47	1.75	0.33	-1.09	0.39	0.34
30	2.02	1.74	0.33	2.02	1.86	0.33	-0.53	0.31	0.34
31	2.34	1.12	0.33	2.34	1.20	0.33	-1.03	1.53	0.34
32	2.36	1.88	0.33	2.36	1.99	0.33	-0.86	0.07	0.34
33	2.01	1.12	0.33	2.01	1.79	0.33	-0.31	0.99	0.34
34	2.05	1.03	0.33	2.05	1.15	0.33	-0.58	1.82	0.34
35	2.27	1.93	0.33	2.27	1.98	0.33	-1.20	-0.41	0.34
36	2.08	1.70	0.33	2.08	1.97	0.33	-0.89	0.09	0.34
37	2.21	1.94	0.33	2.21	1.99	0.33	-1.27	-0.22	0.34
38	2.16	1.23	0.33	2.16	1.95	0.33	-1.25	0.01	0.34
39	2.18	1.44	0.33	2.18	1.84	0.33	-1.09	0.38	0.34
40	2.16	1.77	0.33	2.16	1.93	0.33	-0.52	-0.27	0.34
41	2.09	1.74	0.33	2.09	1.87	0.33	-0.26	-0.56	0.34
42	2.44	1.91	0.33	2.44	1.94	0.33	-1.58	-0.80	0.34
43	2.22	1.54	0.33	2.22	1.80	0.33	-0.72	-0.30	0.34
44	2.03	1.62	0.33	2.03	1.80	0.33	-0.10	-0.18	0.34
45	2.02	1.12	0.33	2.02	1.17	0.33	-0.03	0.82	0.34
46	2.15	1.86	0.33	2.15	1.93	0.33	-0.77	-0.56	0.34
47	2.36	1.69	0.33	2.36	1.88	0.33	-1.36	0.25	0.34
48	2.20	1.82	0.33	2.20	1.89	0.33	-1.31	0.19	0.34
49	2.49	1.60	0.33	2.49	1.79	0.33	-1.55	-0.02	0.34
50	2.04	1.59	0.33	2.04	1.88	0.33	-0.95	-0.37	0.34

Table 22

Lower lure distractor item parameters

Multiple-Choice Model									
Item	a_A	c_A	d_A	a_B	c_B	d_B	a_C	c_C	d_C
1	2.05	1.34	0.33	2.05	1.03	0.33	-0.45	1.42	0.34
2	2.13	1.30	0.33	2.13	0.83	0.33	-0.74	1.15	0.34
3	2.41	1.27	0.33	2.41	0.66	0.33	-1.73	1.91	0.34
4	2.31	1.36	0.33	2.31	1.04	0.33	-1.27	1.13	0.34
5	2.09	1.24	0.33	2.09	1.16	0.33	-0.19	1.24	0.34
6	2.16	1.40	0.33	2.16	1.28	0.33	-1.31	0.36	0.34
7	2.29	1.40	0.33	2.29	1.25	0.33	-1.46	0.89	0.34
8	2.03	1.14	0.33	2.03	0.91	0.33	-0.19	1.27	0.34
9	2.20	1.50	0.33	2.20	1.48	0.33	-0.52	0.65	0.34
10	2.09	1.19	0.33	2.09	0.69	0.33	-0.27	1.60	0.34
11	2.32	1.02	0.33	2.32	0.54	0.33	-1.29	1.95	0.34
12	2.11	1.28	0.33	2.11	0.65	0.33	-0.96	1.58	0.34
13	2.31	1.18	0.33	2.31	0.73	0.33	-1.48	1.46	0.34
14	2.44	1.32	0.33	2.44	0.76	0.33	-1.69	1.14	0.34
15	2.35	1.17	0.33	2.35	0.52	0.33	-1.47	1.85	0.34
16	2.50	1.42	0.33	2.50	0.68	0.33	-1.85	1.32	0.34
17	2.06	1.40	0.33	2.06	1.18	0.33	-0.83	1.39	0.34
18	2.21	1.12	0.33	2.21	0.90	0.33	-1.23	0.99	0.34
19	2.06	1.39	0.33	2.06	1.22	0.33	-0.55	0.82	0.34
20	2.06	1.25	0.33	2.06	0.62	0.33	-0.81	1.75	0.34
21	2.08	1.17	0.33	2.08	0.89	0.33	-1.14	1.10	0.34
22	2.02	1.11	0.33	2.02	0.78	0.33	-0.59	2.03	0.34
23	2.48	1.40	0.33	2.48	1.27	0.33	-1.57	0.64	0.34
24	2.24	1.33	0.33	2.24	0.63	0.33	-1.46	1.80	0.34
25	2.27	1.01	0.33	2.27	0.80	0.33	-1.35	1.98	0.34
26	2.32	1.38	0.33	2.32	0.54	0.33	-1.51	1.78	0.34
27	2.37	1.34	0.33	2.37	1.25	0.33	-1.64	0.97	0.34
28	2.42	1.17	0.33	2.42	1.01	0.33	-1.70	1.06	0.34
29	2.18	1.09	0.33	2.18	0.79	0.33	-1.17	1.98	0.34
30	2.45	1.05	0.33	2.45	0.98	0.33	-1.46	1.30	0.34
31	2.29	1.49	0.33	2.29	0.67	0.33	-1.21	0.98	0.34
32	2.03	1.11	0.33	2.03	0.59	0.33	-0.56	1.67	0.34
33	2.32	1.37	0.33	2.32	0.64	0.33	-0.98	1.17	0.34
34	2.26	1.35	0.33	2.26	1.33	0.33	-0.58	0.75	0.34
35	2.34	1.34	0.33	2.34	1.34	0.33	-0.76	1.32	0.34
36	2.30	1.32	0.33	2.30	0.51	0.33	-0.76	1.38	0.34
37	2.25	1.19	0.33	2.25	0.56	0.33	-1.17	2.22	0.34
38	2.06	1.34	0.33	2.06	0.91	0.33	-0.49	1.52	0.34
39	2.49	1.25	0.33	2.49	0.59	0.33	-1.91	2.08	0.34
40	2.09	1.24	0.33	2.09	0.85	0.33	-1.00	1.76	0.34

Table 22 (continued)

41	2.09	1.06	0.33	2.09	0.79	0.33	-0.51	2.04	0.34
42	2.12	1.14	0.33	2.12	0.66	0.33	-0.59	1.26	0.34
43	2.23	1.35	0.33	2.23	1.20	0.33	-1.34	0.98	0.34
44	2.36	1.49	0.33	2.36	1.02	0.33	-1.50	0.69	0.34
45	2.33	1.13	0.33	2.33	0.94	0.33	-1.46	1.79	0.34
46	2.35	1.15	0.33	2.35	0.60	0.33	-0.78	1.64	0.34
47	2.10	1.22	0.33	2.10	1.03	0.33	-0.60	0.94	0.34
48	2.36	1.21	0.33	2.36	0.79	0.33	-1.46	2.00	0.34
49	2.31	1.46	0.33	2.31	1.32	0.33	-0.69	0.55	0.34
50	2.25	1.12	0.33	2.25	0.54	0.33	-1.02	1.76	0.34

I used the 3-PL model to generate responses of normal functioning items. Using SAS, I randomly selected true difficulty item parameters (b_i) from a standard normal distribution and true discrimination item parameters (a_i) from a distribution of N(0,1), but the lower bound for values was set to 0. I fixed the true guessing parameters to a value of 0.33. I then auditioned item parameters for unrealistic pairings. For example, no item difficulty below -1 is associated with a discrimination above 1. This represents an unrealistic situation in which a very easy item would highly discriminate respondents.

After I selected true item parameters, I merged the true item parameters and true ability parameters into one dataset to calculate the probabilities of responses. Under the TSMCM, for each person, I calculated three probabilities. These were the probabilities that each person selected options A, B, and C. For example, for conditions with a test length of 10 items, 30 probabilities were calculated for each person. The Rand(‘Table’, pA, pB, pC) function in SAS assigned a score depending on the probabilities. Using the TSMCM model, I used SAS to calculate the probability of each option being selected based on the item parameters. Scores are assigned to examinees based on the probability produced by the TSMCM model and their associated ability. Scores were specified as 1, 2, and 3. These scores were translated so that 1 was equal to A, 2 to B, and 3 to C. I then scored each item so that A indicated a score of 1, B a score of 0, and C a score 0. This is because A is the correct answer.

For the remaining normally functioning items, I used SAS and the 3-PL model to calculate a probability for each person selecting the correct answer. I used true ability and the true item parameters to calculate this. For example, for conditions with a test length of 10 items, each person has 10 probabilities. I then randomly selected a value between 0 and 1 from a uniform distribution. If the value is less than the probability, the item is scored 1 (for correct). If this randomly selected value is greater than the calculated probability, then the person receives a score of 0. The final result was a dataset containing a person's score for each item, combining generated responses from the TSMCM and the 3PL.

In addition to the undesirable distractor conditions, I created a control group to generate responses completely based on the 3-PL model. The control group acted as a baseline to facilitate comparisons between experimental conditions. The establishment of a baseline not only aided in comparing undesirable distractors and test length to those from a test with well-functioning items, but the results contribute to the literature in how well the SAS IRT procedure estimation routine recovers ability parameters.

Analysis

The first research question pertains to how well the SAS 9.4 IRT procedure estimates ability using the 3-PL model. Similar to Cole and Paek (2017), who evaluated the recovery of item parameters for multiple types of IRT models (e.g., 1-PL, Rasch, 2-PL, 3-PL, etc.) using the SAS 9.4 IRT procedure, the current study examines how well ability parameters are recovered. To address this question, full responses were simulated under a fixed sample size of 2000 (*N*=2000) using the 3-PL model in all conditions of test length (30, 50, and 100 items). Values of bias and SE were investigated for the recovery of both ability and item parameters. The analysis of the recovery of ability parameters with the 3-PL model in SAS 9.4 was completed first to see how well these parameters would be recovered. This allowed us to answer our other research questions related to the models of interest with confidence in our interpretations.

Proc IRT in SAS 9.4 was used to model the responses by obtaining ability estimates and item parameter estimates. To estimate ability for each person, maximum a posteriori (MAP) was used. This is the default estimation for abilities in SAS 9.4. Item parameter estimates of difficulty (b_i), discrimination (a_i), and guessing (c_i) were estimated for each item using marginal maximum likelihood (MML). A total of 1000 replications were completed for each condition. Harwell et al. (1996) suggest at least 25 replications when performing an IRT simulation study. Based on suggestions from Feinberg and Rubright (2016), a total of 1000 replications were completed to estimate the stability of the simulation results. Harwell et al.'s (1996) recommendation is outdated, and with the advancement of computing capabilities, it is appropriate to increase the number of replications.

After completing the simulation and analysis cycles, the degree of bias in ability estimates was examined. For this study, the outcome of interest was investigated at the total score level. This was completed by calculating the expected response function (ERF; Luecht & Ackerman, 2018) and the root mean square error (RMSE). Luect and Ackerman (2018) suggest using the ERF when there are multiple models involved in the simulation. The generating and analysis models differed so the ERF was used to gauge ability. The true expected total score is calculated as,

$$f(\Sigma x|\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m} x_{jk} P_{jk} \qquad (12)$$

where the true expected total score is the sum of the expected item scores given the true ability and true item parameters.

To estimate the expected total score response, the sum of the estimated expected item scores given estimated ability and estimated item parameters is calculated:

$$\hat{f}\left(\Sigma x\middle|\hat{\theta}\right) = \sum_{j=1}^{n} \sum_{k=1}^{m} x_{jk} \widehat{P_{jk}} \qquad (13)$$

The true and estimated expected total score responses allowed for the comparison of bias in the estimated ability parameters. Average bias for the expected total score (Equation 14) was calculated:

$$\frac{\sum_{r=1}^{R}\left(f(\Sigma x|\theta) - \hat{f}\left(\Sigma x\middle|\hat{\theta}\right)\right)}{R} \qquad (14)$$

where R is the number of replications ($R = 1000$). This index of bias allow researchers to examine the discrepancies between true ability parameters and estimated ability parameter estimates. Values of bias increase as these estimates grow further apart.

Total score bias allows for the examination of the amount of bias in examinees' scores on the entire test. Total score bias takes into account parameter estimates across the entire test. Total score bias indicates the extent to which examinees' true ability and estimated ability produce differing scores on a test based on the true and estimated parameters across all items. Large discrepancies between true expected total score and estimated expected total score mean that examinees' total scores are not accurate representations of their ability. This bias can be attributed to the item parameters. Because total score bias examines the effects of parameter estimates on ability across all items, researchers can determine how the inclusion of items with undesirable distractors impact ability estimates when there are other items that are functioning well.

