5-11-2023

A novel examination of none-of-the-above as it influences examinee item responses

Kathryn N. Thompson
James Madison University

A Novel Examination of None-of-the-Above as it Influences Examinee Item Responses

Kathryn N. Thompson

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2023

FACULTY COMMITTEE:

Committee Chair: Brian C. Leventhal, Ph.D.

Committee Members/ Readers:

Christine E. DeMars, Ph.D.

Dena A. Pastor, Ph.D.

Acknowledgements

I first would like to thank my advisor and mentor, Dr. Brian Leventhal. It is an honor to say that I have worked with you for the past five years. You have helped me and guided me through research, coursework, and graduate assistantship work, and you have been kind, patient, and understanding through everything. I truly cannot express how thankful I am for the opportunities you have given me.

I would next like to thank my incredible committee members, Dr. Christine DeMars and Dr. Dena Pastor. You both have provided me invaluable feedback and assistance through this dissertation, as well as in classes and work. I have learned so much from both of you, and I cannot thank you enough for taking the time to assist me through my graduate studies.

I would also like to thank my mom, dad, Ian, and Zach. You were there for me when I needed to talk through things or needed help. Without your support and love, I do not think I would have been able to accomplish this Ph.D. You have all been such a huge part of this experience, and I am excited to celebrate with you!

Finally, I would like to thank the faculty, staff, and students in the Assessment and Measurement program and the Center for Assessment and Research Studies. I have learned so much, and there have been so many great memories. You are all what makes this program special and worthwhile!

Table of Contents

List of Tables

List of Figures

Abstract

It is imperative to collect validity evidence prior to interpreting and using test scores. During the process of collecting validity evidence, test developers should consider whether test scores are contaminated by sources of extraneous information. This is referred to as construct irrelevant variance, or the "degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (AERA et al., 2014, p. 12). One possible source of construct irrelevant variance is violating item-writing guidelines, such as to "avoid the use of none-of-the-above" in multiple-choice items (Rodriguez, 2016, p. 268).

Numerous studies have been conducted with regards to how none-of-the-above (NOTA) impacts item statistics, such as item difficulty, item discrimination, and test score reliability. The impacts of NOTA on item statistics are mixed and often depend on whether NOTA is the correct or incorrect option. In the case of NOTA as the incorrect option, NOTA tends to be more frequently selected by examinees (Garcia-Perez, 1993; Frary, 1991). This increased selection is hypothesized to be due to a potential selection tendency that examinees possess toward NOTA (Butler, 2018). While this tendency toward selecting NOTA is hypothesized in the literature, there has not yet been a study which tests this hypothesis.

In the current study, I extended previous NOTA literature to explore whether item difficulty varies across groups of examinees who receive a test with NOTA and a test without NOTA, after controlling for examinee ability. I also tested whether there is a hypothesized selection tendency toward NOTA. Overall, as described in previous

research, NOTA resulted in mixed results. I discuss these results, as well as future areas of NOTA research.

Introduction

Tests have a variety of purposes and uses, from predicting whether an examinee should be admitted into college to diagnosing where examinees struggle to learn material. Test developers and test users continually strategize ways to collect evidence in support of test score interpretations, regardless of the type of test. This collection of evidence relates to validity, which is the "degree to which evidence and theory support the interpretations of test scores for proposed uses of a test" (AERA et al., 2014, p. 11). The collection of validity evidence is an ongoing process and provides the basis for an argument in support of test score interpretations (Kane, 2016).

There are five types of validity evidence that test developers and test users may collect to support test score interpretations. These include evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing (AERA et al., 2014). Each type is not required to ensure adequate support of test score interpretations. Rather, only relevant sources of validity evidence should be collected to support each proposed test score interpretation and use.

As an example, consider a university that requires an examinee to earn a passing score on an information literacy competency test prior to graduation. University faculty want to interpret the test score as an indicator of whether examinees are competent in information literacy. Prior to using the test scores as an indicator of competence, the university faculty request assistance from measurement experts. The experts decide to focus on three types of evidence to establish a basis for the information literacy competency argument: the collection of response process data, internal structure, and relations with other variables.

The experts first collect response process validity evidence in the form of cognitive interviews with expert faculty and examinees (Gorin, 2006). To support the claim of information literacy competency, the experts request that information literacy faculty describe their hypotheses about how examinees will correctly answer the multiple-choice items. They compare the proposed response processes to the actual response processes of examinees with individual think-aloud sessions. During the think-aloud sessions, the researcher asks the examinee to describe how they answered the items to determine whether the examinees used information literacy knowledge or a different construct. For example, the researchers notice a theme in examinees' described response processes. Specifically, some examinees explain that they used information from a previous item to answer the current item. Rather than use information literacy knowledge to answer the item, the examinees used context clues. This is evidence of a misalignment between hypothesized and observed response processes. The measurement experts work with information literacy faculty to revise the items so that context clues from one item do not influence an answer on the other items.

The experts also collect internal structure validity evidence to explore the different components that comprise information literacy. The experts, with the assistance of information literacy faculty and literature, hypothesize that information literacy is made up of multiple subfacets. Each item aligns with a certain objective that represents a different component of information literacy. The experts collect examinee responses to examine the relationships between items and conduct a confirmatory factor analysis to formally test dimensionality. Overall, there are six distinct groups of items that comprise

information literacy. These analyses provide evidence in support of the internal structure of information literacy items.

Finally, the experts consider whether information literacy competency scores are related to relevant variables. The experts consult with faculty and determine that relevant course grades (e.g., oral communication) are hypothesized to be positively related with information literacy competency scores. Additionally, the faculty suggest that those who are considered competent in information literacy should also have a greater confidence in information literacy skills. The experts assist the faculty in performing a logistic regression analysis to predict whether examinees with high relevant course grades and high confidence in information literacy statistically significantly predict information literacy competency (i.e., competent or not competent).

While the university faculty are satisfied with the measurement experts' collection of validity evidence to support information literacy competency thus far, they continue to collect evidence to strengthen their claims. Evidently, the collection of validity evidence is an ongoing process, and the goal is to build an argument for test score interpretations (Kane, 2016). Unfortunately, test scores are often confounded with sources of error (e.g., examinees' use of context clues from one information literacy item to answer a different information literacy item). In other words, the construct of interest is contaminated by extraneous information. This extraneous information is troublesome since it influences the interpretations of test scores. This is referred to as construct irrelevant variance, or the "degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (AERA et al., 2014, p. 12). To lessen the consequences of construct

irrelevant variance on test score interpretations, it is imperative to “disentangle the construct-relevant from the construct-irrelevant” (Kyllonen, 2016, p. 200).

There are a wide range of extraneous factors that can influence test score interpretations. One source is item-writing errors, which can be mitigated by using item-writing guidelines. Not following item-writing guidelines may have a significant influence on the validity of test score interpretations (Haladyna et al., 2002; Haladyna & Rodriguez, 2013; Rodriguez, 2016). Continuing with the previous example, the information literacy item in Figure 1 violates a multiple-choice item-writing guideline. The question ends with the article ‘an’, which narrows the correct option to A (expert). The grammatical inconsistency found in Figure 1 can cue examinees into selecting the correct answer (Haladyna et al., 2002). This simple item-writing error is an issue since the information literacy test scores are “systematically influenced to some extent by [grammatical ability]” (AERA et al., 2014, p. 12).

In some instances, it is simple to identify the source of construct irrelevant variance and how it will influence test scores. This includes other multiple-choice item-writing guidelines, such as writing the options of equal length and avoiding nonsensical options (Haladyna et al., 2002). Such options can result in measurement of test-wiseness, which confounds the primary construct of interest. Therefore, the score interpretation is contaminated with an examinee’s ability to recognize certain test characteristics and item formats. In turn, this results in the examinee’s test score being inflated, or increased, in such a way that is not due to the primary construct of interest (Millman et al., 1965; Sarnacki, 1979).

However, in many testing situations, it is challenging to predict the effect that construct irrelevant variance will have on test scores. The violation of the item-writing guidelines described above is associated with the construct of test-wiseness, which commonly results in an increase in test scores (Sarnacki, 1979). The effect of violating certain item-writing guidelines, such as to "avoid using the option none of the above [NOTA]" (Haladyna & Rodriguez, 2013) is not as consistent nor predictable. Mullins (1963), Hughes and Trimble (1965), and Caldwell and Pate (2013) have alluded to the irrelevant information that may contaminate test scores when NOTA is present as an option in a multiple-choice item. Specifically, NOTA is theorized to invoke different examinee cognitive response processes when present as an option. While some argue that NOTA results in the recall of relevant information to correctly answer the item (e.g., Butler, 2018; Gross, 1994), others argue that NOTA necessitates the recognition of information to correctly answer the item (e.g., Caldwell & Pate, 2013; Rodriguez, 1997). Furthermore, Butler (2018) proposes that the presence of NOTA is different from other response options and cues examinees to select it. In other words, some examinees may select NOTA simply because it is present in a multiple-choice item (i.e., selection tendency). While these theories were proposed based on results of empirical research, they have not been formally tested.

Violating an item-writing guideline is likely to result in irrelevant variance which negatively influences the interpretations of test scores (Haladyna et al., 2002). While violation of certain item-writing guidelines results in a predictable influence on test scores (e.g., test-wiseness results in an increase in test scores), the use of NOTA as an option is not as simple. NOTA does not have consensus nor sufficient empirical evidence

to suggest its use as a multiple-choice option, especially as it relates to its influence on test score interpretations. Therefore, there is a need to examine NOTA with construct irrelevant variance in mind to determine the effect it has on the validity of test score interpretations.

In the current study, my aim is to increase empirical evidence for or against the use of NOTA in multiple-choice items. I focus on the collection of validity evidence as it impacts examinee response process by utilizing two psychometric models. I begin by examining the cognitive response demands that NOTA places on examinees as they answer multiple-choice items by examining differences in item difficulty. I then model the tendency to select NOTA as a second construct that is separate from the primary construct of interest. My goal is to use these two methods to explore the effects of NOTA in a novel way. Rather than use only item statistics to make conclusions about the use of NOTA, I use both item statistics and psychometric models to explore whether the presence of NOTA results in construct irrelevant variance.