The empirical standard error (SE) was also calculated for each ability in each condition after 1000 replications. Similar to bias, I used expected total score as a proxy for ability. I calculated SE by first finding the average of the estimated expected total score ($\overline{\hat{f}\left(\Sigma x \middle| \hat{\theta}\right)}$) then found the squared distance of each of the 1000 expected total score estimates from the mean and finally divided by number of replications minus one and take the square root:

$$SE = \sqrt{\frac{\hat{f}\left(\Sigma x \middle| \hat{\theta}\right) - \hat{f}\overline{\left(\Sigma x \middle| \hat{\theta}\right)}^{2}}{R-1}} \tag{17}$$

Similar to bias, this was computed at each level of true ability.

Total score bias and SE values were compared using a fully-crossed 4x3x3. I evaluate the results using an ANCOVA. Due to the large sample (*N*=2000), effect sizes were examined along with statistical significance. For example, less than 1% of the variance may be statistically significant, but 1% is not necessarily practically significant. If the partial variance explained was greater than 6%, (Cohen, 1988), I considered the effect as practically significant. A continuous covariate was also added to account for the linear and quadratic effects of true ability.

Chapter 4. Results

Results of the recovery of ability parameters using the IRT procedure in SAS 9.4 (TS1M4) precede the analyses containing undesirable distractors. I calculated total score bias and standard errors along the ability continuum for each condition (i.e., test length) to determine the degree of bias and inefficiency in the parameter estimates. I used analysis of covariance (ANCOVA) to analyze the effect of varying conditions on bias and efficiency of parameter estimates.

PROC IRT Parameter Recovery

It is important to evaluate the recovery of ability parameters in order to understand the results of this simulation study. I first examine the 3-PL model's capabilities of recovering the ability parameter estimates for a typical multiple-choice test. This analysis was completed prior to evaluating the recovery of ability estimates when including the manipulated distractors to determine how well the ability parameters are recovered in a normal condition. Three conditions of varying levels of test length (30 items, 50 items, and 100 items) were generated using the 3-PL model. The procedure was able to converge to a solution using maximum a posteriori to estimate abilities for all three levels of test length for all replications when analyzing the data generated with the 3-PL model.

To explore the utility of the IRT procedure, I compared total score bias across conditions of test length while controlling for true ability by using an ANCOVA. Total score bias was averaged across 1000 replications. Sample size was held constant across all conditions (*N*=2000). Recall, expected total score is used as a proxy for latent ability. To make conditions comparable, I transformed total score bias to percent correct bias based on the number of items per condition. Not only did this facilitate interpretation and comparison, but also the number of items impacted the degree of total score bias. Total score bias was not accurately represented without changing it to a percent correct value because more items attenuated the total score bias values. Without conversion to percent correct bias, we would expect larger total score bias when there are more items. Raw score absolute comparison would indicate a bias of 2 on expected score when there 30 items is equivalent to a total score bias of 2 when there are 100 items. However, these percent correct differences are 7% and 2% which are quite different. I will refer to this value as percent correct bias.

Table 23 provides results of the ANCOVA, where percent correct bias acted as the dependent variable, test length as an independent variable, and true ability as a covariate. I included a squared term of true ability to account for the curvilinear relationship between percent correct bias and true ability. Partial eta-squared was used to determine practical significance since the large sample size contributed to the statistical significance of all main effects and interactions. I considered medium effect (i.e., partial $\eta^2 \geq .06$) sizes as being practically significant (Cohen, 1988).

Table 23

Baseline ANCOVA results for percent correct bias

Source	df	SS	MS	F	*p*	Partial η^2
Test length	2	24.68	12.34	628.22	<.001	**.17**
True ability	1	34958.57	34958.57	1780049	<.001	**.99**
True ability^2	1	24.54	24.54	1249.38	<.001	**.17**
True ability x Test length	2	4303.73	2151.87	109570	<.001	**.97**
True ability^2 x Test length	2	1.19	0.60	30.31	<.001	.01

Note. Practically significant effects are bolded.

The interaction between the squared true ability term and test length only accounted for a small proportion of the variance ($F(2, 5991) = 30.31, p < .001, \eta_p^2 =$

.01). In other words, the effect of test length on percent correct bias does not depend on values of true ability as a curvilinear term. The effect of test length on percent correct bias does depend on values of true ability ($F(2, 5991) = 109570, p < .001, \eta_p^2 = .97$). The main effects, true ability squared ($F(1, 5991) = 1249.38, p < .001, \eta_p^2 = .17$), true ability ($F(1, 5991) = 1780049, p < .001, \eta_p^2 = .99$), and test length ($F(2, 5991) = 628.22, p < .001, \eta_p^2 = .17$) were also found to be practically significant.

The interaction term of true ability and test length was found to be practically significant. As previously stated, this interaction indicates that the effect of test length on percent correct bias is conditional upon values of true ability while holding true ability squared constant. As test length increases (i.e., 30 items to 50 items to 100 items), percent correct bias decreases. Figure 24 provides a graphical depiction of the varying levels of test length with percent correct bias means across replications plotted for each level of ability.

The 100-item test results in the lowest absolute percent correct bias. Low ability examinees' true ability is overestimated. As theta increases toward an ability of 0, examinees' bias is close to 0 (i.e., there is no bias in percent correct score). Percent correct bias for the 100-item test results in increasingly positive values, indicating that examinees' expected percent correct score with true abilities between 1 and 2 are underestimated.

A 50-item test results in less biased expected percent correct scores across true ability in comparison to the 30-item test. Percent correct bias of the 50-item test for examinees with a true ability at -2 results in less of an overestimate than examinees with a true ability of -1. However, percent correct bias of the 50-item test becomes less of an overestimate of expected percent correct scores at true ability of -1 to 0. From 0 to 2, the 50-item tests results in an increasing underestimate of bias for expected percent correct scores.

The 30-item test results in the highest absolute percent bias across all test lengths. The 30-item test results in percent correct bias that flattens out for examinees with a true ability between -2 and -1, then increases as true ability increases. The 30-item test best estimates ability for examinees with a true ability of 0. High ability examinees' true score is underestimated for the 30-item test.

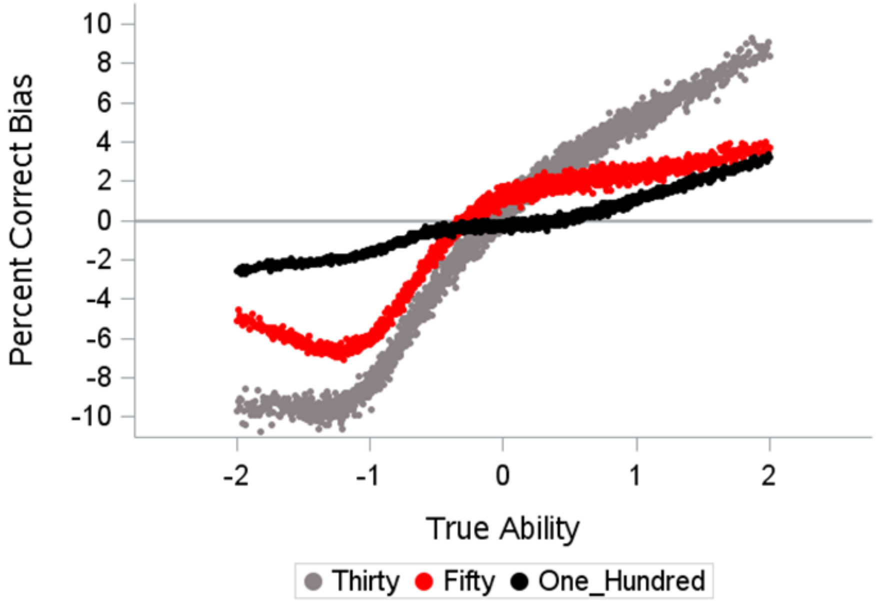

Figure 24. Percent correct bias as a function of true ability grouped by test length.

As seen in Figure 24, longer tests indicate less percent correct bias across true ability. Across all test lengths, low ability examinees' true ability is overestimated while high ability examinees' expected percent correct scores are underestimated. We observe this trend because, when MAP is employed to estimate abilities, abilities are drawn in closer to the mean. Examinees in the middle of the ability continuum (i.e., around a true ability of 0) have the least percent correct bias across all conditions of test length. The

50-item and 100-item tests result in varying degrees of percent correct bias for low ability examinees. At a true ability of -2, the 100-item test results in lower absolute percent correct bias than the 50-item test. However, the 50-item and 100-item tests have similar degrees of percent correct bias for high ability examinees. At a true ability of 2, examinees' expected percent correct scores are underestimated, as estimated by similar amounts of bias. The differences in these patterns explain the significant interaction of test length and true ability.

I also examined standard error to evaluate the efficiency in estimates of ability. In order to put this on a comparable metric, I analyzed percent correct standard error. Table 24 provides the ANCOVA results, using standard error as a dependent variable. Test length was entered as the independent variable, true ability was entered as covariate, as well as a squared true ability term to account for a curvilinear relationship. Partial eta-squared was used to determine practical significance since the large sample size contributed to the statistical significance of all main effects and interactions. I considered medium effect (i.e., partial $\eta^2 \geq .06$) sizes as being practically significant (Cohen, 1988).

Table 24
Baseline ANCOVA results for standard error

Source	df	SS	MS	F	*p*	Partial η^2
Test length	2	2752.81	1376.40	366176	<.001	**.992**
True ability	1	251.69	251.69	66958.6	<.001	**.918**
True ability^2	1	106.86	106.86	28428.4	<.001	**.826**
True ability x Test length	2	32.34	16.17	4301.81	<.001	**.589**
True ability^2 x Test length	2	7.14	3.57	949.11	<.001	**.241**

Note. Practically significant effects are bolded.

All main effects and interactions were found to be practically significant. Standard error differs across various levels of test length, controlling for true ability ($F(2, 5991) = 366176, p < .001, \eta_p^2 = .992$). Standard error also differs across the levels of both true ability ($F(1, 5991) = 66958.6, p < .001, \eta_p^2 = .918$) and true ability squared ($F(2, 5991) = 28428.4, p < .001, \eta_p^2 = .826$) after controlling for test length. While the effect of test length on standard error depends on the true ability ($F(2, 5991) = 4301.81, p < .001, \eta_p^2 = .589$), the interaction of test length and true ability squared was also practically significant ($F(2, 5991) = 949.11, p < .001, \eta_p^2 = .241$). The differences between standard error associated with test lengths at lower levels of ability are greater than the differences between standard error associated with higher levels of ability.

Figure 25 provides a graphical depiction of the relationship between mean standard error across replications and true ability. This relationship is curvilinear. The change in differences of standard error between test lengths across true ability shows the interaction of test length and true ability. Across all levels of test length, standard error tends to be highest for low ability examinees and lowest for high ability examinees. A 30-item test results in the greatest amount of standard error, while a 100-item test results in the least amount of standard error. As true ability increases, the distances between test lengths decrease. Examinees with a true ability of 2 have smaller differences across test length in comparison to examinees with true abilities of -2.

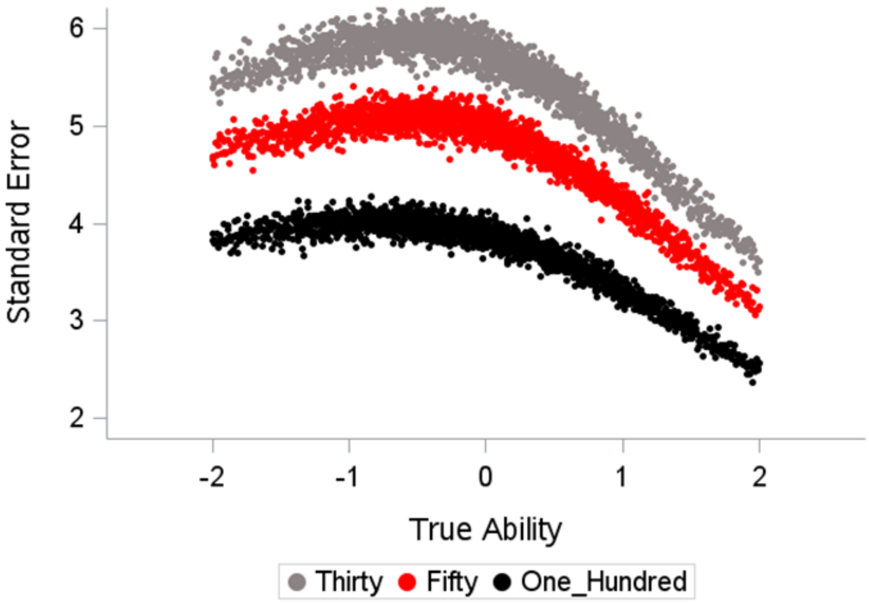

Figure 25. Standard error as a function of true ability grouped by test length.

After analyzing 3-PL ability estimates without manipulated distractors, I focus my attention on the analysis with undesirable distractors.

Percent Correct Bias

I analyzed the generated item responses using ANCOVA to examine the effects of undesirable distractors on percent correct bias. There were three categorical independent variables: 1) undesirable distractor type, 2) test length, and 3) percentage of items containing undesirable distractors. There were four types of undesirable distractors, including implausible, equivalent, upper lure, and lower lure. Test length contained three levels (30 items, 50 items, and 100 items), and the percentage of items containing undesirable also had three levels (10%, 30%, and 50%). Additionally, true ability and true ability squared were used as covariates. The ANCOVA was fully crossed (4x3x3) with the inclusion of main effects and interactions. Practical significance, in addition to statistical significance, was examined. I considered medium effect (i.e., partial $\eta^2 \geq .06$) sizes as being practically significant (Cohen, 1988).

Table 25 provides results of the ANCOVA. Based on the cut-off of $\eta_p^2 \geq .06$, three main effects and five interactions were found to be practically significant, while the other 15 main effects and interactions were not practically significant.

Table 25
ANCOVA results for percent correct bias

Source	df	SS	MS	F	*p*	Partial η^2
L	2	55.22	27.61	210.12	<.0001	.006
P	2	96.13	48.06	365.77	<.0001	.010
T	3	854.86	284.95	2168.51	<.0001	**.083**
A	1	367125.58	367125.58	2793852	<.0001	**.975**
A2	1	782.40	782.40	5954.12	<.0001	**.076**
L x P	4	21.033	5.26	40.02	<.0001	.002
L x T	6	30.87	5.15	39.16	<.0001	.003
L x A	2	4500.51	22500.25	171299	<.0001	**.827**
L x A2	2	50.45	25.23	191.96	<.0001	.005
P x T	6	219.80	36.63	278.78	<.0001	.023
P x A	2	193.66	96.83	736.90	<.0001	.020
P x A2	2	227.25	113.63	864.70	<.0001	.024
T x A	3	9007.77	3002.59	22849.90	<.0001	**.488**
T x A2	3	1347.95	449.32	3419.32	<.0001	**.125**
L x P x T	12	15.37	1.28	9.74	<.0001	.002
L x P x A	4	44.51	11.13	84.68	<.0001	.005
L x P x A2	4	6.97	1.74	13.25	<.0001	.001
L x T x A	6	1034.37	172.40	1311.94	<.0001	**.099**
L x T x A2	6	33.94	5.66	43.05	<.0001	.004
P x T x A	6	1187.52	197.92	1506.19	<.0001	**.112**
P x T x A2	6	392.73	65.45	498.11	<.0001	.040
L x P x T x A	12	179.42	14.95	113.78	<.0001	.019
L x P x T x A2	12	21.51	1.79	13.64	<.0001	.002

Note. Test length = L; Percentage = P; Undesirable distractor type = T; True ability = A; True ability^2 = A2; practically significant effects are bolded.

Percent Correct Bias: Insignificant Effects. Percent correct bias does not differ across levels of test length after controlling for percentage, undesirable distractor type, and true ability ($F(2, 71892) = 210.12, p < .001, \eta_p^2 = .006$). Percent correct bias also does not differ across levels of percentage after controlling for test length, undesirable distractor type, and true ability ($F(2, 71892) = 365.77, p < .001, \eta_p^2 = .010$). The two-way interactions of test length and percentage ($F(4, 71892) = 40.02, p < .001, \eta_p^2 =$

.002), test length and undesirable distractor type ($F(6, 71892) = 39.16, p < .001, \eta_p^2 =$ .003), test length and true ability squared ($F(2, 71892) = 191.96, p < .001, \eta_p^2 =$ .005), percentage and undesirable distractor type ($F(6, 71892) = 278.78, p < .001, \eta_p^2 =$ .023), percentage and true ability ($F(2, 71892) = 736.90, p < .001, \eta_p^2 =$ .020), and percentage and true ability squared ($F(2, 71892) = 864.70, p < .001, \eta_p^2 =$ .024) were not practically significant. In other words, percent correct bias does not differ across varying levels of, for example, test length and percentage, after controlling for undesirable distractor type and true ability.

Five of the seven three-way interactions were also not practically significant. There was no three-way interaction effect on percent correct bias among test length, percentage, and undesirable distractor type ($F(12, 71892) = 9.74, p < .001, \eta_p^2 =$ .002). Similarly, there was no practically significant interaction among test length, percentage, and true ability ($F(4, 71892) = 84.68, p < .001, \eta_p^2 =$ .005), test length, percentage, and true ability squared ($F(2, 71892) = 13.25, p < .001, \eta_p^2 =$ .001), and percentage, undesirable distractor type, and true ability squared ($F(6, 71892) = 498.11, p <$ $.001, \eta_p^2 =$ .040). Finally, both four-way interactions of test length, percentage, undesirable distractor type, and true ability ($F(12, 71892) = 113.78, p < .001, \eta_p^2 =$ .019) and test length, percentage, undesirable distractor type, and true ability squared ($F(12, 71892) = 13.64, p < .001, \eta_p^2 =$ .002) were not practically significant.

Percent Correct Bias: Practically Significant Effects. The main effect of distractor type ($F(3, 71892) = 2168.51, p < .001, \eta_p^2 =$ .083), true ability ($F(1, 71892) = 2793852, p < .001, \eta_p^2 =$ .975), and true ability squared

($F(1, 71892) = 5954.12, p < .001, \eta_p^2 = .076$) were practically significant. The effect of true ability is moderated by test length on percent correct bias ($F(2, 71892) = 171299, p < .001, \eta_p^2 = .827$). The effect of undesirable distractor type on percent correct bias is dependent on true ability ($F(3, 71892) = 22849.90, p < .001, \eta_p^2 = .488$) and true ability squared ($F(3, 71892) = 3419.32, p < .001, \eta_p^2 = .125$). There are two three-way interactions with practically significant effects on percent correct bias, which are discussed in detail below.

Percent Correct Bias: Type x Percentage x True Ability Interaction. The first interaction is between undesirable distractor type, percentage, and true ability ($F(6, 71892) = 1506.19, p < .001, \eta_p^2 = .112$). The differences seen in percent correct bias for each type of undesirable distractor at each level of percentage vary across true ability. In particular, the lower lure distractor results in greatest amount of percent correct bias across true ability. In contrast, the differences in percent correct bias of the implausible, equivalent, and upper lure distractors differ across true ability. To better understand the three-way interaction, I examined Figures 26, 27, and 28 in which the relationship between true ability and percent correct bias among the four distractor type is separated by 10%, 30%, and 50% of items containing undesirable distractors, respectively.

In Figure 26, 10% of items contain an undesirable distractor. The equivalent distractor results in the lowest absolute bias of all undesirable distractor types across true ability. Low ability examinees (-2) receive underestimates of their expected percent correct scores. From a true ability of -2 to 0, percent correct bias increases to 0, indicating no difference between the true and estimated expected percent correct scores. As true ability increases for the equivalent distractor, percent correct bias becomes positive. From true abilities of 0 to 2, the equivalent distractor results in underestimates of expected percent correct scores.

The implausible and upper lure distractors follow very similar trends. In Figure 26, the lines overlap. This is especially evident towards the low end of true ability. The implausible and upper lure distractors both result in percent correct biases meaning expected percent correct scores are overestimated. However, the implausible distractor results in less absolute bias in comparison to the upper lure distractor. As true ability increases from -2 to 0, percent correct bias transitions from negative to 0. As true ability increases from 0 to 2, percent correct bias becomes positive for both the implausible and upper lure distractors. However, the implausible distractor results in greater absolute bias than the upper lure distractor.

The lower lure distractor results in the highest absolute bias across true ability. This is especially evident at the ends of the true ability continuum. Low ability examinees receive overestimates of their expected percent correct scores. As true ability increases from -2 to 0, percent correct bias becomes less negative and closer to 0. At a true ability of 0, the lower lure distractor results in no differences between the true and estimated expected percent correct scores. Percent correct bias then becomes positive as true ability increases from 0 to 2. Examinees with a true ability of 2 answering items with lower lure distractors receive underestimates of their expected percent correct scores that are much higher than the other undesirable distractor types.

Around a true ability of 0, all undesirable distractor types result in no percent correct bias. Across the entirety of the true ability continuum, the equivalent distractor results in the lowest absolute percent correct bias, followed by the implausible and upper lure distractors. The lower lure distractor results in the highest absolute percent correct bias across true ability. Toward the lower end of the true ability continuum the undesirable distractor types have greater differences than at the higher end of the true ability continuum. In other words, distractor type has more effect on percent correct bias low ability examinees in comparison to high ability examinees. Guessing could be playing a role in this result.

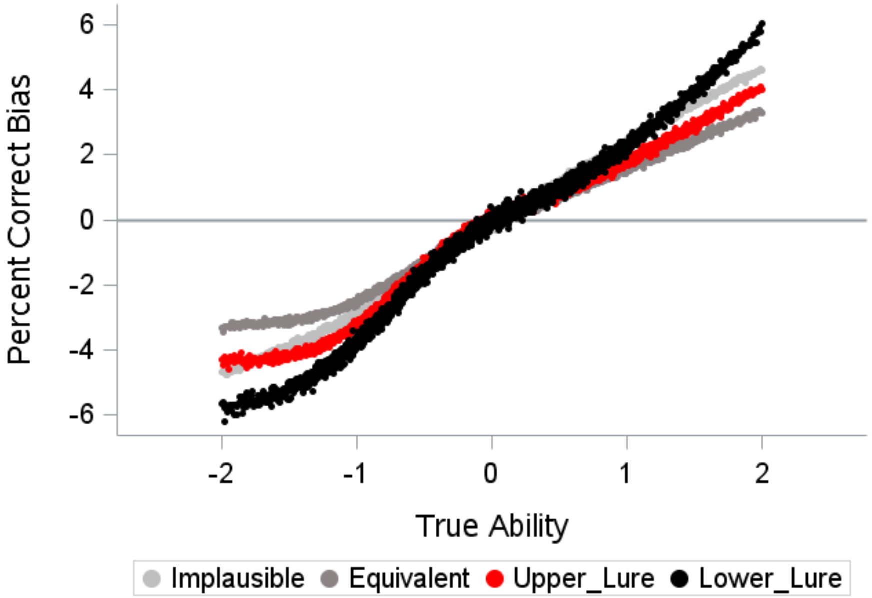

Figure 26. Relationship between percent correct bias and true ability with 10% of items containing each type of undesirable distractors.

Figure 27 displays the relationship between percent correct bias and true ability with 30% of items containing each type of undesirable distractor. The implausible distractor has the lowest absolute percent correct bias across ability, with notable exception at low ability. At a true ability of -2, the implausible and upper lure distractor are approximately equal in terms of percent correct bias. At a true ability of -2, the percent correct bias associated with the upper lure distractor dips below percent correct bias for the implausible distractor. Both distractors follow a similar trend when true ability is 0, where percent correct bias is equal to 0. Although, if we examine a true ability of 2, the implausible distractor results in less percent correct bias in comparison to the upper lure distractor. The equivalent distractor results in higher absolute percent correct bias in comparison to the implausible and upper lure distractors, but it has lower percent correct bias than the lower lure distractor.

If we examine true ability between -2 and -1, all percent correct bias for undesirable distractor types begin to level out. In other words, the percent correct bias is constant between true abilities of -2 and -1. From -1 in a positive direction, percent correct bias becomes increasingly less negative. For true abilities greater than 0, percent correct bias for all undesirable distractor types increase to positive values. Ultimately for high ability examinees, the lower lure distractor results in higher percent correct bias, followed by the equivalent, upper lure, and implausible distractors, respectively.

Similar to the 10% condition, lower values of true ability are overestimated while higher values are underestimated in the 30% condition. However, the 30% condition results in higher overall absolute percent correct bias than the 10% condition. If we examine true ability at 2 in the Figure 27, we observe that the equivalent distractor results in the lower absolute percent correct bias than that of the 10%. In Figure 27, the implausible distractor results in the lowest absolute percent correct bias, but the absolute percent correct bias is higher than the 10% condition.