Literature Review

Various nuisance factors can impact examinee item responses, which, in turn, impact test score interpretations. Following item-writing guidelines is one method to combat the negative impacts of construct irrelevant variance. In the case of writing multiple-choice items, guidelines are abundant. One schema (Haladyna & Rodriguez, 2013; Rodriguez, 2016) documented 22 multiple-choice item-writing guidelines. They are classified into five categories: (1) content concerns, (2) format concerns, (3) style concerns, (4) writing the stem, and (5) writing the options (Haladyna & Rodriguez, 2013; Rodriguez, 2016).

The first three multiple-choice item-writing guideline categories concern general aspects of the multiple-choice item (i.e., content, format, and style) and comprise ten of the 22 guidelines. The last two categories concern specific components of the multiple-choice item (i.e., stem and options) and comprise 12 of the 22 guidelines. The stem is a statement, or question, that prompts the examinee to select an answer. The options are further divided into the correct answer and incorrect answers, or distractors (Gierl et al., 2017). In Figure 1, the stem is the statement (i.e., “Scholarly sources are written for an”) which prompts the examinee to consider the options (i.e., A, B, C, and D). The correct answer is option A (“expert”), and the distractors are options B (“general audience”), C (“college student”), and D (“middle school student”).

The greatest number of multiple-choice item-writing guidelines fall under concerns for writing of the options (i.e., 10 of the total 22). Rodriguez (2016) states that “some evidence exists for guidelines” (p. 268), such as to “use only options that are plausible and discriminating”, “avoid the use of none-of-the-above, all-of-the-above, and

I don't know", and "avoid giving clues to the right answer" (p. 268). While these guidelines are considered valid based on item psychometric evidence (i.e., item difficulty, item discrimination, and test score reliability), there have been fewer empirical examinations of multiple-choice item-writing guidelines in recent years (Haladyna & Rodriguez, 2013; Rodriguez, 2016), especially as they influence the interpretation of test scores. Rather than focus on item difficulty, item discrimination, and test score reliability to support the use of an item-writing guideline, researchers should also focus on testing sources of construct irrelevant variance that may arise due to breaking an item-writing guideline.

One item-writing guideline that has received more empirical attention than others, yet has the least amount of consensus, is the use of NOTA as an option (Haladyna et al., 2002; Haladyna & Downing, 1989a; 1989b). This lack of consensus occurs in theoretical, empirical, and synthesis research. Specifically, there are inconsistencies of NOTA's influence on psychometric item qualities, which results in varying conclusions on whether to include NOTA in multiple-choice items. Hypothetically, while item difficulty may increase[1] in two different NOTA studies, researchers offer different conclusions about whether to use NOTA. In other words, it is unclear whether, for example, a decrease in item difficulty due to NOTA is evidence to use it in multiple-choice items. This inconsistency between evidence and interpretation appears in both empirical and synthesis research.

[1] Item difficulty in this context is defined as the proportion of examinees of select the correct option. An increase in item difficulty tends to indicate a harder item while a decrease tends to indicate an easier item. However, in the context of NOTA research, researchers often describe increases in item difficulty as a harder item and decreases as an easier item.

While Haladyna and Downing (1989a) originally concluded to "avoid [NOTA]" (p. 44) due to an overall increase in item difficulty and decrease in item discrimination and test score reliability (Haladyna & Downing, 1989b), Haladyna et al. (2002) later reported that "[NOTA] should be used carefully" (p.312). Knowles and Welch (1992) reported similar findings in their meta-analysis with regards to NOTA's impact on item discrimination. Yet, the authors advocate for the use of NOTA. This contrasts with Haladyna et al.'s (2002) suggestion to carefully use NOTA even though, on average, in Knowles and Welch's study, item difficulty did not change. Finally, Rodriguez (1997) performed a meta-analysis with a variety of item writing guidelines that have the least consensus, which included the use of NOTA. Most frequently, NOTA resulted in increased item difficulty and decreased test score reliability. Similar to Knowles and Welch (1992), on average, item discrimination did not change as a result of NOTA. Rather, item discrimination decreased in half of the studies and increased in the other half. Therefore, Rodriguez (1997) concluded that test developers should carefully use NOTA, which aligns with Haladyna et al.'s (2002) research.

One important caveat that Rodriguez (1997) identifies is whether the effects of NOTA are different depending on whether NOTA is the correct or incorrect option. For studies that included NOTA as the correct option, item difficulty increased (Bonyton, 1950; Dudycha & Carpenter, 1973; Frary, 1991; Wesman & Bennet, 1946). Rodriguez (1997) suggests that when NOTA is the correct answer, examinees must retrieve knowledge to consider whether each of the response options is correct or incorrect. Upon rejecting all distractors, the examinee selects NOTA. Additionally, the presence of NOTA may cause examinees to consider each response option more carefully than when

it is not present. This indicates that selection of NOTA is based on the careful consideration of each option due to the construct of interest. While a similar response process is proposed about NOTA as a distractor, findings with NOTA as a distractor are slightly different than findings with NOTA as the correct answer.

In studies where NOTA is the incorrect option, frequency of NOTA selection tended to increase compared to the fourth option of items where NOTA was not present (Frary, 1991; Garcia-Perez, 1993). Frary (1991) collected item-level data from a variety of disciplines in a higher education setting. He compared items statistics between items that contained NOTA (i.e., NOTA items) and items that did not contain NOTA (i.e., non-NOTA items) within the same test. On average, across all tests, when NOTA was the incorrect option, the NOTA items were not only more difficult but also had higher NOTA selection compared to non-NOTA items and fourth option selection. The differences in proportion of selection for the NOTA options and non-NOTA options suggested that examinees were drawn to select NOTA rather than the correct option. Garcia-Perez (1993) found similar results with NOTA as the incorrect option. While Garcia-Perez's (1993) goal was to improve examinee ability estimation, he describes similar results to Frary (1991) and explains that NOTA selection depends on how examinees interact with NOTA items. Specifically, the inclusion of NOTA is expected to result in mixed effects on item statistics, which are likely due to how examinees process NOTA as an option.

In contrast to the response process associated with NOTA as the correct option, when NOTA is the incorrect option, examinees simply need to recognize the correct answer (Caldwell & Pate, 2013; Frary, 1991; Garcia-Perez, 1993). When an examinee only needs to recognize rather than recall the correct answer, "knowledge of the correct

answer is not an absolute requirement" (Caldwell & Pate, 2013, p.3). Additionally, Caldwell and Pate (2013) theorize that an examinee may wrongly select NOTA as a distractor simply because it is present. The examinee is drawn to NOTA even when, in truth, the examinee possesses a high level of the underlying construct.

Most previous empirical work focuses on item difficulty and other psychometric item characteristics (i.e., item discrimination and test score reliability). However, some researchers make note of the influence of NOTA on a different outcome. Garcia-Perez (1993) focused on the efficiency of examinee ability estimates. Estimating examinee ability as finite-state scores, he concluded that the inclusion of NOTA results in more precise estimates of ability (i.e., smaller confidence intervals about ability estimates). In contrast to other researchers, Garcia-Perez (1993) recommends the use of NOTA due to the increased precision of ability estimates, especially when latent ability is of interest. While there has not been a replication of Garcia-Perez's (1993) study, which focused on NOTA as it influences the precision of ability estimates, Dochy et al. (2001) examined the influence of NOTA using log-linear models. The log-linear models enabled Dochy et al. (2001) to consider how the selection of NOTA changes as ability increases. Specifically, Dochy et al. (2001) noted that examinees with a lower ability tended to select NOTA more often than higher ability examinees. Similar to Garcia-Perez (1993), this resulted in a greater focus on NOTA as it has to do with examinee scores rather than psychometric item qualities.

In more recent years, relative to the most current NOTA meta-analysis (i.e., Haladyna et al., 2002), focus has been placed on NOTA as it influences learning (e.g., Blendermann et al., 2020; Butler, 2018; DiBattista et al., 2014; DeVore et al., 2016; Jang

et al., 2014; Odegard & Koen, 2007). This literature includes a focus on the utility of multiple-choice items as they help examinees learn and retain information while testing. Additionally, similar to Garcia-Perez (1993) and Dochy et al. (2001), there is some reference to NOTA as it relates to examinee scores. Unfortunately, in these studies, little is mentioned outside of what has already been found by previous NOTA researchers.

Butler (2018) describes one way to consider the influence of NOTA on examinees as they answer multiple-choice items. Specifically, Butler (2018) refers to the influence of NOTA on examinees as a selection bias and theorizes that frequency of selection increases as the presence of NOTA on a test increases. A similar finding has been made with regard to all-of-the-above (AOTA). Although it is not the same as NOTA, Harasym et al. (1998) refer to this as a cuing effect. Simply the presence of AOTA causes, or cues, examinees to select it. This is problematic as selection is not due to the construct of interest but rather due to construct-irrelevant selection bias.

Gross (1994) furthers the theoretical argument against NOTA. In addition to a potential cuing effect, examinees may simply recognize NOTA as the correct answer. Examinees do not necessarily know the correct answer but are able to identify that the other response options are incorrect and thus select NOTA as the correct answer (Gross, 1994). This conflicts with the explanation Rodriguez (1997) describes as examinees are not using retrieval of knowledge to select the correct answer but simply recognizing that the other response options are incorrect (Blendermann et al., 2020; Butler, 2018; DiBattista et al., 2014; Jang et al., 2014; Odegard & Koen, 2007). This may be an issue as examinees do not necessarily know the correct option but only what is incorrect.

In addition to theories of examinee cognitive response processes when NOTA is present, others have also proposed that the use of NOTA in multiple-choice items results in construct irrelevant variance due to cuing or selection bias, as suggested by Butler (2018). Mullins (1963) compared the use of NOTA in multiple conditions. First, two tests of differing content (i.e., vocabulary and spatial reasoning) were administered where NOTA was varied as the correct option or as a distractor. The tests either included or excluded NOTA. The correlations amongst test scores including NOTA resulted in stronger, positive correlations in comparison to a test excluding NOTA. Mullins (1963) described that "it appears that something is being measured by [NOTA] not contained in the ability score" (p.157).