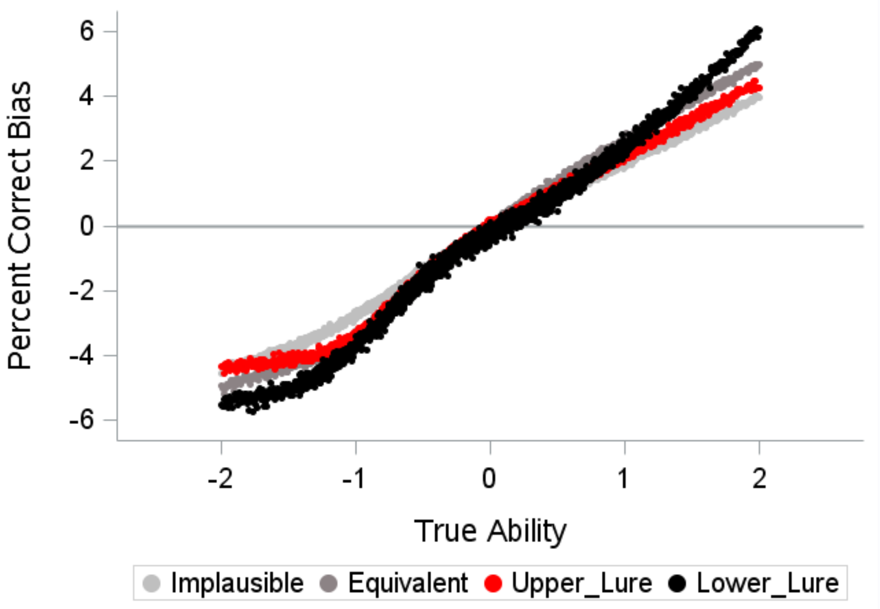

Figure 27. Relationship between percent correct bias and true ability with 30% of items containing each type of undesirable distractors.

Figure 28 displays the relationship between percent correct bias and true ability with 50% of items containing undesirable distractors. The implausible distractor results in the lowest absolute percent correct bias across true ability. The equivalent distractor overestimates ability to a greater degree than the implausible distractor for true abilities between -2 and -1. The equivalent distractor also tends to overestimate expected percent correct score for abilities greater than 0, but it results in the same percent correct bias as the implausible distractor at a true ability of 2. The upper lure and lower lure distractors result in almost equal absolute percent correct biases as the lines for each overlap. The upper lure distractor results in slightly less negative and less positive percent correct bias values across ability than the lower lure distractor, but these differences are negligible.

In comparison to the 30% condition, percent correct bias levels off for the 50% condition at the lower end of the true ability continuum. Between true abilities of -2 and -1, percent correct bias does not change for each distractor type. For example, the equivalent distractor results in constant percent correct bias between true abilities -2 and -

1 (approximately -4%). However, the flat slope of the percent correct bias for undesirable distractor type line discontinues as true ability increases.

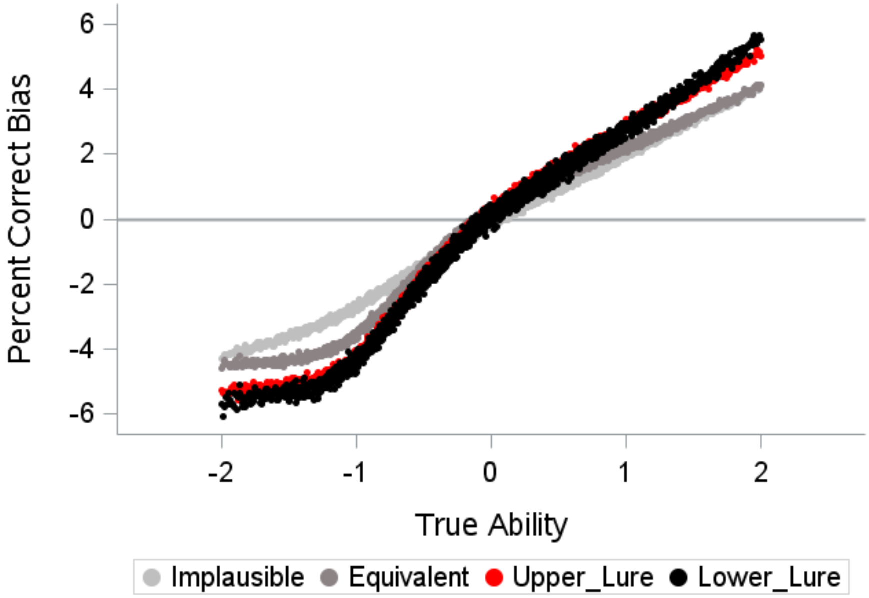

Figure 28. Relationship between percent correct bias and true ability with 50% of items containing each type undesirable distractors.

In summary, the effect of undesirable distractor type on percent correct bias is moderated by the percentage of items containing undesirable distractors, controlling for true ability and test length. Across all levels of percentage, low ability examinees' expected percent correct scores were overestimated while high ability examinees' expected percent correct scores were underestimated. Aside from a true ability at 0, as percentage increased from 10% to 30% to 50%, the absolute percent correct bias increased. The lower lure distractor consistently resulted in the highest absolute percent correct bias across percentage. When 50% of items contained an upper lure distractor, it resulted in especially high absolute percent correct bias. When 10% of items contained an equivalent distractor, the absolute percent correct bias was lowest amongst undesirable distractor type. However, in both conditions where 30% and 50% of items contained an implausible distractor, the absolute percent correct bias was lowest. Overall, when controlling for test length and true ability, as percentage of items containing undesirable distractors increases, percent correct bias also increases with some exceptions depending on the undesirable distractor type.

Percent Correct Bias: Test Length x Type x True Ability Interaction. The second three-way interaction that is practically significant is between test length, type of undesirable distractor, and true ability ($F(6, 71892) = 1311.94, p < .001, \eta_p^2 = .099$). The effect of undesirable distractor type on percent correct bias depends on the test length and true ability after controlling for the percentage of items containing undesirable distractors. In particular, the differences of percent correct bias amongst the undesirable distractor types changes as true ability increases. With the shortest test, the differences among each distractor type are very apparent. However, the differences between the implausible, equivalent, and upper lure distractors diminish when compared to the lower lure distractor. Figures 29, 30, and 31 display the relationship between percent correct bias and true ability by the type of undesirable distractor across each level of test length (i.e., 30 items, 50 items, and 100 items).

Figure 29 reveals that the lower lure distractor results in the greatest absolute percent correct bias for a 30-item test across true ability. The implausible and upper lure distractors result in percent correct bias that is equal when true ability is at -2, but the percent correct bias for the implausible distractor becomes less negative (i.e., less biased) at a faster rate as ability increases. The equivalent distractor results in the lowest absolute percent correct bias at a true ability of -2, but the implausible distractor results in less absolute percent correct bias as a true ability increases from -1.5. Overall, percent correct bias for the all undesirable distractor types overestimate expected percent correct scores for low ability examinees. The lower lure, upper lure, and equivalent distractors' percent correct bias values are consistent between a true ability of -2 and -1. In contrast, the percent correct bias of the implausible distractor becomes continuously increases.

After a true ability of -1, all undesirable distractor types result in percent correct bias that is less negative. At a true ability of 0, there are no differences in percent correct bias; all undesirable distractor types for a 30-item test at a true ability of 0 result in no percent correct bias. As true ability increases, all undesirable distractor types underestimate expected percent correct scores. The lower lure distractor results in the largest percent correct bias when true ability is greater than 0. The implausible and upper lure distractors result in equal percent correct bias values, and the equivalent distractor results in the lowest percent correct bias values when true ability greater than 0.

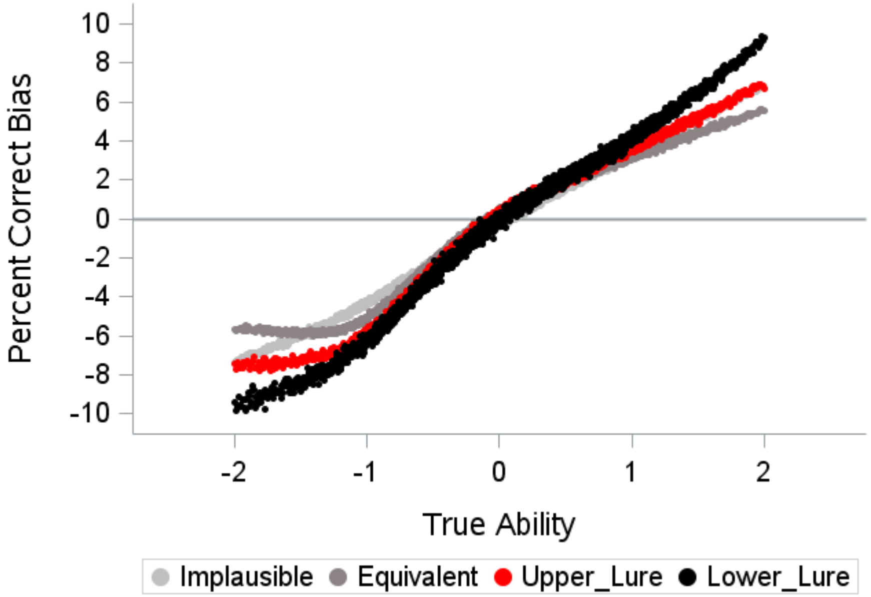

Figure 29. Relationship between percent correct bias and true ability of 30 items for each type of undesirable distractor.

As displayed in Figure 30, with a test length of 50 items, the lower lure distractor results in the largest absolute percent correct bias across true ability. Additionally, across true abilities, the equivalent, upper lure, and implausible distractors result in very similar absolute percent correct bias. At a true ability of -2, the equivalent distractor results in higher absolute percent correct bias values in comparison to the upper lure and implausible distractors. However, as true ability increases to 2, the equivalent, upper lure, and implausible distractors result in equal overestimation (when true ability is less than 0) and underestimation (when true ability is greater than 0) of expected percent correct scores.

With the lower lure distractor, we see constant percent correct bias for true abilities between -2 to -1, similar to as we did with the 30-item condition. The upper lure and equivalent distractors are also relatively flat, but this is not as pronounced as in the 30-item condition. The implausible distractor results in percent correct bias that continuously becomes less negative as true ability approaches 0.

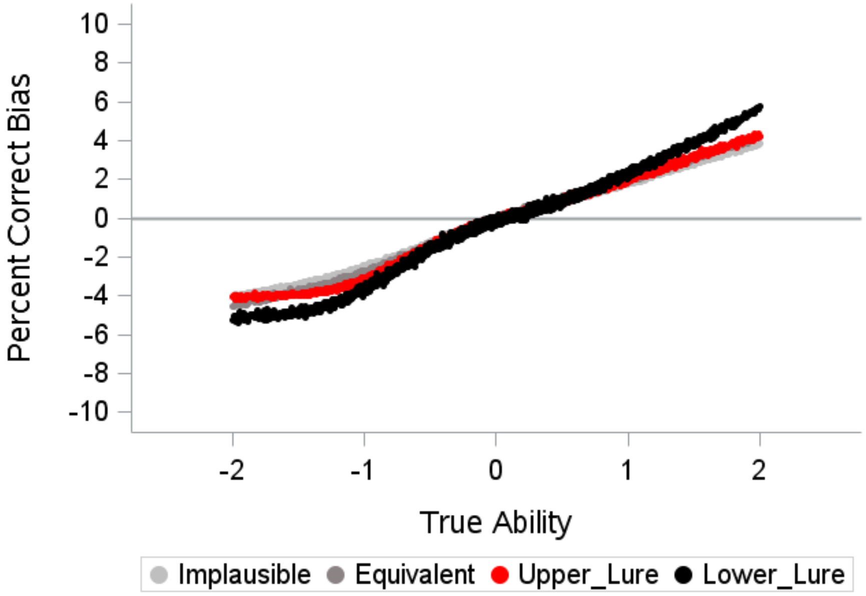

Figure 30. Relationship between percent correct bias and true ability for 50 items for each type of undesirable distractor.

Figure 31 displays the relationship between average percent correct bias over replications and true ability for a 100-item test grouped by undesirable distractor type. There are very slight differences between the undesirable distractor type, but these differences are negligible. Compared to the 30-item and 50-item tests, the 100-item test results in the least amount of percent correct bias across the true ability continuum.