Hughes and Trimble (1965) referred to NOTA as a complex alternative or response option. They hypothesized that the use of complex response options could invoke a separate construct of test-wiseness. As stated previously, Harasym et al. (1998) extended this point but applied it to AOTA. If extended to NOTA, perhaps presenting NOTA as a response option will result in a potential bias in examinee responses since some examinees are more test-wise than others.

One way to explore whether the presence of NOTA is associated with an irrelevant construct is to explore the presence of differential item functioning (DIF). An item is flagged as exhibiting DIF if the probability of selecting the correct option differs across groups, after controlling for ability. In the context of NOTA, an item would exhibit DIF if examinees with the same ability respond to an item differently based on whether they receive a test with NOTA or without NOTA. Haladyna and Downing (2004) describe DIF as a source of construct irrelevant variance since examinees respond

differently to the same item even if they share the same ability level on a construct. Therefore, differences in examinee responses are not due to the construct of interest. Rather, a separate construct due to the presence of NOTA could be contributing to the way in which examinees respond. As others have alluded to (e.g., Bulter, 2018; Hughes & Trimble, 1965; Mullins, 1963), the presence of NOTA is hypothesized to invoke a separate construct that influences examinee responses. DIF detection methods could be useful in identifying whether examinees respond differently to an item simply due to the presence of NOTA. If an item that includes NOTA is flagged for DIF, then it could be inferred that the presence of NOTA results in a construct separate from the primary construct of interest.

Another way to explore whether the presence of NOTA is associated with an irrelevant construct is to explore a potential NOTA selection tendency. Thus far, examinees are hypothesized to only use one, simultaneous step (i.e., recognition or recall) to answer an item when NOTA is present. This response process can be statistically evaluated using a divide-by-total structure model (Thissen et al., 1989; Thissen & Steinberg, 1984). Such models include the nominal response model (NRM; Bock, 1972), where it is hypothesized that one or more underlying constructs influence an examinees' simultaneous comparison across all of the response options (Deng & Bolt, 2016; Thissen et al., 1989; Thissen & Steinberg, 1984). An examinee assigns a probability to each response option based on their confidence in the response option being correct. The examinee then selects the response option that is assigned the highest probability of being correct as the correct answer. It is important to note that examinees are not necessarily aware of this process. Rather, examinees implicitly perform this process as they answer

multiple-choice items. However, the response process associated with NOTA could be more complicated.

In contrast to the divide-by-total model, the response process hypothesis underlying a sequential model follows a stepwise structure rather than a simultaneous structure. Assuming a sequential response process, an examinee approaches responding to an item in a series of steps. During these steps, the examinee does not view the options holistically. Rather, an examinee considers certain options one at a time and eliminates options sequentially (DeBoeck & Partchev, 2012; Deng & Bolt, 2016). Such a response process can be statistically evaluated through the use of item response trees (IRTrees). Sequential models differ from divide-by-total models in that examinees' "unique proficiency dimensions" (Deng & Bolt, 2016, p. 244) are separated into steps, whereas divide-by-total models condense processing of response options into one step.

Previous applications of a sequential response model include the modeling of intuitive and deliberate reasoning (Böckenholt, 2012a) and recognition and correction of sentence errors (Deng & Bolt, 2016). Böckenholt (2012a) introduced the Cognitive-Miser Response (CMR) Model as a way to model two separate examinee response tendencies: intuitive and deliberative reasoning. Examinees are hypothesized to either select an option based on an immediate response, which decreases the cognitive load placed on an examinee, or select an option based on a careful response, which increases the cognitive load placed on an examinee. While this immediate response, or intuitive reasoning, allows the examinee to save cognitive resources through the use of heuristics, these "quick and plausible judgements" (Böckenholt, 2012a, p. 388) are not always correct. For example, Böckenholt (2012a) describes an item from Frederick's (2005) Cognitive

Reflections Task (CRT). An examinee is given that "a bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?". An examinee who relies on intuitive reasoning will select the option of "10 cents". While "10 cents" is reasonable for a majority of examinees in terms of intuition, it is not the correct answer. An examinee who employs deliberative reasoning "may realize that the difference between $1 and 10 cents is less than $1" (Böckenholt, 2012a, p. 388). To disentangle these response tendencies, Böckenholt (2012a) utilized an IRTree. Intuitive or deliberate reasoning are hypothesized to be two separate latent traits. An examinee who selects a specific option (e.g., 10 cents in the CRT item) is hypothesized to exhibit inhibitory control. If an examinee does not select this option, then the examinee is hypothesized to exhibit deliberate reasoning through selection of a different option.

Deng and Bolt (2016) extended Böckenholt's (2012a) CMR model to a sentence correction task. Examinees were provided a sentence with an underlined portion, where they were asked to select an option with the same underlined portion as in the sentence (i.e., no sentence correction needed) or an option with the underlined portion changed. The examinee needed to identify the option with the correct underlined portion. Similar to Böckenholt (2012a), Deng and Bolt (2016) specified a two-step process where an examinee first recognized that the underlined portion was correct or incorrect. If the examinee identified that the underlined portion was incorrect, then the examinee moved to the correction step. Deng and Bolt (2016) referred to this model as a sequential response model for multiple-choice items (SRM-MC) as examinees completed a stepwise process to select the correct answer. They described it as a sequential, stepwise process since the examinee's first step was to recognize whether the first option was correct,

which was always the same underlined portion found in the sentence stem. The examinee moved to the second step if they deemed the underlined portion in the sentence stem to be incorrect (i.e., the examinee identified that a correction needed to be made).

While Böckenholt's (2012a) CMR model and Deng and Bolt's (2016) SRM-MC are applied in different contexts, the models share a communality. Specifically, there is a need to consider whether a unique latent trait influences examinees' steps when answering multiple-choice items. As alluded to by some researchers who examined the influence of NOTA on psychometric item qualities (e.g., Butler, 2018; Caldwell & Pate, 2013; Harasym et al., 1998; Hughes & Trimble, 1965; Mullins, 1963), a separate latent trait from the one of interest may influence examinees' responses to multiple-choice items.

There has yet to be a study to determine whether NOTA invokes a selection tendency separate from the latent trait of interest. Much of the current NOTA literature has focused on how the presence of NOTA in multiple-choice items influences learning (e.g., Blendermann et al., 2020; Butler, 2018; DiBattista et al., 2014; Jang et al., 2014), but the conclusions were made based on item difficulty (i.e., proportion of correct selection) which is a consistent theme to previous NOTA research. Unfortunately, the methods in which NOTA has been tested do not disentangle whether it is NOTA that results in differing selection frequency. Further, there have only been speculations about how NOTA influences examinees' responses to multiple-choice items, but these speculations have not been empirically tested. With the creation of new models that assume certain steps in examinees' response processes to multiple-choice items, such as

Deng and Bolt's (2016) SRM-MC, it is time to revisit NOTA with emphasis placed on examinees rather than solely on psychometric item qualities.

The collection of validity evidence is lacking for NOTA as it has to do with construct irrelevant variance. It is imperative to ensure that the presence of NOTA is not associated with contaminating examinee test scores. If there is a separate, irrelevant latent trait that influences examinees responses to multiple-choice items, which is due to NOTA, then it is not possible to make valid interpretations of test scores (AERA, APA, & NCME, 2014). To strengthen the argument for or against the use of NOTA, I propose two research questions:

1. Is there a difference in item difficulty for examinees who receive a test with NOTA compared to examinees who receive a test without NOTA, after controlling for ability?
2. Do examinees exhibit a tendency to select NOTA when present as a distractor in multiple-choice items?

Methodology

In the methodology section, I describe the participants, measures, and analysis plan with the goal of addressing two research questions:

1. Is there a difference in item difficulty for examinees who receive a test with NOTA compared to examinees who receive a test without NOTA, after controlling for ability?
2. Do examinees exhibit a tendency to select NOTA when present as a distractor in multiple-choice items?

Participants

Convenience sampling was used to recruit participants from the James Madison University Department of Psychology Participant Pool during the Fall 2022 semester. All participants were 18 years of age or older. Participants were able to select which study they would like to participate in by browsing a list of study names on the Participant Pool website. Participants who signed-up for the current study were shown the study title prior to completing the study, which was "Student Perceptions of Cognitive Demands Invoked by Test-Taking and the Relationship with Test Scores".

After signing-up, participants were provided a link to begin the study to receive credit for their classes. Participants had the option to complete the study at any point prior to November 30, 2022. Once examinees decided to participate, they were directed to Qualtrics and provided a consent form. Participants were asked to read a consent form prior to participating in the study. If the participant did not consent to being in the study, the participant was routed to the end of the survey. If the participant did consent to being in the study, the participant was randomly assigned to complete a test with NOTA or

complete a test without NOTA. Qualtrics has a feature that allows for random assignment of examinees to one of the two testing conditions.

Measures

A total of three measures were administered to participants. The first measure was an information literacy test. The second and third measures were the Effort subscale of the Student Opinions Survey and the Perceptions of None-of-the-Above Survey. First, participants were asked to report their perceived effort while completing the information literacy test. Second, participants were asked to report their perceptions of NOTA. The purpose for each measure, as well as reliability and validity evidence associated with the scores is provided in the following sections.

Information Literacy Test

The information literacy test consists of 30 multiple-choice items with four options each. None of the options on the original test include NOTA. The test is typically administered to investigate value added gains through a first-year general education curriculum, a population similar to the one in the current study. In addition to administering the original information literacy test to a group of participants, I administered a new version of the information literacy test to a separate group of participants. The new version consists of the same 30 multiple-choice items with 10 items modified to contain NOTA as a distractor. The goal in creating this new version is to isolate the effects of NOTA as a distractor on examinee performance. Therefore, examinees were randomly assigned to either complete the original information literacy test or the test which included NOTA as a distractor. The number of items which included NOTA was selected by reviewing previous NOTA literature. Overall, there is

large variability in the number of items that had distractors replaced with NOTA. Specifically, as few as 20% (e.g., Pachai et al., 2015) to as many as 40% (e.g., Frary, 1991) of items replaced an option with NOTA. With such variability, I decided to select 33% of items (i.e., 10 of 30 items) to replace an option with NOTA.