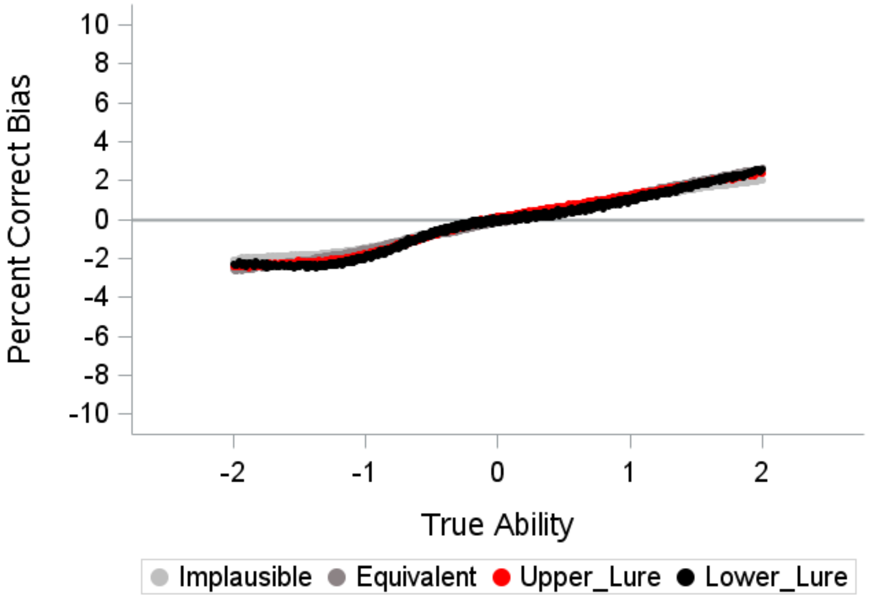

Figure 31. Relationship between percent correct bias and true ability for 100 items for each type of undesirable distractor.

Overall, expected percent correct scores for low ability examinees is overestimated while expected percent correct scores for high ability examinees is underestimated. The 30-item test results in the highest absolute percent correct bias. The lower lure results in the highest absolute percent correct bias for the 30-item and 50-item test, but the distractor tends to produce similar amounts of absolute percent correct bias in comparison to the other distractor types for the 100-item test. While the equivalent distractor results in the least amount of bias for the 30-item test, the implausible, upper lure, and equivalent distractors result in equal amounts of absolute percent correct bias for the 50-item test. The 100-item test results in the least amount of percent correct bias for all types of undesirable distractors.

Standard Error

Standard error of true expected percent correct score and estimated expected percent scores, hereafter referred to as standard error, was analyzed using an ANCOVA to examine the effects of test length, percentage of items containing undesirable distractors, and type of undesirable distractor with a covariate of true ability. Squared true ability was also included in the model to examine if the relationship between standard error and true ability had a curvilinear relationship. Table 26 presents the results of the standard error ANCOVA.

Table 26
ANCOVA results for standard error

Source	df	SS	MS	F	*p*	Partial η^2
L	2	28635.98	14317.99	723062	<.0001	**.953**
P	2	105.07	52.53	2653.01	<.0001	**.069**
T	3	1545.87	515.29	26022.3	<.0001	**.521**
A	1	139.32	139.32	7035.62	<.0001	**.089**
A2	1	641.84	641.84	32413.1	<.0001	**.311**
L x P	4	30.10	7.52	380.00	<.0001	.021
L x T	6	429.82	71.64	3617.65	<.0001	**.232**
L x A	2	22.68	11.34	572.71	<.0001	.016
L x A2	2	46.87	23.43	1183.38	<.0001	.032
P x T	6	339.66	56.61	2858.85	<.0001	**.193**
P x A	2	1258.57	629.28	31779.0	<.0001	**.469**
P x A2	2	13.64	6.82	344.46	<.0001	.010
T x A	3	774.09	258.03	13030.6	<.0001	**.352**
T x A2	3	201.11	67.04	3385.42	<.0001	**.124**
L x P x T	12	82.12	6.84	345.61	<.0001	.055
L x P x A	4	94.15	23.54	1188.63	<.0001	**.062**
L x P x A2	4	0.88	0.22	11.06	<.0001	.001
L x T x A	6	61.36	10.23	516.42	<.0001	.041
L x T x A2	6	18.47	3.08	155.42	<.0001	.013
P x T x A	6	178.55	29.76	1502.78	<.0001	**.111**
P x T x A2	6	41.16	6.86	346.40	<.0001	.028
L x P x T x A	12	9.43	0.79	39.67	<.0001	.007
L x P x T x A2	12	5.04	0.42	21.19	<.0001	.004

Note. Test length = L; Percentage = P; Undesirable distractor type = T; True ability = A; True ability^2 = A2; practically significant effects are bolded.

Standard Error: Insignificant Effects. The effect of test length on standard error does not depend on the percentage of items containing undesirable distractors, controlling for true ability and type of undesirable distractor ($F(4, 71892) = 380.00, p < .001, \eta_p^2 = .021$). The effect of test length on standard error also does not depend on true ability ($F(2, 71892) = 572.71, p < .001, \eta_p^2 = .016$) or true ability squared ($F(2, 71892) = 1183.38, p < .001, \eta_p^2 = .032$), controlling for percentage of items containing undesirable distractors and type of undesirable distractor. The effect of percentage on standard error does not depend on true ability squared, controlling for test length and type of undesirable distractor ($F(2, 71892) = 344.46, p < .001, \eta_p^2 = .010$).

Six of the three-way interactions were not practically significant. Standard error does not differ among the varying levels of the interaction between test length, percentage, and undesirable distractor type, controlling for true ability ($F(12, 71892) = 345.61, p < .001, \eta_p^2 = .055$). Standard error also does not differ among varying levels of the interaction between test length, percentage, and true ability squared, controlling for type of undesirable distractor ($F(4, 71892) = 11.06, p < .001, \eta_p^2 = .001$). The interactions of test length and type of undesirable distractor with true ability ($F(6, 71892) = 516.42, p < .001, \eta_p^2 = .041$), as well as with true ability squared ($F(6, 71892) = 155.42, p < .001, \eta_p^2 = .013$), were not practically significant. Finally, standard error does not differ among the varying levels of the interaction between percentage, undesirable distractor type, and true ability squared, controlling for test length ($F(6, 71892) = 346.40, p < .001, \eta_p^2 = .028$).

Both four-way interactions were not significant. The interaction between test length, percentage, undesirable distractor type, and true ability does not have a practical effect on standard error ($F(12, 71892) = 39.67, p < .001, \eta_p^2 = .007$). Additionally, the interaction of test length, percentage, undesirable distractor type, and true ability squared does not have a practical effect on the standard error ($F(12, 71892) = 21.29, p < .001, \eta_p^2 = .004$).

Standard Error: Practically Significant Effects. All main effects were practically significant. For example, the effect of test length on standard error explained the largest proportion of variance, after partialing out the variance explained by all other effects ($F(2, 71892) = 723062, p < .001, \eta_p^2 = .953$). Although still practically significant, the effect of percentage on standard error explained the smallest proportion of variance of all main effects, after partialing out the variance explained by all other effects ($F(2, 71892) = 2653.01, p < .001, \eta_p^2 = .069$).

Five of the two-interaction were practically significant. The effect of percentage on standard error that depends on true ability explained the most variance of the two-way interactions, after partialing out the variance explained by all other effects ($F(2, 71892) = 31779.0, p < .001, \eta_p^2 = .469$). The effect of test length on standard error does depend on true ability squared, controlling for percentage and undesirable distractor type ($F(3, 71892) = 3385.42, p < .001, \eta_p^2 = .124$). Two of the three-way interactions were practically significant.

Standard Error: Length x Percentage x True Ability Interaction. I first examine the interaction between length, percentage, and true ability ($F(4, 71892) = 1188.63, p < .001, \eta_p^2 = .062$). The differences in standard error for each test length become more variable as the percentage of items containing undesirable distractors increases. In particular, the shortest test results in the greatest amount of standard error while the longest test results in the least amount of standard error. Figure 32 shows the relationship between standard error and true ability for 10% of items containing undesirable distractors grouped by test length. A test of 30-items with 10% of items containing undesirable distractors results in the largest amount standard error across true ability, a 50-item test results in the second largest amount of standard error, and a 100-item test results in the least amount of standard error. Standard error is consistent for all three test lengths for true abilities less than 0. The resulting standard error for a 30-item increases slightly at a true ability of 0, but standard error decreases as true ability increases to 2. The standard error for a 50-item test follows a similar pattern as the standard error for a 30-item test. However, the slight increase in standard error at a true ability of 0 is not as pronounced. A 100-item test results in standard error that remains constant from -2 to 0, but then decreases as true ability increases from 0.

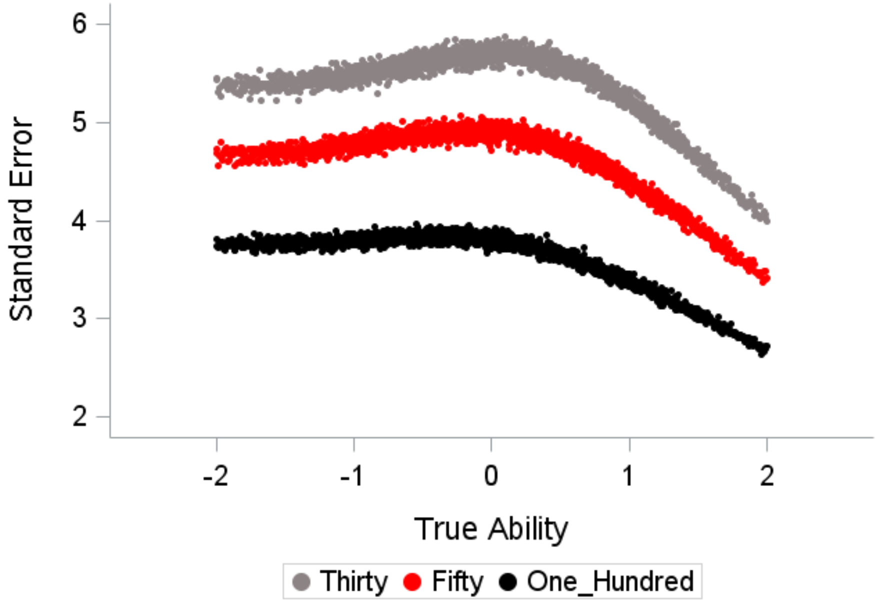

Figure 32. Relationship between standard error and true ability with 10% of items containing undesirable distractors for each test length.

Figure 33 displays the relationship between standard error and true ability when 30% of items contain undesirable distractors for test length. At a true ability, each test length results in similar trends of standard error. Along the true ability continuum, standard error decreases between true abilities -2 and -1, increases between true abilities of -1 and .5, and decreases between .5 and 2.

In comparison to a test where 10% of items contain undesirable distractors, the 30% condition produces increasingly varying levels of standard error across true ability for each level of test length. While true abilities of -2 share similar values of standard error for each test length for the 10% and 30% conditions, true abilities of -1 have much less standard error standard error for each test length in the 30% condition than in the 10% condition. However, when true ability is near 2, there is larger standard error in the 30% condition then the 10% condition for all test lengths.

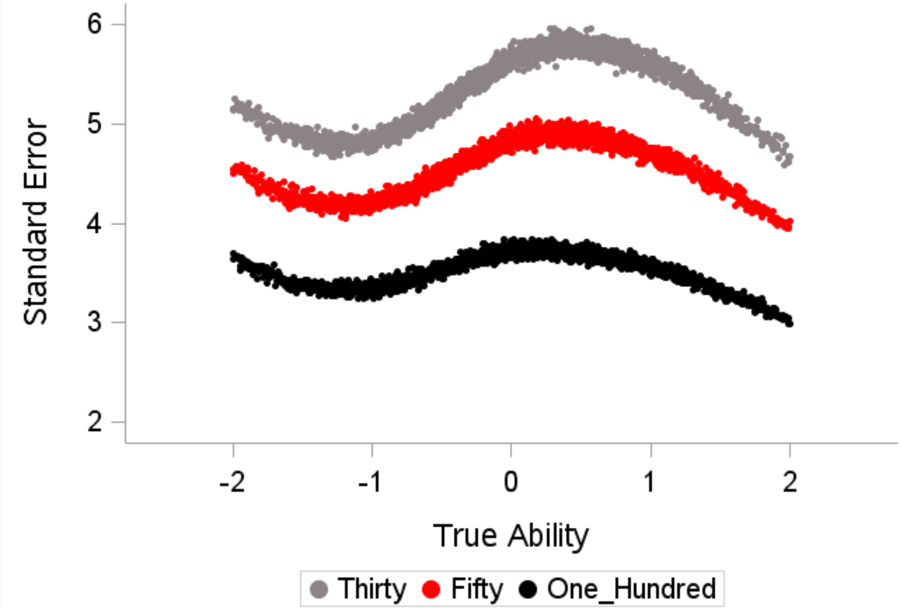

Figure 33. Relationship between standard error and true ability with 30% of items containing undesirable distractors for each test length.