The decision to replace distractors from the original information literacy test with NOTA was made based on NOTA literature and practical utility of NOTA. In terms of NOTA literature, researchers have performed similar replacements with the goal of comparing the performance of NOTA to the performance of non-NOTA items (e.g., Pachai et al., 2015; Dochy et al., 2001; Frary, 1991). Additionally, in these studies, researchers hypothesize that NOTA as a distractor is associated with an increase in the frequency of selection in comparison to items where NOTA was not used as a distractor. In terms of practicality, NOTA is a common filler distractor when it is challenging to develop appropriate, well-functioning distractors (Pachai et al., 2015). For example, item developers may be able to think of two plausible options but developing a third plausible option is challenging. As a result, item developers use NOTA so that an item has four options. In other words, NOTA is used when there are already good distractors or in place of a potentially weak distractor. For these reasons, distractors on the information literacy test were replaced with NOTA rather than being added as a fifth option.

Item and distractor analyses from previous information literacy test administrations informed which distractors to replace with NOTA. I examined the item and option difficulty and discrimination values. Item difficulty was calculated as the proportion of examinees who selected the correct option, and option difficulty was calculated as the proportion of examinees who selected each distractor (Haladyna, 2016).

Values range from 0 to 1, where lower values indicate fewer examinees answered the item correctly and higher values indicate more examinees answer the item correctly (Gierl et al., 2017; Haladyna & Downing, 1993). Item and option discrimination were calculated as the corrected point-biserial correlations (i.e., excluding the item score from total score when calculating the point-biserial correlation for a specific item), which is the correlation between the item score (correct or incorrect) or option (selected or not selected) and total test score (Haladyna, 2016). Values range from -1 to 1, where negative values indicate a negative relationship between total score and item score or option selection, and positive values indicate a positive relationship between total score and item score or option selection. The correlation between a distractor and total score should have a negative relationship (i.e., as total score increases, distractor selection decreases) while the correlation between the correct option and total score should have a positive relationship (i.e., as total score increases, correct option selection increases) (Gierl et al., 2017; Haladyna & Downing, 1993).

I determined which distractors to replace with NOTA by identifying poor functioning distractors. Poor functioning distractors are those that do not perform in appropriate ways, such as higher examinee proportion of selection compared to the correct option, a positive point-biserial correlation, or when fewer than 5% of examinees select the distractor. The latter distractor is known as an implausible distractor since very few examinees select it (Haladyna & Downing, 1993). The implausible distractor definition was used as criteria for selecting which distractor to replace with NOTA. In the case where there was more than one implausible distractor, the distractor with the point-biserial correlation nearest 0 was selected for replacement. When distractors perform in

this way, they should be removed and/or edited. I selected these replacement criteria to reflect situations where item developers have trouble developing a final distractor. Thus, the other distractors have much better qualities and higher selection frequency. Table 1 provides the item and distractor analysis results.

In total, 10 distractors were replaced with NOTA. NOTA was always the final (i.e., fourth) option. Within some items, the implausible distractor was not the fourth option. For example, within one item, the implausible distractor was option A. To include NOTA as the fourth option, option A was removed. The remaining three options were then shifted from options B, C, and D to options A, B, and C, respectively. NOTA was then included as the fourth option. I refer to the items that contained NOTA (i.e., manipulated items) as the NOTA items. For the test that did not contain NOTA, the implausible distractor was moved to the fourth option to make comparisons with the NOTA option. I refer to the items that did not include NOTA (i.e., unmanipulated items) as the non-NOTA items. Item stem wording was slightly edited so that the inclusion of NOTA would make sense as an option. The impact of the presence of NOTA will be discussed when describing the Perceptions of None-of-the-Above attitudinal measure.

Student Opinions Survey: Effort

The Student Opinions Survey (SOS) consists of 10 attitudinal items. Participants respond on a 5-point Likert scale from 1 (“Strongly agree”) to 5 (“Strongly disagree”). The SOS consists of two subscales: Effort and Importance (Sundre & Thelk, 2007). In the current study, participants only completed the five item Effort subscale to explore whether they gave effortful responses while completing the information literacy test. A low Effort score is an indication that the participant did not put forth effort while

completing the information literacy test. If participants reported low Effort scores (i.e., a total score less than 15), then the item response was filtered from analyses. This cut score is suggested by Sundre and Thelk (2007).

Perceptions of None-of-the-Above Survey

I developed the Perceptions of None-of-the-Above Survey to examine participants' perceptions of NOTA. The focus of NOTA research has been on item statistics, but the way in which participants perceive NOTA has not been examined. The purpose of including this attitudinal measure is to give context to examinee item responses. Figure 2 includes the five survey items. It is important to note that examinees were not allowed to go back to the information literacy test. Participants first report the number of items of which they believe contained NOTA on the previously completed information literacy test on a scale of 1 to 30 (i.e., the total number of items on the information literacy test). Participants then describe their thought process when NOTA is present. Finally, participants report their attitudes toward NOTA on a 5-point Likert scale from 1 ("Strongly agree") to 5 ("Strongly disagree").

Analyses

Research Question 1

The goal of the first research question is to examine whether there are differences in item difficulty depending on whether an examinee completed an information literacy test with NOTA, controlling for ability. To answer this research question, I analyzed item responses using a multilevel Rasch model to detect whether differential item functioning (DIF) is present on items which contain NOTA. Recall that because examinees were randomly assigned to receive a test with NOTA (NOTA group) or a test without NOTA

(non-NOTA group), I assumed that examinees in each group have, on average, similar information literacy ability. Due to sampling error in the measurement of information literacy scores, the average scores will not be the same. Any differences in item difficulty on the experimental items (i.e., items that are manipulated to contain NOTA for the NOTA group) are hypothesized to be due to the inclusion of NOTA.

To test whether DIF is present, I fit a series of multilevel Rasch models. I used a multilevel framework as the data is hierarchical in structure. Specifically, items are nested within examinees. In addition to the hierarchical structure of the data, one advantage to using a multilevel modeling perspective to model item responses is the simultaneous estimation of examinee ability and the relationship between examinee ability and predictors (Pastor, 2003). Specifically, these relationships are more accurate through simultaneous estimation compared to separate estimation of abilities and the relationship with predictors.

To estimate the models, I used the nlmixed procedure in SAS 9.4M7 (SAS Institute Inc., 2020). The nlmixed procedure is used to fit nonlinear mixed models and, by default, uses maximum likelihood estimation by adaptive Gauss-Hermite quadrature. The outcome was whether the examinee selected the correct option (1) or one of the incorrect options (0) for all models, but the predictors varied between models. Across all models, the first level is the item-level and the second level is the examinee-level. I describe each of the models in the following section. In the Appendix, I provide an example of the data structure required to analyze the item responses using the nlmixed procedure.

Model 1a. Unconditional Model. In the first model, or unconditional model, I estimated a multilevel Rasch model with no examinee-level predictors. This model is

equivalent to a Rasch model, where items are considered fixed. The log-odds of the probability of examinee j correctly answering item i is

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{30j}X_{30ij} \quad (1)$$

The examinee intercept, β_{0j}, is the overall effect across all items for examinee j. As there are a total of 30 information literacy items, there are a total of 30 item effects (β_{ij}). The X_{ij} values represent an effect code for a given item, where -1 indicates the item.

At the second level, examinees are considered random. An ability, u_{oj}, is estimated for each examinee. Examinee abilities are normally distributed with a mean of 0 and variance of $var(u_{0j})$. To model the variation in examinee ability, the examinee intercept β_{0j} at level-1 in Equation 1 is equal to u_{0j} (see Equation 2). Because the item effects, β_{ij}, are fixed, the item effects at level-2 are simply equal to γ_{i0}.

$$\begin{cases} \beta_{0j} = u_{0j} \\ \beta_{1j} = \gamma_{1,0} \\ \beta_{2j} = \gamma_{2,0} \\ \quad \vdots \\ \beta_{30j} = \gamma_{30,0} \end{cases} \quad (2)$$

The equations at both levels can also be modeled as a single equation, where the log-odds of the probability examinee j correctly answering item i is

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = u_{0j} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2i} \ + \cdots + \gamma_{30,0}X_{30ij} \quad (3)$$

It is important to note that the item effects are synonymous with item difficulties in the Rasch model due to the way in which I coded the X_{ij} values for each item (see Appendix). To obtain an item difficulty estimate (i.e., γ_{i0}) equivalent to that under the Rasch model, the items must be coded as -1 or 0. Replacing the item codes for a specific item with the X_{ij} terms in Equation 3 result in the Rasch model item difficulties. Using

item 1 as an example, X_{1i} is replaced with -1 while the other X_{ij} terms for item 2 to item 30 are replaced with 0.

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = u_{0j} + \gamma_{10}(-1) + \gamma_{20}(0) + \cdots + \gamma_{30,0}(0) \quad (4)$$

This simplifies Equation 3 to be the log-odds of the probability of examinee j correctly answering item 1

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = u_{0j} - \gamma_{10}, \quad (5)$$

which is equivalent to the Rasch model item difficulty estimates. Therefore, I interpret the γ_{i0} terms as item difficulty parameter estimates rather than an item easiness parameter estimates.

Model 1b. Impact Model. In the second model, or impact model, I estimate a multilevel Rasch model with one predictor at level-2. The predictor is included at the examinee-level (i.e., level-2) to examine whether there is an overall difference in information literacy ability between examinees in the NOTA and non-NOTA groups. The first level is equal to Equation 1. At the second level, a random effect for the intercept (ability) is still included, but, in contrast to the unconditional model at level-2, the impact predictor is included:

$$\begin{cases} \beta_{0j} = \gamma_{01}(NOTA)_j + u_{0j} \\ \beta_{1j} = \gamma_{1,0} \\ \beta_{2j} = \gamma_{2,0} \\ \vdots \\ \beta_{30} = \gamma_{30,0} \end{cases} \quad (6)$$

The impact estimate, γ_{01}, represents the logit difference in abilities between the NOTA and non-NOTA groups. The NOTA group was coded as 1 $(NOTA)_j = 1$), and the non-NOTA group was coded as 0 $(NOTA)_j = 0$). If the impact estimate is positive, then

the average NOTA group ability is γ_{01} logits higher than the average non-NOTA group ability. If the impact estimate is negative, then the average NOTA group ability is γ_{01} logits lower than the average non-NOTA group ability. I tested whether the impact estimate was statistically significantly different than 0 (i.e., whether there was a statistically significant difference in average information literacy ability between the NOTA and non-NOTA groups) with an $\alpha = .05$. If the p-value of the impact estimate associated with the t-test is less than .05, then the impact estimate is statistically significantly different than 0. The equations at both levels can be written as a single equation,

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \gamma_{01}(NOTA)_j + u_{0j} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \cdots + \gamma_{30,0}X_{30ij} \quad (7)$$

The impact estimate from Model 1b (γ_{01}) was subsequently used in Model 1c.