Figure 34 shows the relationship between standard error and true ability when 50% of items contain undesirable distractors grouped by test length. Similar to 10% and 30% conditions, all levels of test length follow similar trends. However, in contrast to the 10% and 30% conditions, true abilities near -2 have less standard error across all test lengths, and true abilities near -1 for the 50% condition also have less standard than the other percentage conditions. The maximum for standard error in the 50% condition occurs near a true ability of .5, and is greater than the standard error in the 10% and 30% conditions at similar abilities. Finally, true abilities near 2 for the 50% condition result in higher standard error compared to the same abilities in the 10% and 30% conditions for a given test.

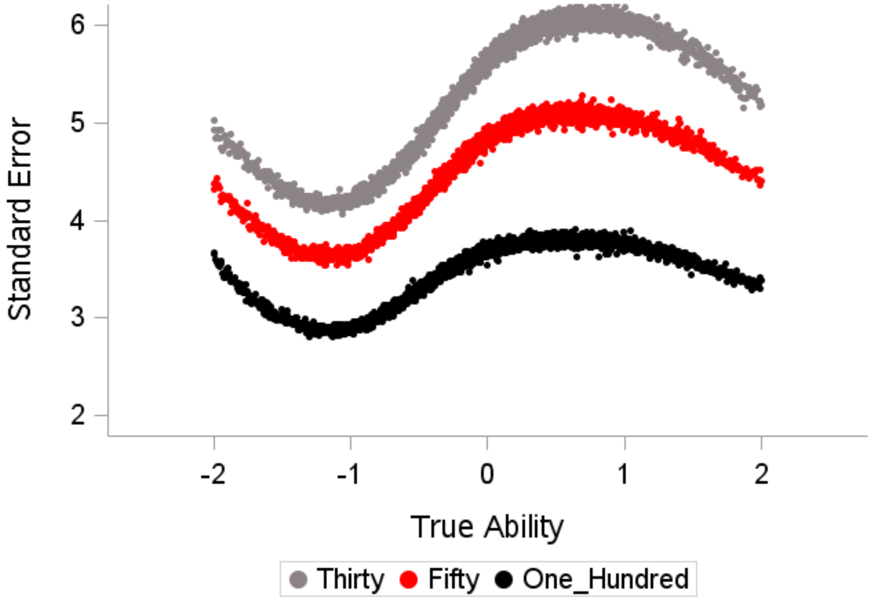

Figure 34. Relationship between standard error and true ability with 50% of items containing undesirable distractors for each test length.

Overall, the relationship between standard error and true ability is moderated by the percentage of items containing undesirable distractors and test length, controlling for type of undesirable distractor. As test length increases, less standard error is produced across true ability. The relationship between standard error and true ability, however, is dependent upon the percentage of items containing undesirable distractors at various levels of test length. A test containing 50% of items with undesirable distractors produces more extreme values of standard error across the true ability continuum. While a test where 10% of items contain undesirable distractors has a standard error that decreases as true ability increases, the 30% and 50% conditions change across true ability. Finally, a 100-item test produces the largest amount of standard error and changes the most as a function of true ability.

Standard Error: Type x Percentage x True Ability Interaction. The interaction between undesirable distractor type, percentage, and true ability also had a practically significant effect on standard error ($F(6, 71892) = 1502.78, p < .001, \eta_p^2 = .111$). As the percentage of items containing undesirable distractors increases, the differences in the amount of standard error for each undesirable distractor type varies across true ability. The relationship between standard error and true ability is displayed in Figure 35, where 10% of items contain each type of undesirable distractor. Across ability, the implausible distractor results in the least amount of standard error. A test where 10% of items contain a lower lure distractor results in a lower amount of standard error in comparison to the implausible, upper lure, and equivalent distractors when ability is less than -1. However, between a true ability of 0 and 2, the lower lure distractor results in higher standard error than the other three undesirable distractors types. For ability less than 0, the equivalent distractor, followed by the upper lure distractor, results in the largest amount of standard error, but the relationships between standard error and true ability decreases as true ability becomes greater than 0. Compared to low ability examinees, when ability is greater than 0, lower standard errors occur for all undesirable distractors when only 10% of items contain undesirable distractors.

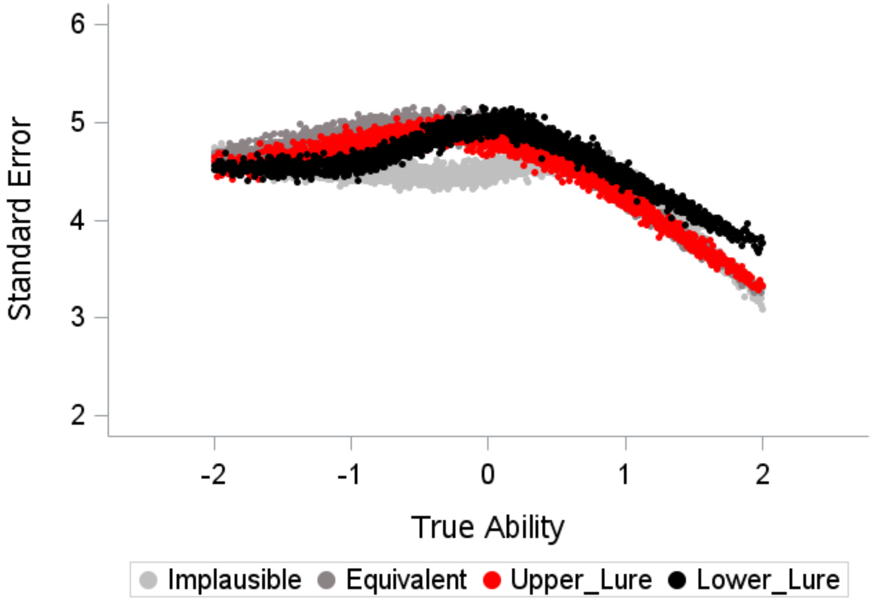

Figure 35. Relationship between standard error and true ability with 10% of items containing each type undesirable distractors.

Figure 36 displays the relationship between standard error and true ability with 30% of items containing an undesirable distractor. The implausible and lower lure distractors change the most with respect to standard error across true ability when comparing back to the 10% condition. For ability near -2, the implausible distractor results in the greatest amount of standard error. As true ability increases from -2 to -1, the standard error for the implausible distractor decreases, similar to the lower lure distractor. However, for these distractor types, standard error then increases as ability increases from -1 to 0. Standard error for the implausible distractor peaks at a true ability of 1 but decreases as true ability increase. In contrast to the implausible distractor, standard error for the upper lure distractor peaks at a true ability of 0 and decreases as true ability increases toward 2.

In comparison to the 10% condition, there is a larger difference in standard error when 30% of items on a test contain the equivalent distractor and 30% of items contain the upper lure distractor. When true ability is between -2 and 0, the equivalent distractor results in a greater amount of standard error than the upper lure distractor. However, this difference in standard error dissipates as true ability increases from 0 to 2.

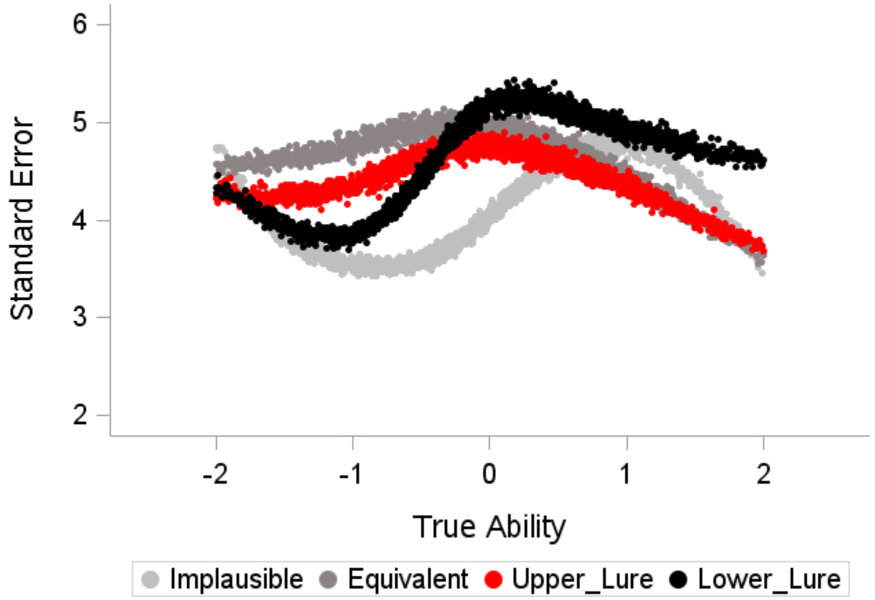

Figure 36. Relationship between standard error and true ability with 30% of items containing each type undesirable distractors.

The relationship between standard error and true ability when 50% of items contain an undesirable distractor is shown in Figure 37. While standard error resulting from tests with implausible distractors is greater than standard error from tests with any other type of distractor when true ability is near -2, it results in the least amount of standard error for abilities between about -1.7 to .7. The standard error resulting from the implausible distractor is greater than the resulting standard error from the upper lure and equivalent distractors for abilities greater than .7. In general, the standard error resulting from tests with 50% of items having implausible distractors follows a sinusoidal pattern. On average, the implausible distractor results in the least amount of standard error across true ability.

The lower lure distractor follows a similar sinusoidal pattern as the implausible distractor, but it tends to result in greater standard error across the true ability continuum.

This curvilinear pattern ceases for abilities greater than 0. After reaching the point of maximum standard error at an ability of 0, standard error remains flat and relatively high as ability increases to an ability of 2. Standard error for the lower lure distractor then decreases slightly as true ability approaches 2.

In comparison to the 10% and 30% conditions, the equivalent and upper lure distractors have even greater differences in standard error when 10% of items contain an undesirable distractor type. While the equivalent distractor for the 50% condition results in similar amounts of standard error as in the 30% condition, the upper lure distractor results in lower amounts of standard error for the 50% condition in comparison to the 30% condition. Even though the equivalent distractor and the upper lure distractor result in similar sinusoidal patterns followed by flattening out standard error patterns as the lower lure, the curvilinear pattern's maximum and minimum differences are not as dramatic.

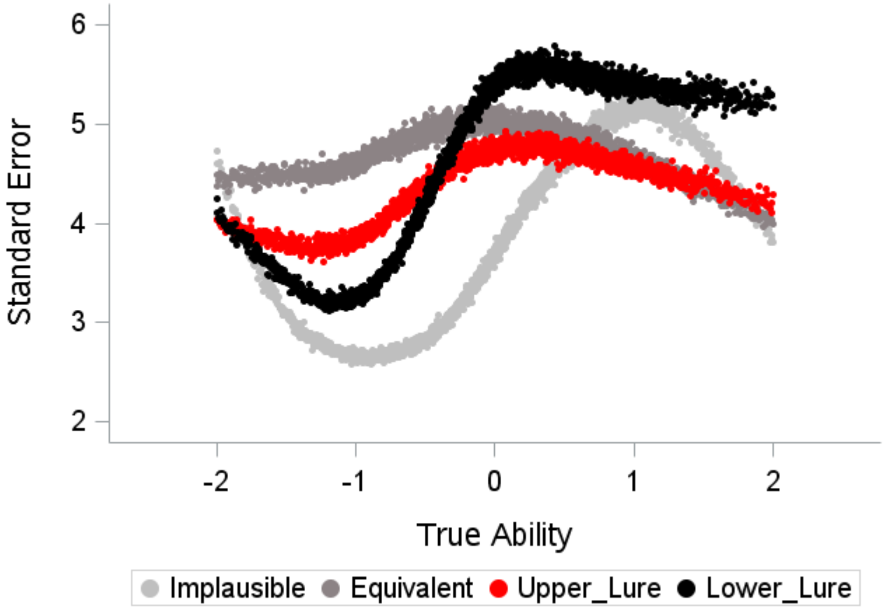

Figure 37. Relationship between standard error and true ability with 50% of items containing each type undesirable distractors.

Overall, the relationship between standard error and true ability differs with respect to the percentage of items containing undesirable distractors and the type of undesirable distractor, controlling for test length. For each undesirable distractor type, the relationship between standard error and true ability of a test where 10% of items contain the undesirable distractor tends to peak at a true ability of 0 but decrease as ability increases. The lower lure distractor, on average, results in the largest amount of standard error, while the implausible distractor results in the lowest amount of standard error across ability. The equivalent and upper lure distractors tend to remain the same with respect to standard error across true ability, but the amount of standard error tends to increase as the percentage of items containing an undesirable distractor increases.

Chapter 5. Discussion

The purpose of this study was to determine how undesirable distractors impact estimates of ability. In the following sections, I discuss the results of my research questions, provide recommendations for test developers based on these results, outline limitations of this study, and consider next steps for this line of research.