Model 1c. DIF Model. In the third model, or DIF model, I estimated a multilevel Rasch model with 10 predictors at level-2. The 10 predictors were included at the examinee-level to examine whether item effects vary as a function of examinees being in the NOTA group. The first level is equal to Equation 1. At the second level, a random effect for the intercept (ability) is still included, but, in contrast to the unconditional model at level-2, a DIF coefficient $\gamma_{i,1}$ is estimated for the 10 information literacy items that contained NOTA. The impact estimate γ_{01}, which is estimated in Model 1b and entered into Equation 8, is subtracted from the 10 DIF estimates γ_{i1} to control for impact. In Equation 8, I include effects for each of the items to display which items contain NOTA (i.e., items 1, 3, 6, 10, 11, 13, 14, 15, 23, and 30).

The DIF estimates γ_{i1} represent the difference in item difficulty between the NOTA and non-NOTA groups after controlling for differences in ability between the

NOTA and non-NOTA groups (i.e., impact). Because the NOTA group was coded as 1 $(NOTA)_j = 1$), and the non-NOTA group was coded as 0 $(NOTA)_j = 0$), a negative DIF coefficient indicates that the item was easier for the NOTA group than the non-NOTA group. A positive DIF coefficient indicates that the item was harder for the NOTA group than the non-NOTA group. As with the impact estimate, I tested whether the DIF coefficient were statistically significantly different than 0 (i.e., whether there was a statistically significant difference in item difficulty between the NOTA and non-NOTA groups, controlling for impact) with an $\alpha = .05$. If the p-value of the DIF coefficient was less than .05, then the DIF estimate is statistically significantly different than 0 and the item is flagged as exhibiting DIF. Additionally, I did not make adjustments to account for an increase in Type I error rate for the multiple statistical significance tests.

Model Comparison. In addition to examining whether DIF estimates were statistically significantly different than 0, I compare model-data fit with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), likelihood ratio tests (LRTs), and changes in variance of examinee abilities. I first compare the AIC and BIC values across Model 1a, Model 1b, and Model 1c, and the model with the smallest AIC and BIC is evidence of superior model-data fit. I also compare model-data fit using LRTs. Using deviances and degrees of freedom associated with each model, I perform a series of model comparisons. If there is a statistically significant difference (i.e., $p < .05$) between models, then this is evidence of superior model-data fit of the more complex model. Finally, I compare the changes in variance of examinee abilities across models. As more predictors are added to a model, the variance in examinee abilities is expected to decrease due to the inclusion of predictors (i.e., impact parameter and DIF parameters).

$$
\begin{cases}
\beta_{0j} = u_{0j} \\
\beta_{1j} = \gamma_{1,0} + (\gamma_{1,1} - \gamma_{01})(NOTA)_j \\
\beta_{2j} = \gamma_{2,0} \\
\beta_{3j} = \gamma_{3,0} + (\gamma_{3,1} - \gamma_{01})(NOTA)_j \\
\beta_{4j} = \gamma_{4,0} \\
\beta_{5j} = \gamma_{5,0} \\
\beta_{6j} = \gamma_{6,0} + (\gamma_{6,1} - \gamma_{01})(NOTA)_j \\
\beta_{7j} = \gamma_{7,0} \\
\beta_{8j} = \gamma_{8,0} \\
\beta_{9j} = \gamma_{9,0} \\
\beta_{10j} = \gamma_{10,0} + (\gamma_{10,1} - \gamma_{01})(NOTA)_j \\
\beta_{11j} = \gamma_{11,0} + (\gamma_{11,1} - \gamma_{01})(NOTA)_j \\
\beta_{12j} = \gamma_{12,0} \\
\beta_{13j} = \gamma_{13,0} + (\gamma_{13,1} - \gamma_{01})(NOTA)_j \\
\beta_{14j} = \gamma_{14,0} + (\gamma_{14,1} - \gamma_{01})(NOTA)_j \\
\beta_{15j} = \gamma_{15,0} + (\gamma_{15,1} - \gamma_{01})(NOTA)_j \\
\beta_{16j} = \gamma_{16,0} \\
\beta_{17j} = \gamma_{17,0} \\
\beta_{18j} = \gamma_{18,0} \\
\beta_{19j} = \gamma_{19,0} \\
\beta_{20j} = \gamma_{20,0} \\
\beta_{21j} = \gamma_{21,0} \\
\beta_{22j} = \gamma_{22,0} \\
\beta_{23j} = \gamma_{23,0} + (\gamma_{23,1} - \gamma_{01})(NOTA)_j \\
\beta_{24j} = \gamma_{24,0} \\
\beta_{25j} = \gamma_{25,0} \\
\beta_{26j} = \gamma_{26,0} \\
\beta_{27j} = \gamma_{27,0} \\
\beta_{28j} = \gamma_{28,0} \\
\beta_{29j} = \gamma_{29,0} \\
\beta_{30} = \gamma_{30,0} + (\gamma_{30,1} - \gamma_{01})(NOTA)_j
\end{cases}
\tag{8}
$$

Research Question 2

The goal of the second research question is to examine whether examinees exhibit a tendency to select NOTA. To answer this research question, I compared model-data fit of a two-parameter logistic (2PL) IRT model (i.e., a model without the inclusion of a selection tendency toward NOTA) and an item response tree model (IRTree; i.e., a model

with the inclusion of a selection tendency toward NOTA). I only retained item responses from examinees who received a test with NOTA (NOTA group) as I would not be able to model a selection tendency toward NOTA for examinees who received a test without NOTA (non-NOTA group). Recall that examinees in the NOTA group completed an information literacy test with 10 items that contained NOTA. While the goal of the second research question is to explore whether examinees have a selection tendency towards NOTA, in practice, the primary construct of interest would be information literacy. In the following sections, I refer to information literacy as the trait of interest (TOI) and the selection tendency towards NOTA as the NOTA selection tendency (NST).

The IRTree that I used is similar to Deng and Bolt's (2016) SRM-MC. However, rather than isolate propensity of selection of the first option, I isolate propensity of selection towards the NOTA option. In the current study, the NOTA option was both the last option and incorrect option across all 10 NOTA items. The reader is directed to the literature review for greater detail of the SRM-MC.

In the following sections, I describe two models to test the NST trait and estimation of the models. Under the first model, only the TOI is estimated. This model is known as the Non-NOTA Selection Tendency Model. In other words, only the TOI influences examinee item responses. Under the second model, both the NST and TOI traits are estimated. This model is known as the NOTA Selection Tendency Model. Not only does the TOI influence examinee item responses, but the NST trait also influences examinee item responses. After presenting the models, I discuss how I will estimate the models, evaluate fit indices to determine which model has better model-data fit, and utilize various measures to establish validity of the NST and TOI.

Model 2a. Non-NOTA Selection Tendency Model. In the first model, or the non-NOTA selection tendency model, I estimated a 2PL IRT model. Under this model, I assumed that there was no selection tendency toward NOTA and that only the latent TOI is hypothesized to influence item responses across the 30 information literacy test items. Specifically, the probability of examinee i correctly answering item j is

$$P(x_{ij} = 1|\theta_i, \alpha_j, b_j) = \frac{1}{1+e^{-\alpha_j(\theta_i - b_j)}}. \quad (9)$$

Under the 2PL model, θ_i is an examinee ability parameter and represents the examinees proficiency on the latent TOI. Additionally, α_j is an item discrimination parameter and b_j is an item difficulty parameter. The a_j parameter represents how well the item is able to distinguish amongst examinees across the TOI continuum, and the b_j parameter indicates the location on the TOI continuum where examinees have 50% chance of correctly answering item i (de Ayala, 2009). Given that there are a total of 30 items on the information literacy test, I estimated 30 α_j parameters and 30 b_j parameters.

Model 2b. NOTA Selection Tendency Model. In the second model, or the NOTA Selection Tendency Model, I fit an IRTree model to the data. Under this model, I assumed that there was a selection tendency toward NOTA and that both the NST and TOI are hypothesized to influence item responses. This assumption is specific to the 10 items which contain NOTA. The item responses on the other 20 items on the information literacy test that do not contain NOTA are hypothesized to only be influenced by the latent TOI. In the following sections, I describe how item responses to the 10 items that contain NOTA are modeled with the IRTree. The item responses to the other 20 items that do not contain NOTA are simply modeled with the 2PL IRT model as described in the non-NST model.

Figure 3 provides a visual of the hypothesized response process modeled using the IRTree. At Stage 1, examinees are hypothesized to either select NOTA with probability, P_{ijNST}, or not select NOTA with probability, $1 - P_{ijNST}$. At this stage, examinee decisions to select NOTA are assumed to be driven by the NST latent trait (θ_{iNST}). If an examinee does not select NOTA, they are hypothesized to go to Stage 2. At this stage, the TOI is assumed to be driving the selection of the correct option with probability, P_{ijTOI} or the incorrect answer with probability, $1 - P_{ijTOI}$.

At each stage, I modeled decisions using the 2PL IRT model. At Stage 1, the probability of examinee i selecting NOTA on item j is

$$P_{ijNST} = \frac{1}{1+e^{-a_{jNST}\left(\theta_{iNST}-b_{jNST}\right)}} \tag{10}$$

The parameter a_{jNST} is the discrimination parameter for item j at Stage 1, which represents the slope of the item category characteristic curve at a probability .50. Higher values of a_{jNST} indicate that the item does a better job at distinguishing examinees who have high NST from examinees who have low NST. In contrast, lower value of a_{jNST} indicate that the item does a poor job at distinguishing amongst examinees who have differing levels of NST. The parameter b_{jNST} is the difficulty parameter for item j at Stage 1, which represents the point on the NOTA ability continuum (θ_{iNST}) where examinees have a 50% chance of selecting NOTA. Items with high b_{jNST} values require examinees had higher NST (θ_{iNST}) to answer the item correctly than items with low b_{jNST} values.