Research Questions

Parameter Recovery. I first examined the capabilities of recovering ability estimates using the 3-PL IRT model in SAS 9.4 (TS1M4). Previous research by Cole and Paek (2017) also employed a simulation study to evaluate the recovery of parameters using the IRT procedure, but the authors focused on item parameters rather than ability parameters. In order to address this question, I first generated data for three test lengths, 30 items, 50 items, and 100 items, each for a sample of 2000 examinees. Difficulty, discrimination, and pseudo-guessing item parameters of the 3-PL were randomly sampled in each replication to improve generalizability of the simulation results. Initially, no distractor properties were considered. Rather, I first explained recovery in baseline conditions. To analyze recovery of ability parameters in the default estimation routines of the IRT procedure in SAS, I calculated bias and standard error of ability over 1000 replications. Because the distractor analysis would eventually require us of different generating models from analysis models, based on suggestions from Leuhct and Ackerman (2018), I used expected total score as a proxy for true and estimated ability. Further, to make results comparable across varying test lengths, expected total score was converted to expected percent correct score. I examined the differences between ability estimates and true ability averaged over replications. Standard error was calculated as way to examine the spread of true ability across replications. In other words, I looked at the spread of the ability estimates to gain information about the efficiency of these estimates.

I begin by discussing the bias of the ability estimates for the three baseline conditions of varying test length. In general, bias tended to be greatest toward the tails of true ability. This is likely due to the way in which I generated the item parameters. Difficulty was randomly sampled from a standard normal distribution. Ability and difficulty are on the same scale, meaning that we have the most test information around a true ability of 0 since this the mean of the sampled difficulty distribution. More information near a true ability of 0 is associated with less bias. Therefore, we see more bias toward the tails of true ability and less bias in the middle of true ability.

Bias was different at each level of test length and across true ability. Longer tests (i.e., 100 items) resulted in less biased ability estimates. In contrast, shorter tests (i.e., 30 items) resulted in more biased ability estimates. On average, the ability estimates for a shorter test, over 1000 replications, are further away from true ability than that of the longer tests across the entirety of the ability continuum. This finding is consistent with previous research that a longer test will result in less biased ability estimates (De Ayala, 2013).

Additionally, guessing may have an impact on the amount of bias that each test length resulted in. A longer test resulted in a constant slope for bias as true ability increased. In contrast, the 30-item and 50-item tests result in bias for which the rate of change is a function of true ability. The 30-item test resulted in a negative bias that remained constant from true abilities of -2 to -1, but after -1, bias increased linearly as true ability increased. We can observe a similar trend with the 50-item test. It is likely that estimation of low abilities is impacted by guessing. Guessing may have contributed to the slight increase in negative bias for true abilities of -2 to -1.

Similar to bias, standard error differed at each level of test length and across true ability. Longer tests resulted in the least amount of standard error across true ability. While shorter tests resulted in the greatest standard error across true ability, standard error tended to be lowest for a true ability in the tails and greatest for a true ability near 0. This may seem like a contrary result as in the IRT paradigm standard error tends to be lowest for abilities where we have the greatest information. we see that standard error is highest for the true abilities where we have the most information. This pattern is occurring because expected percent correct score was used as a proxy for ability. Thus, we see evidence of similar floor and ceiling effects that occur for standard error in CTT.

Similar to Cole and Paek’s (2017) evaluation of the IRT procedure in SAS for item parameters, ability parameters tended to be unbiased and efficient. However, shorter tests result in more biased and less efficient ability estimates. I recommend the use of the IRT procedure in SAS to recover ability parameters, but I would suggest that test developers attempt to create longer tests for less biased and more efficient ability estimates.

Bias: Undesirable Distractors. The remaining three research questions address concerns about the effects that the percentage of items containing undesirable distractors, the test length, and the type of undesirable distractor have on ability estimates. Overall, lower levels of ability tended to be overestimated while higher levels of ability tended to be underestimated. This occurred because I used MAP to estimate abilities, which draws abilities in towards the mean of ability. First, I investigated the interaction among these factors on bias of abilities. There were two practically significant three-way interactions: (1) percentage, undesirable distractor type, and true ability and (2) test length, undesirable distractor type, and true ability. I discuss each interaction below.

The first interaction of percentage, undesirable distractor type, and true ability indicates that the bias of ability estimates is dependent on what distractor type, the percentage of items containing the distractor, and the true ability of the examinee. The lower lure distractor consistently resulted in ability estimates that contained the most bias. Bias for ability estimates associated with the lower lure distractor tended to increase with percentage, indicating that the more items containing a lower lure distractor, the more bias there is in estimating examinee ability.

As the percentage of items containing undesirable distractors increased (10% to 30% to 50%), bias in ability estimates increased. However, the rank ordering of the type of undesirable distractor resulting in the least amount of bias was not consistent across percentages. The implausible distractor resulted in the least amount of bias in ability estimates for the 30% and 50% conditions but not the 10% condition. A surprising result is that a distractor that produces no information results in the least amount of bias for the 30% and 50% conditions. Upon further investigation, this result is likely due to the other distractor in the item. Figure 21 provides the item characteristic choice curves for a multiple-choice item containing an implausible distractor. The second distractor (option C) was defined to identify the model given item parameter constraints. However, this distractor is considered an informative distractor (Samejima, 1988). A distractor taking this form has been evidenced to provide discrimination information for examinees with abilities in the center of the distribution. While the implausible distractor does not discriminate amongst examinees' abilities, the inclusion of this informative distractor is likely the cause of the lowest absolute bias for the test with the implausible distractor. In other words, the addition of this desirable informative distractor is likely the reason ability estimates are closer to true ability in comparison to the other undesirable distractor conditions.

The equivalent distractor resulted in varying degrees of bias across true ability for each level of percentage. When 10% of items contained the equivalent distractor, bias was lowest across true ability. Although, the equivalent distractor resulted in greater amounts of bias compared to the implausible distractor as percentage increased. The upper lure distractor follows a similar trend, but the upper lure distractor resulted in less bias than the equivalent distractor for the 10% and 30% conditions and more bias than in the 50% condition. Overall, both the equivalent and upper lure distractors resulted in greater amounts of bias than the implausible distractor; however, both distractors resulted in less bias than that of the lower lure distractor.

There was also a practically significant three-way interaction effect on bias between test length, undesirable distractor type, and true ability. In other words, the effects of test length on bias differed for undesirable distractor types and across true abilities, controlling for percentage. Similar to the baseline conditions, a longer test resulted in less bias and the shortest test resulted in more bias. Trivial differences existed in the resulting bias between the baseline conditions and undesirable distractor conditions for all undesirable distractor types when the test contained 100 items. In other words, if the test is long, the effect of undesirable distractors in items tends to be minimized.

Concerning the shorter tests (i.e., 30 items and 50 items), undesirable distractor types did result in varying degrees of bias from baseline. The lower lure distractor is especially problematic on shorter tests; it results in much greater bias in ability than the other undesirable distractors for a test of 30. The lower lure distractor also results in the largest amount of bias for the 50-item test in comparison to the other undesirable distractor types but less bias than that of the 30-item test. The implausible distractor resulted in the least amount of bias for the 30-item and 50-item tests. However, differences between the implausible distractor and equivalent and upper lure distractors were negligible for the 50-item test. Regardless of the positive or negative effects of distractor on bias, these effects on ability estimates tend to disappear when there is a greater amount of items.

Standard Error: Undesirable Distractors. There were two significant three-way interactions when examining efficiency of ability estimates: (1) percentage, undesirable distractor type, and true ability, and (2) percentage, test length, and true ability.

The first interaction shows that the effect of percentage on standard error differs across undesirable distractor type and true ability, controlling for test length. Generally, tests containing items with higher percentages of undesirable distractors result in higher standard error of estimates across the true ability spectrum. However, this observation is not necessarily true for the implausible distractor. Standard error actually decreases as the percentage of items containing an implausible distractor increases across true ability. Across replications, ability estimates are more similar to one another when more items contain the implausible distractor. This may be due to the informative distractor. The equivalent and upper lure distractors result in somewhat similar amounts of standard error across true ability and percentage. The lower lure distractor tends to result in the largest amount of standard error across true ability, especially as the percentage of items containing undesirable distractors increases.

The second interaction was between percentage, test length, and true ability. The effect of percentage on standard error differs across test length and true ability, controlling for undesirable distractor type. The amount of standard error associated with true ability is not constant across all levels of test length and increasing percentage of items containing undesirable distractors. A test where 10% of items contain undesirable distractors results in the greatest amount of standard error for abilities at the center of the continuum and the least amount of standard error for high ability examinees. Although we have most information toward the center of the ability continuum, I calculated standard error on the percent correct metric. This is the reason we see a ceiling effect for standard error at the center of ability. A test where 30% and 50% of items contain undesirable distractors results in the largest amount of standard error for true abilities greater than an ability of 0 and the least amount of standard error for true abilities less than an ability of 0. In other words, as percentage increases, ability estimates are both less and more accurate across the continuum of ability. This pattern is likely the result of guessing. We expect varying percent correct scores for low ability when examinees are able to guess the correct answer. Therefore, as seen with higher standard errors, we expect more variability in ability estimates for these low ability examinees.

Recommendations

In general, I recommend that test developers should pay careful attention to the functioning of distractors. Many researchers have expressed that distractors are an important part of the multiple-choice item (Gierl, Bulut, Guo, & Zhang, 2017; Sideridis, Tsaousis, & Harbi, 2016; Thissen, Steinberg, & Fitzpatrick, 1989). I believe that this research supports that argument. The implausible distractor contained a second distractor that was informative. An informative distractor is one that divides examinees into groups based upon the selection of the options (Thissen, Steinberg, & Fitzpatrick, 1989; Samejima, 1988). For example, low ability examinees have equal probability of selecting all the choices, examinees in the middle of the ability distribution have the highest probability of selecting the informative distractor, and high ability examinees have the highest probability of selecting the correct answer. In the current study, conditions with the implausible distractor resulted in the least biased and most efficient ability estimates across the ability continuum. This was likely due to the inclusion of the informative distractor in the implausible distractor condition when there is a clear separation of groups of examinees based on which choice they select. Therefore, the implausible distractor could be alright to include in multiple-choice items but it is the the second distractor being informative that is crucial.

I suggest that test developers should be aware of lower lure distractors in their multiple-choice items. The lower lure distractor consistently resulted in the greatest bias and standard error. In other words, the inclusion of a lower lure distractor resulted in ability estimates that were both further from an examinees' true ability and more variable over replications in comparison to the implausible, equivalent, and upper lure distractors.

For this reason, test developers should certainly be hesitant in interpreting an examinees ability estimate when a lower lure distractor is present, especially for a shorter test.

Test developers should use longer tests when possible. De Ayala (2013) agrees with this finding. If a test does contain any type of undesirable distractor, the negative effects on ability estimates are lessened with the inclusion of more items. A 100-item test always resulted in less bias and more efficient estimates of ability. In fact, the 100-item test containing undesirable distractors had very similar results of bias and standard error to the baseline condition of a 100-item test that contained well-functioning items.

While test developers can examine distractor functioning by performing an item analysis, this is completed after the creation of the distractors. If a test contains a large percentage of undesirable distractors, especially for a short test, there are bound to be biased and inefficient ability estimates. We want scores that are best representative of an examinees' ability, but the inclusion of undesirable distractors would not allow for this. This issue is especially problematic when the test is high-stakes. If the examinee's score is not representative of their true ability due to the inclusion of undesirable distractors during the scoring of the test, then this has the possibility of drastically effecting that examinee's future.

However, how are we to create good distractors if we are unable to evaluate their functioning until after examinees have completed the test? I recommend that test developers focus on Haladyna et al.'s (2002) item-writing rules. In particular, there are 14 guidelines that pertain to the writing of multiple-choice item options. Not following these guidelines could align with the writing of certain undesirable distractors. For example, Haladyna et al. (2002) suggest having only one right answer. This guideline might seem obvious, but certain distractors may be “somewhat” correct. If a “somewhat” correct distractor draws in some of the examinees with a high ability, then we might see a lower lure distractor on our test.

Limitations and Next Steps

The current study does provide evidence that undesirable distractors can have effects on ability estimates. However, there are some limitations of this study that should be addressed. In addition to limitations, I discuss next steps for this line of research. To begin, I used a fixed sample size (*N*=2000) for all conditions. Typically, item parameters are most affected by large sample size, but if I had included smaller samples, how would this effect the bias and efficiency of the results? The large sample in this study allows for the generalization of results to large-scale assessment, but not all multiple-choice tests are taken by large samples of examinees. For example, investigation with a smaller sample would allow for the generalization of results to classroom assessments.