In Stage 2, the probability of examinee i selecting the correct answer on item j is

$$P_{ijTOI} = \frac{1}{1+e^{-a_{jTOI}\left(\theta_{iTOI}-b_{jTOI}\right)}}. \tag{11}$$

The parameter a_{jTOI} is the discrimination parameter for item j at Stage 2, which represents the slope of the item characteristic curve at a probability of .50. Higher values of a_{jTOI} indicate that the item does a better job at distinguishing amongst examinees with high and low TOI. In contrast, low values of a_{jTOI} indicate that the item does a poor job at distinguishing amongst examinees with high and low TOI. The parameter b_{jTOI} is the difficulty parameter for item j at Stage 2, which represents the point on the TOI continuum (θ_{iTOI}) where examinees have a 50% chance of selecting the correct option.

As displayed in Figure 3, the arrows that extend from each latent trait are represented in probabilistic terms. The hypothesized stage decisions are independent, resulting in the probability of terminal choices equal to products of the branch probabilities. Specifically, the probability of examinee i selecting NOTA on item j is

$$P\left(U_{ij} = NOTA \middle| \theta_{iNST}, \theta_{iTOI}\right) = P_{ijNST}. \quad (12)$$

The probability of selecting the correct option is

$$P\left(U_{ij} = correct \middle| \theta_{iNST}, \theta_{iTOI}\right) = \left(1 - P_{ijNST}\right)\left(P_{ijTOI}\right), \quad (13)$$

and the probability of selecting an incorrect option that is not NOTA is

$$P\left(U_{ij} = incorrect \middle| \theta_{iNST}, \theta_{iTOI}\right) = \left(1 - P_{ijNST}\right)\left(1 - P_{ijTOI}\right). \quad (14)$$

Model Estimation. Both models were estimated under a Bayesian framework using the Markov chain Monte Carlo (MCMC) procedure in SAS 9.4M7 (SAS Institute Inc., 2020). For both models, I set uninformative prior distributions for the discrimination (i.e., $a \sim lognormal(0,1)$ with a lower bound of 0) and difficulty (i.e., $b \sim N(0,4)$) with initial values of 1 and 0, respectively. I set the prior distributions for abilities to follow a multivariate normal distribution with a mean vector of **0**. I also specified the variances of the prior to be 1 and freely estimated the correlation among the traits. I used an

uninformative prior distribution for the correlation between θ_{iNST} and θ_{TOI}, $\rho_{\theta_{iNST}\theta_{iTOI}} \sim N(0, 1)$ with a lower bound of -1 and upper bound of 1. The total number of iterations was set to 100,000 with a burn-in of 20,000.

Model Convergence. To evaluate the convergence of parameters to stable posterior distributions, I evaluated both visual and statistical criteria. In terms of visual criteria, I examined trace plots and autocorrelation plots for all parameters. Trace plots display the degree to which sampled parameters across iterations traverse the posterior parameter space, and autocorrelation plots display the degree of dependence of sampled parameters across lags (Stone & Zhu, 2015). If a trace plot displays sampled parameters that adequately vary across iterations (i.e., randomly traverse the parameter space), then this is indication that the Markov chain has mixed well. If an autocorrelation plot displays correlations across lags that become close to 0, then this is indication that Markov chain is efficient (Stone & Zhu, 2015). In terms of statistical criteria, I examined the Geweke (1992) statistic. The Geweke statistic is a measure of convergence which allows for the comparison of parameter means across iterations (Stone & Zhu, 2015). Using the Geweke statistic, statistical significance is calculated using a *z*-test comparing the first 10% of iterations to the last 50% of iterations, where a *p*-value less than or equal to .05 indicates evidence of a lack of convergence. A *p*-value above .05 indicates evidence of convergence.

Model Comparison. After evaluating convergence and trait validity interpretations, I examined model-data fit. I compared the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) associated with both models. The DIC is

expressed as the posterior mean of deviance across iterations ($\overline{D}(\boldsymbol{\delta})$) with the addition of a penalty for model complexity (p_D):

$$DIC = \overline{D}(\boldsymbol{\delta}) + p_D \tag{15}$$

As described by Stone and Zhu (2015), the posterior mean of the deviance across iterations ($\overline{D}(\delta)$) is a measure of model-data fit, which uses the log likelihood function ($-2logL(D|\boldsymbol{\delta})$) and sampled $\boldsymbol{\delta}$ parameter values at each iteration. The penalty for model complexity (p_D) is also computed using the $\boldsymbol{\delta}$ parameter values. Specifically, p_D is the difference between the posterior mean of deviances across iterations and the deviance from the log likelihood function for the posterior mean of the parameter estimates ($D(\bar{\delta})$).

The model associated with the smaller DIC indicates a better fitting model. The amount of difference in the DIC between models is associated with more meaningful differences. Specifically, Stone and Zhu (2015) cite Lunn et al. (2013), who state that differences greater than 10 indicate "important" differences between models, differences between 5 and 10 indicate "substantial" differences between models, and differences less than 5 indicate that the models have similar fit. I used the DIC values from the Model 2a and Model 2b to determine which model to champion. If Model 2b had a smaller DIC than Model 2a, then I championed Model 2b. If Model 2b has a larger DIC than Model 2a, then I championed Model 2a. Further, I followed Lunn et al.'s (2013) criteria to describe the degree of model-data fit differences.

Trait Validity. To adequately infer that the θ_{iNST} and θ_{iTOI} estimates were a NOTA selection tendency and information literacy, respectively, under Model 2a and Model 2b, I estimated the relationship between the latent traits and relevant variables. I first estimated the correlation between θ_{iNST} and the total NOTA selection. I next

estimated the correlation between θ_{iNST} and attitudinal NOTA measures. Total NOTA selection was calculated as the total number of times an examinee selected NOTA and ranged from 0 (i.e., the examinee did not select NOTA) to 10 (i.e., the examinee selected NOTA across all NOTA 10 items). I hypothesize that total NOTA selection and θ_{iNST} share a strong, positive relationship. The attitudinal NOTA items from the NOTA Perception Survey items were reverse scored so that "Strongly disagree" was recoded as 1 and "Strongly agree" was recoded as 5. The attitudinal items asked examinees to report whether they "Strongly disagree" to "Strongly agree" to being drawn to select NOTA, avoid selecting NOTA, and indifference toward selection NOTA. I hypothesized that examinees who reported higher endorsement of being drawn to select NOTA have higher θ_{iNST} (i.e., positive correlation) while examinees who reported higher endorsement of avoiding NOTA have lower θ_{iNST} (i.e., negative correlation). Indifference toward NOTA is hypothesized to be unrelated to θ_{iNST}.

For the θ_{iTOI}, I estimated the correlation between θ_{iTOI} and total number correct on the information literacy test. The 2-PL IRT model is used to estimate θ_{iTOI}, so total number correct will not be perfectly related to TOI. However, I hypothesized that θ_{iTOI} and total number correct had a strong, positive correlation.

Results

The results are organized in three sections. The first section is titled Descriptive Statistics. I report and compare descriptive statistics for the cognitive and non-cognitive measures between the group of examinees who received a test with NOTA and the group of examinees who did not receive a test with NOTA. Specifically, I filter examinees based on effort scores to ensure valid responses. I then examine group differences in average scores and perform an item analysis to explore differences in item statistics between the NOTA and non-NOTA groups. I also report descriptive statistics for the attitudinal measures. The second section is titled Research Question 1. This section is divided into three parts. Each part includes results of the three multilevel Rasch models discussed in the Methodology section (i.e., Unconditional Model, Impact Model, and DIF Model). The third section is titled Research Question 2. This section is also divided into three parts. The first two parts include results of the Non-NOTA Selection Tendency Model (2PL) and the NOTA Selection Tendency Model (IRTree). In the third part I focus on model comparison results between the Non-NOTA Selection Tendency Model and the NOTA Selection Tendency Model.

Descriptive Statistics

Effort Scores

A total of 395 examinees participated in the current study. Of the 395 examinees, 195 completed a test with NOTA (i.e., NOTA group), and 200 completed a test without NOTA (i.e., non-NOTA group). To ensure valid responses, I filtered examinees based on their reported effort put forth during the test. Effort scores were reverse scored so that a score of 5 corresponded to higher effort and a score of 1 corresponded with lower effort.

An effort total score was calculated for each examinee, and examinees who scored below 15 were removed from the sample for lack of effort. On average, the NOTA group ($M = 19.72, SD = 3.78$) reported slightly lower effort scores than the non-NOTA group ($M = 20.36, SD = 3.03$). However, this difference was not statistically significant, $t(393) = -1.86, p = .063$. A total of 17 examinees were removed from the NOTA group, and a total of 9 examinees were removed from the non-NOTA group. This resulted in a total sample size of 369 examinees with 178 examinees in the NOTA group and 191 in the non-NOTA group.

Information Literacy Scores

Total Scores. After filtering examinees based on effort, I performed an item and distractor analysis in the classical test theory framework to examine differences in item functioning between the NOTA and non-NOTA groups. Overall, the coefficient alpha was smaller for the NOTA group ($\alpha = .687$) compared to the non-NOTA group ($\alpha = .748$). On average, the NOTA group ($M = 20.39, SD = 3.90$) scored .49 points higher than the non-NOTA group ($M = 19.90, SD = 4.40$), which was not a statistically significant difference, $t(367) = 1.12, p = .263$.

Item Analysis Results. I performed an item analysis to compare item difficulty (i.e., the proportion of examinees who selected the correct option) and item discrimination (i.e., the point-biserial correlation between correct option selection and total score) between the NOTA and non-NOTA groups for the 30 information literacy items. In the current study, I assumed that NOTA would not influence examinee item responses to non-NOTA items. Therefore, I hypothesized there to be larger differences in NOTA item statistic differences between the NOTA and non-NOTA groups compared to

the non-NOTA item statistics. The item difficulty and item discrimination are presented Figure 4 and Figure 5, respectively. I visually inspected the panels in Figure 4 to determine whether there were meaningful differences in item difficulty and item discrimination between the NOTA and non-NOTA groups across the NOTA and non-NOTA items.

Figure 4 displays the item difficulties for the NOTA and non-NOTA groups. Item difficulty differences greater than .05 were evidence of meaningful differences between the NOTA and non-NOTA groups as a difference of .05 is approximately 20 examinees in the overall sample. In other words, there was a .05 difference in the proportion of individuals in each group who selected the correct option. For example, fewer NOTA group examinees (.71) selected the correct option on item 1 compared to the non-NOTA group examinees (.79). The greatest item difficulty difference was associated with item 6, where NOTA group examinees (.84) more frequently selected the correct option compared to the non-NOTA group examinees (.72). Overall, six (i.e., items 1, 6, 11, 15, 23, and 30) of the 10 NOTA items exhibited meaningful item difficulty differences, and only four (i.e., items 7, 20, 27, and 28) of the 20 non-NOTA items exhibited meaningful item difficulty differences. Given the study is on the effects of NOTA, I primarily focused on differences in item statistics for experimental NOTA items.

To gain more insight, I examined the direction of the meaningful differences. In other words, I investigated which group had a higher frequency of selecting the correct option. Of the NOTA items associated with meaningful item difficulty differences (i.e., items 1, 6, 11, 15, 23, and 30), the NOTA group more frequently selected the correct option on four of the items (i.e., items 6, 15, 23, and 30). For example, on item 15, 87%

of NOTA group examinees as compared to only 81% of non-NOTA group examinees selected the correct option. On all four of the non-NOTA items associated with meaningful item difficulty differences, the NOTA group more frequently selected the correct option. Generally, when meaningful differences existed, the NOTA group tended to select the correct answer more frequently than the non-NOTA group.

Figure 5 displays item discrimination for the NOTA and non-NOTA groups. To make interpretations feasible, I highlight the item discrimination differences greater than .10. Squaring .10 results in an R^2 of .01, which Cohen (1988) describes as a small effect. While considered a small effect, I use a value of .10 as a meaningful effect to describe the differences in item discrimination. As an example, the NOTA group item discrimination (.29) and non-NOTA group item discrimination (.47) for item 1 differed by .18, indicating that the point-biserial correlation between total score and correct option selection was meaningfully weaker for the NOTA group. Item 11 exhibited the greatest item discrimination difference, where item discrimination for the NOTA group (.08) was lower than item discrimination for the non-NOTA group (.33). Five (i.e., items 1, 3, 11, 12, and 15) of the 10 NOTA items and eight (i.e., items 2, 7, 13, 16, 20, 22, 26, and 27) of the 20 non-NOTA items were associated with meaningful item discrimination differences.

Finally, I examined whether the NOTA group item discrimination was greater than or less than the non-NOTA group item discrimination for items with meaningful differences. Four (i.e., item 1, 3, 11, and 12) of the five NOTA items associated with meaningful item discrimination differences exhibited lower item discrimination for the NOTA group as compared to the non-NOTA group. Six (i.e., items 7, 13, 16, 20, 22, and

26) of the eight non-NOTA items associated with meaningful item discrimination differences exhibited lower item discrimination for the NOTA group as compared to the non-NOTA group. Overall, the NOTA group tended to result in lower item discrimination than the non-NOTA group.

Distractor Analysis Results. I next performed a distractor analysis on the NOTA items to examine the differences in option statistics between the NOTA and non-NOTA groups. The option difficulty and discrimination differences for the NOTA items between the NOTA and non-NOTA groups are presented in Figure 6. In this section, I focus on the fourth option as the only difference on the information literacy test between the NOTA and non-NOTA groups. Similar to the previous section, I compare the NOTA option and fourth option by identifying meaningful differences using the values of .05 and .10 as cut-offs for the proportion of selection and option discrimination, respectively.

Of the 10 NOTA items, three items (i.e., items 1, 6, and 11) display evidence of meaningful fourth option frequency selection differences. Whereas the NOTA option was selected more frequently than the fourth option for item 1 and item 11, the NOTA option was selected less frequently than the fourth option for item 6.

Of the 10 NOTA items, five items (i.e., items 1, 3, 10, 11, and 14) display evidence of meaningful differences between the NOTA option discrimination and the fourth option discrimination. Whereas NOTA option discrimination was closer to 0 than fourth option discrimination for items 1, 3, 11, and 14, NOTA option discrimination was more negative than fourth option discrimination for item 10. Interestingly, for item 3, NOTA option discrimination was positive and close to 0 (.04), but fourth option

discrimination was negative (-.51). Overall, the NOTA option discriminated amongst examinees less than the fourth option in the non-NOTA group.

When examining just items in the NOTA group, there was variation in the frequency of NOTA selection (see the NOTA options in Figure 6). For example, item 1 had the greatest NOTA selection ($n = 36$), and five of the 10 items had fewer than 5 examinees select the NOTA option (i.e., items 3, 6, 10, 15, and 30). Total NOTA selection was low across items (see Figure 7). Of the 178 NOTA group examinees, 123 NOTA group examinees did not select NOTA, 39 selected NOTA once, 12 selected NOTA twice, 3 selected NOTA three times, and only 1 selected NOTA nine times out of the 10 NOTA opportunities.

Perceptions and Attitudes Toward NOTA

In addition to the information literacy items, examinees reported their perception of the number of items that contained NOTA and their attitudes toward NOTA. The perceived number of NOTA items reported by the non-NOTA and NOTA groups are displayed in Figure 8. A greater percentage of non-NOTA group examinees (76%) than NOTA group examinees (26%) reported that 0 of the items contained NOTA. For the NOTA group, 8% of examinees reported that one item contained NOTA, 6% reported that two items contained NOTA, 9% reported that five items contained NOTA, and 8% reported that 10 items contained NOTA. For the non-NOTA group, 7% of examinees reported that one item contained NOTA, 7% reported that two items contained NOTA, 1% reported that five items contained NOTA, and 1% reported 10 items contained NOTA. In general, a greater percentage of the NOTA group examinees reported more perceived NOTA items.

Overall, for the non-NOTA group, most examinees accurately perceived that there were no NOTA items present on their information literacy test. However, for the NOTA group, there was a greater discrepancy between the actual (i.e., 10) and perceived number of NOTA items. While most examinees did notice that NOTA was present, only 8% of NOTA group examinees accurately reported that there were 10 NOTA items on the information literacy test.

In addition to perceived number of NOTA items, examinees reported whether they felt drawn toward selecting NOTA, an avoidance toward selecting NOTA, and an indifference toward selecting NOTA. Examinees responded to the NOTA attitudinal survey on a 5-point Likert scale from 1 ("Strongly agree") to 5 ("Strongly disagree") after reverse scoring responses so that higher values indicated stronger endorsement of the statement. Table 2 provides average examinee responses on these three attitudinal items.

The non-NOTA group ($M = 3.08, SD = 1.10$) compared to the NOTA group ($M = 2.56, SD = 1.05$) reported significantly stronger endorsement of feeling drawn toward selecting NOTA, $t(367) = 5.08, p < .001$. In contrast, the non-NOTA group ($M = 2.88, SD = .97$) compared to the NOTA group ($M = 3.26, SD = .98$) reported significantly weaker endorsement of avoidance toward selecting NOTA, $t(367) = -3.74, p < .001$. Both the non-NOTA group ($M = 2.99, SD = .93$) and NOTA group ($M = 3.01, SD = .93$) reported feeling neutral about an indifference toward NOTA, $t(367) = -.17, p = .865$. In other words, the NOTA and non-NOTA groups differed in feeling drawn to and avoidance of selecting NOTA. I should note that examinees completed the NOTA attitudinal survey after completing the information literacy test. The presence of NOTA on the information literacy test may have influenced responses

from the NOTA group. This effect, and how it potentially influenced responses to information literacy NOTA items, is explored in detail in the Discussion.

I further examined the relationship between the attitudinal item responses and the total number of times an examinee selected NOTA. There was a positive, weak relationship between total NOTA selection and being drawn toward selecting NOTA ($r = .243,\ p < .001$), indicating that higher NOTA selection is associated with higher feelings of being drawn toward selecting NOTA. There was a negative, weak relationship between total NOTA selection and avoidance toward selecting NOTA ($r = -.256, p < .001$), indicating that higher NOTA selection is associated with lower feelings of an avoidance toward selecting NOTA. Total NOTA selection and indifference toward selecting NOTA also had a negative, weak relationship, but it was not statistically significant ($r = -.119, p = .099$).

Research Question 1

The goal of the first research question is to test whether item difficulty of the NOTA items is different between the NOTA and non-NOTA groups, after controlling for differences in information literacy ability. To answer this research question, I present results from three multilevel Rasch models: Model 1a (i.e., Unconditional Model), Model 2a (i.e., Impact Model), and Model 3a (i.e., DIF Model).

Model 1a. Unconditional Model

Under the unconditional model, item difficulty estimates were fixed across examinees ($N = 368$) and no predictors were included in the model. A total of 190 examinees were in the non-NOTA group and 178 examinees were in the NOTA group. Of interest were the fixed effect item difficulty parameters associated with each of the 30

items ($\gamma_{1,0}, \gamma_{2,0}, \ldots, \gamma_{30,0}$) and the variability in the random effect associated with examinees (u_{oj}), which resulted in a total of 31 parameter estimates. Table 3 contains the fixed effects and variability in random effect parameter estimates, standard errors, and t-values from Model 1a. The item difficulty parameters from Model 1a are estimated using the total sample. In other words, the estimates are common, or fixed, across all examinees, regardless of whether they were in the NOTA or non-NOTA group. The hardest item was item 27 ($\gamma_{27,0} = 1.69$), where examinees with an ability of 1.69 had a 50% chance of correctly answering item 27. The easiest item was item 13 ($\gamma_{13,0} = -3.41$), where examinees with an ability of -3.41 had a 50% chance of correctly answering the item.

Model 1b. Impact Model

Under the impact model, item difficulties were fixed across examinees ($N = 395$), and a predictor to examine impact (γ_{01}), or the difference in abilities between the NOTA and non-NOTA groups, was included at the second level. The parameter of interest was the impact parameter, which was then used when estimating the DIF model (i.e., Model 1c) to control for impact.