I also fixed the number of options to three. Although research suggests that the optimal number of options in multiple-choice items is three (Haladyna, Rodriguez, & Stevens, 2019; Baghaei & Amrahi, 2011; Vyas & Supe, 2008; Rodriguez, 2005; Crehan, Haladyna, & Brewer, 1993; Haladyna & Downing, 1988), this is not necessarily a requirement. Considering four or more options would allow better generalization of results to more multiple-choice testing situations.

In this study, I only considered undesirable distractor types. Although research is limited regarding undesirable distractors types, there may be others that are common but have yet to be investigated. Besides undesirable distractors negatively impacting ability estimates, good distractors, such as an informative distractor, may result in less biased and more efficient estimates of ability. The implausible distractor condition alluded to this. It was likely positive effects of the informative distractor, not the effects of the implausible distractor that resulted in low bias and standard error. This effect and other positive effects necessitate further exploration.

Aside from the multiple-choice model (Thissen & Steinberg, 1984) to generate undesirable distractors, other models could have been used such as Bock's (1972) nominal response model. I also fixed the guessing parameter for the multiple-choice model, meaning that guessing was immediately assumed for low ability examinees with equal likelihood of each choice. This may not generalize to real-world testing situations. Perhaps guessing is not a concern for some test developers. In this case, the nominal response model would be appropriate to use compared to Thissen and Steinberg's multiple-choice model. Furthermore, in this study, I analyzed item responses using the 3-PL model. Using the multiple-choice model to estimate ability may result in differing results.

Additionally, each replication within each condition had randomly sampled item parameters. This introduced more random error that limits the generalization of these results. With varying item parameters for each condition, it would have been appropriate to equate the tests for all replications in each condition. However, I assumed that all replications in a condition were equal. In general, shorter tests generally have larger standard errors (i.e., shorter tests impact standard errors more than longer tests). The difference in standard error evidenced between the longer tests and shorter tests may be overestimated. This is because the randomly selected item parameters for a shorter test are less likely to "even" out randomly. Larger tests are not as highly impacted by the varying item parameters across replications, but there is still some random error which causes standard error to be greater than it should be. Thus, the differences in standard error between long tests and short tests are likely to exist, but may not be as dramatic as evidenced here.

It is also important to investigate what steps test developers can take to investigate each type of undesirable distractor. I suggest using IRT methods for evaluating distractor functioning, but small-scale multiple-choice tests are unable to use such advanced methods. In situations with smaller samples, it is still beneficial to use trace plots to examine undesirable distractor functioning. For example, Haladyna and Downing (1993) used trace plots to identify undesirable distractors. Although the current research does not address the examination of distractor functioning in CTT, it is an important consideration for testing situations with small samples. Perhaps undesirable distractors impact ability estimates in different ways in the CTT framework compared to the IRT framework.

Although there are some limitations of this study, I believe that these results are best generalized to large-scale assessment testing situations where multiple-choice tests are administered to large samples of examinees (i.e., N=2000). This research supports that there are negative impacts of undesirable distractors on estimates of ability. Although I am hesitant to generalize to other testing situations (e.g., smaller samples, variable guessing, etc.), I believe that these results support the argument that distractors are an important part of the multiple-choice item. So much so that poor functioning distractors negatively impact ability estimates.

References

Aiken, L. R. (1982). Writing multiple-choice items to measure higher-order educational objectives. *Educational and Psychological Measurement*, *42*(3), 803-806.

Allen, S., & Sudweeks, R. R. (2001, April). *Identifying and Managing Local Item Dependence in Context-Dependent Item Sets*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). New York: Longman.

Andrés, A. M., & del Castillo, J. D. (1990). Multiple-choice tests: Power, length, and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology, 43*, 57–71.

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*(2), 153-170.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109-128.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*(2), 192.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York, NY: Guilford Publications.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician*, *56*(4), 299-303.

Beaujean, A. A. (2018). Simulating data for clinical research: A tutorial. *Journal of Psychoeducational Assessment*, *36*(1), 7-20.

Bennett, R. E. (1991). On the meanings of constructed response. *ETS Research Report Series*, *1991*(2), 1-46.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy, & Practice, 18*(1), 5-25.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*(1), 77-92.

Biggs, J. (1999). *Teaching for quality learning at university*. Buckingham, UK: Society for Research into Higher Education and Open University Press.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement,* 11(4), 385-395.

Black, P. & Wiliam, D. (2003). In praise of educational research: formative assessment. *British Educational Research Journal, 29*(5), 623-637.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: new roles, new means: The 63rd yearbook of the National Society for the Study of Education (part II)* (Vol. 69(2), pp. 26-50). Chicago, IL: University of Chicago Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners (Handbook I: Cognitive Domain).* New York: Longmans Publishing.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51.

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry, 34*(4), 310-316.

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, & Evaluation, 8*(9), 1-4.

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

Brown, A., & Itzig, J. (1976).*The introduction of humor and anxiety in academic test situations*. Southern Methodist University. ERIC Document Reproduction Service No. ED 152 783.

Burton, S. Sudweeks, Merrill, P. & Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Brigham Young University: Testing Services and the Department of Instructional Science.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323-331.

Cole, K. M., & Paek, I. (2017). PROC IRT: A SAS procedure for item response theory. *Applied psychological measurement*, *41*(4), 311-320.

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, *26*(2), 543-551.

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*(1), 241-247.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1990) *Essentials of psychological testing*. New York, NY: Harper Collins Publisher.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education*, *44*(1), 109-117.

DeVellis, R.F. (2006). Classical test theory. *Medical Care, 44*(11), 50-59.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research, 5*(2), 1-17.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 1-4.

Dolly, J. P., & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement, 46*(3), 619-625.

Downing, S. M. (1992). True-false and alternate-choice item formats: A review of research. *Educational Measurement: Issues and Practice*, *11*(3), 27-30.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77*(10), 103-104.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133-143.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education*, *8*(2), 187-207.

Doyle, K. O. (1974). Theory and practice of ability testing in ancient Greece. *Journal of the History of the Behavioral Sciences,10*(2) 202-212.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, *13*(4), 574-595.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing*, *23*(2), 198-228.

Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, *19*(2), 171-179.

Ebel, R. L. (1971). How to write true-false test items. *Educational and Psychological Measurement, 31*(2), 417-426.

Ebel, R. L. (1981, April). *Some advantages of alternate-choice test items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*(4) 267-278.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49.

Frary, R. B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation, 4*(11), 3.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357-364.

Frisbie, D. A. (1990). *The evolution of the multiple true false item format*. Paper presented at a symposium at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 5,* 21-26.

Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education*, *4*(1), 67-83.

Frisbie, D. A., & Druva, C. A. (1986). Estimating the reliability of multiple true-false tests. *Journal of Educational Measurement*, *23*(2), 99-105.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19*(1) 29-35.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082-1116.

Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement*, *45*(1), 1-13.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*(1), 21-25.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*(1), 73-88.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. New York, NY: Routledge.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51-78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999-1010.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333.

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

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are multiple-choice items too fat?. *Applied Measurement in Education, 32*(4), 350-364.

Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation & the Health Professions, 12*(1), 97-106.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53*(11), 1043-1055.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Measurement methods for the social sciences series, Vol. 2. Fundamentals of item response theory.* Thousand Oaks, CA: Sage Publications, Inc.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education, 62*(2), 143-157.

Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of" all of the above" on the reliability and validity of multiple-choice test items. *Evaluation & the Health Professions*, *21*(1), 120-133.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35-41.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125.

Hickson, S., Reed, W. R., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: Data from the classroom. *Educational Assessment, 17*(4), 200-213.

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple choice items. *Educational and Psychological Measurement*, *25*(1), 117-126.

Kelly, F. J. (1916). The Kansas Silent Reading Tests. *Journal of Educational Psychology*, *7*(2), 63.

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent ttem sets. *Applied Measurement in Education*, *16*(3), 207-222.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education*, *21*(3), 571-585.

Knowles, S. L., & Welch, C. A. (1992). A Meta-Analytic Review of Item Discrimination and Difficulty in Multiple-Choice Items Using" None-Of-The-Above". *Educational and Psychological Measurement*, *52*(3), 571-577.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212-218.

Kreiter, C. D., & Frisbie, D. A. (1989). Effectiveness of multiple true-false items. *Applied Measurement in Education*, *2*(3), 207-216.

Kurz, T. B. (1999, January). *A review of scoring algorithms for multiple-choice tests*. Paper presented at Annual Meeting of the Southwest Educational Research Association, San Antonio, TX.

Lane, S., Raymond, M. R., & Haladyna, T. M. (2015). *Handbook of test development.* New York, NY: Routledge.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement*, *43*(3), 675-685.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. *ETS R&D Connections*, *11,* 1-8.

Lord F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. New York, NY: Information Age Publishing.

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice*, *37*(3), 65-76.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*(3), 234-250.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Maher, M. H. K., Barzegar, M., & Ghasempour, M. (2016). The relationship between negative stem and taxonomy of multiple-choice questions in residency pre-board and board exams. *Research and Development in Medical Education, 5*(1), 32.

Maihoff, N. A., & Mehrens, W. A. (1985, April). *A comparison of alternate-choice and true-false item forms used in classroom examinations.* Paper presented at the Annual Researchers Meeting of the National Council on Measurement in Evaluation, Chicago, IL.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

McClellan, C. A. (2010). Constructed-response scoring—doing it right. *R&D Connections*, *13*, 1-7.

McMorris, R. F., Boothroyd, R. A., & Pietrangelo, D. J. (1997). Humor in educational testing: A review and discussion. *Applied Measurement in Education*, *10*(3), 269-297.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707-726.

Nguyen, T. H., Han, H. R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient-Patient-Centered Outcomes Research*, *7*(1), 23-35.

Osterlind, S. J. & Wang, Z. (2018). Item response theory in measurement, assessment, and evaluation for higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (2nd ed., pp.191-200). New York, NY: Routledge.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education, 7*(1), 1-7.

Pinglia, R. S. (1994). A psychometric study of true–false, alternate-choice, and multiple-choice item formats. *Indian Psychological Review, 42*(1), 21-26.

Rachor, R.E. and Gray, G.T. (1996, April) *'Must all stems be green? A study of two guidelines for writing multiple choice items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford: Nielsen & Lydiche.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133-144.

Rich, C. E., & Johanson, G. A. (1990, April). *An item-level analysis of "none of the above"*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Roberts, D. M. (1993). An empirical study on the nature of trick test questions. *Journal of Educational Measurement, 30*(4), 331-344.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13.

Samejima, F. (1979). A *New Family of Models for the Multiple Choice Item* (Research Report #79–4). Knoxville, TN: University of Tennessee, Department of Psychology.

Samejima, F. (1988). Comprehensive latent trait theory. *Behaviormetrika, 15*(24), 1-24.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39–83). Chicago, IL: Rand McNally.

Sideridis, G., Tsaousis, I., & Harbi, K. A. (2016). Improving measures via examining the behavior of distractors in multiple-choice tests: assessment and remediation. *Educational and Psychological Measurement, 77*(1), 82-103.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*(1), 73-98.

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Cary, NC: SAS Institute, Inc.

Strang, H. R. (1980). Effect of technically worded options on multiple-choice test performance. *The Journal of Educational Research*, *73*(5), 262-265.

Tamir, P. (1993). Positive and negative multiple-choice items: How difficult are they? *Studies in Educational Evaluation*, 19, 311-332.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*(8), 662-671.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198-206.

Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466-478.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13*(3) 201-214.

Thissen, D., Cai, L., and Bock, R. D. (2010). The nominal categories item response

model. In *Handbook of polytomous item response theory models: developments and applications*, (Eds.) M. Nering and R. Ostini (pp.43-75) New York, NY: Taylor and Francis.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*(3), 385-395.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161-176.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247-260.

Thorpe, G.L. and A. Favia, 2012. Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship*: 1-34.

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, *91*(9), 1426-1431.

Tractenberg, R. E., Gushta, M. M., Mulroney, S. E., & Weissinger, P. A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, *18*(5), 945-961.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, *1*(3), 355-369.

Trewin, S. A. (2007). Robert Yerkes' multiple-choice apparatus, 1913-1939. *The American Journal of Psychology*, *120*(4), 645-660.

Vyas, R., & Supe, A. (2008). Multiple choice questions: a literature review on the optimal number of options. *Natl Med J India*, *21*(3), 130-3.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*(2), 191-208.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103-118.

Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Journal of Nursing Education*, *45*(6), 212-219.

Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology*, *2*(2), 147.

Yerkes, R.M. (1921). Psychological examining in the United States army. *Mem. Nat. Acad. Sci.*, *15*, 1-890.