The item difficulty parameter estimates under Model 1a and Model 1b changed slightly (e.g., the difference in $\gamma_{1,0}$ between Model 1a and Model 1b was .02), but the estimates were perfectly correlated ($r = 1.00$). In other words, the rank ordering of the item difficulty estimates did not change from Model 1a to Model 1b. The impact parameter estimate was .09 ($SE = .09$), indicating that the average NOTA group ability was . 09 logits greater than the average non-NOTA group ability. While the estimate was not statistically significantly different than 0, $t(368) = 1.07, p = .28, CI: [-.08, .27]$, I

used the impact estimate in the DIF model to control for NOTA and non-NOTA group differences in ability.

Model 1c. DIF Model

Under the DIF model, 10 predictors were included at the examinee-level to examine whether item effects varied as a function of examinees being in the NOTA group. The impact estimate (i.e., $\gamma_{01} = .09$) from Model 1b was subtracted from the DIF estimate to control for differences in average examinee ability between the NOTA and non-NOTA groups. In contrast to Model 1a, there are two sets of item difficulty estimates based on the NOTA group and non-NOTA group as item difficulty was free to vary between groups for the NOTA items. The non-NOTA group and NOTA group item difficulty estimates are presented in Table 3 under Model 1c. Model 1a item difficulty estimates were strongly related to the Model 1c non-NOTA group item difficulty estimates ($r = .997$). The Model 1a item difficulty estimates were also strongly related to the Model 1c NOTA item difficulty estimates ($r = .996$).

A total of 10 DIF parameters were estimated (i.e., $\gamma_{1,1}$, $\gamma_{3,1}$, $\gamma_{6,1}$, $\gamma_{10,1}$, $\gamma_{11,1}$, $\gamma_{13,1}$, $\gamma_{14,1}$, $\gamma_{15,1}$, $\gamma_{23,1}$, $\gamma_{30,1}$). The DIF parameter estimates represent the difference in item difficulty between the NOTA and non-NOTA groups, after controlling for impact. Three of the 10 DIF parameter estimates were statistically significant. Specifically, item 1, $t(368) = 2.71, p = .007$), item 6, $t(368) = -2.30, p = .022$, and item 11, $t(368) = 2.71, p = .007$, were flagged for DIF.

The DIF estimates associated with item 1 ($\gamma_{1,1} = .351$) and item 11 ($\gamma_{11,1} = .351$) were both positive, indicating that item 1 and item 11 were harder for the NOTA group, after controlling for impact. In other words, NOTA group examinees needed a

higher ability to have the same probability of correctly answering item 1 and item 11 as compared to non-NOTA group examinees (see Figure 9 and Figure 10, respectively). To obtain the NOTA group item difficulty estimate, I added the non-NOTA group item difficulty estimate and the DIF estimate. Table 4 contains the item difficulty estimates from Model 1a, the DIF estimates from Model 1c, and the item difficulty estimates for the non-NOTA and NOTA groups. For item 1, I added 0.351 to the non-NOTA group item 1 difficulty estimate ($\gamma_{1.0} = -1.49$) to obtain the NOTA group item 1 difficulty estimate of -1.14. For item 11, I added 0.351 to the non-NOTA group item 11 difficulty estimate (i.e., $\gamma_{1.0} = -0.24$) to obtain the NOTA group item 11 difficulty estimate of 0.11.

The DIF estimate associated with item 6 ($\gamma_{6,1} = -0.636$) was negative, indicating that item 6 was easier for the NOTA group, after controlling for impact. In other words, NOTA group examinees had a higher chance of correctly answering item 6, even though they may have the same ability as non-NOTA group examinees (see Figure 11). As with items 1 and 11, I added the non-NOTA group item difficulty estimate ($\gamma_{6.0} = -1.08$) and the DIF estimate ($\gamma_{6,1} = -.636$) to obtain the NOTA group item 6 difficulty estimate (-1.71).

Model Comparisons

In addition to examining whether the DIF estimates were statistically significantly different than 0, I compared model-data fit with information criteria indices (i.e., AIC and BIC), LRTs, and changes in variance of examinee abilities. Table 5 contains the model comparison indices. The AIC and BIC are smallest for Model 1a, indicating superior fit to Model 1b and Mode lc. The LRTs align with the AIC and BIC conclusions. Model 1b (i.e., a model with an impact parameter) did not fit better than Model 1a (i.e., a model with no predictors), $X^2(1) = 1, p = .317$. Additionally, Model 1c (i.e., a model with DIF parameters) did not fit better than Model 1a, $X^2(10) = 18, p = .055$. The variance in examinee abilities (i.e., $var(u_{0j})$) for Model 1a, Model 1b, and Model 1c are provided in Table 3. There was little to no difference in variation of examinee abilities across models. Based on the model-data fit indices and variation in examinee abilities across models, Model 1a exhibits superior fit compared to Model 1b and Model 1c. In other words, there is evidence to suggest that the inclusion of the impact parameter and DIF parameters did not contribute to improving model-data fit.

Research Question 2

The goal of the second research question is to test the hypothesis that there is a selection tendency toward NOTA. I present item and person parameter results of the 2PL IRT model (i.e., Model 2a: Non-NOTA Selection Tendency Model) and the IRTree model (i.e., Model 2b: NOTA Selection Tendency Model). I then compare model-data fit of Model 2a and Model 2b.

Model 2a: Non-NOTA Selection Tendency Model

Prior to examining item and person parameters, I ensured that parameters converged to a stable posterior distribution by examining Geweke statistics, autocorrelation plots and lag values, and trace plots using initial MCMC algorithm parameters (e.g., 100,000 iterations). Multiple parameters had Geweke statistic p-values less than .05, indicating that convergence was not achieved. I also explored the autocorrelation plots and lag values, which were all close to 0. This was an indication that later iterations in the chain were not dependent on previous iterations. Further, the trace plots for these parameters tended to stay in one area of the parameter space rather than bounce around to ensure adequate coverage. Therefore, I increased the number of iterations from 100,000 to 200,000. This resulted in no Geweke statistic p-values less than .05, adequate coverage of the parameter space as evidenced by trace plots, and autocorrelation plots and lags with low correlations.

Table 6 includes the expected a posterior (EAP), standard deviation, and 95% Highest Posterior Density (HPD) intervals of the item discrimination and item difficulty posterior distributions. The average item discrimination estimate was 0.92 with the most discriminating item being item 19 ($a_{19} = 3.47$) and the least discriminating item was item 8 ($a_8 = 0.264$). The average item difficulty was -0.88 with the easiest item being item 22 ($b_{22} = -3.10$) and the hardest item was item 27 ($b_{27} = 3.21$).

In addition to 30 item discrimination and 30 item difficulty parameters, a person parameter was estimated for each examinee. In the context of Model 2a, theta represented information literacy ability or the trait of interest (i.e., θ_{TOI}). TOI EAP estimates ranged from -2.70 to 1.93 with a mean of 0.06 and standard deviation of 0.79 (see Figure 12).

Model 2b: NOTA Selection Tendency Model

There were similar convergence issues for Model 2b as with Model 2a (i.e., Geweke statistic p-values were less than .05). As with Model 2a, I increased the number of iterations from 100,000 to 200,000. This resulted in no Geweke statistic p-values less than .05, adequate coverage of the parameter space as evidenced by trace plots, and autocorrelation plots depicting lags with low correlations.

Table 7 includes the EAP, standard deviation, and 95% Highest Posterior Density (HPD) intervals of the stage-level item discrimination and item difficulty posterior distributions for Model 2b. In contrast to Model 2a where one item discrimination parameter and one item difficulty parameter were estimated per item, two item discrimination parameters and two item difficulty parameters were estimated for each of the 10 NOTA items under Model 2b. Specifically, Stage 1 item discrimination and item difficulty parameters were associated with the NST trait (θ_{NST}) while Stage 2 item discrimination and item difficulty parameters were associated with the TOI (θ_{TOI}). I first interpret the Stage 2 item parameters as the interpretations are similar to Model 2a item parameter interpretations.

At Stage 2, the average item discrimination estimate was 0.94. The most discriminating item was item 19 ($a_{19TOI} = 3.56$), indicating that item 19 did well at distinguishing amongst examinees on the TOI. The least discriminating item was item 8 ($a_{8TOI} = 0.26$), indicating that item 8 did not distinguish well amongst examinees. The average item difficulty was -0.95. The easiest item was item 6 ($b_{6TOI} = -3.18$), and the hardest item was item 27 ($b_{27TOI} = 3.25$).

At Stage 1, the average item discrimination was 2.37. The most discriminating item was item 10 ($a_{10NST} = 4.95$), and the least discriminating item was item 1 ($a_{1NST} = 0.79$). The average item difficulty was 2.76. The lowest item difficulty was item 11 ($b_{11NST} = 2.12$), indicating that examinees with a θ_{NST} of 2.12 had 50% chance of selecting NOTA on item 11. The highest item was item 3 ($b_{3NST} = 3.69$), indicating that examinees with a θ_{NST} of 3.69 had a 50% chance of selecting NOTA on item 3.

Model-Data Fit

Table 8 displays model-data fit statistics for Model 2a and Model 2b. A smaller DIC value indicates better model-data fit. The DIC under Model 2a is smaller than the DIC under Model 2b, which indicates that Model 2a had better model-data fit than Model 2b. The difference between DIC values is 295.525, indicating that Model 2a is the superior model (Lunn et al., 2013).

NOTA and TOI Trait Validity

The goal of the second research question was to test whether there is support for a construct representing the NST. As represented in Figure 3, there are two hypothesized latent traits estimated under Model 2b for each examinee. The first trait is the NOTA selection tendency trait (NST), and the second trait is the information literacy trait (TOI). I explored the relationships between the traits and relevant variables to support valid interpretations of the NST and TOI estimates as NOTA selection tendency and information literacy, respectively. I first calculated the correlation between the θ_{TOI} estimates under Model 2b and information literacy total score. I then calculated the correlation between the θ_{TOI} estimates under Model 2b and the θ_{TOI} estimates under Model 2a. Next, I calculated a correlation between the θ_{NST} estimates and total NOTA

selection. Finally, I examined the correlation between θ_{NST} estimates and NOTA attitudinal measures.

The θ_{TOI} estimates ranged from -2.61 to 1.97 with a mean of 0.06 and standard deviation of 0.79 (see Figure 13). TOI estimates under Model 2b were strongly related to total score on the information literacy test ($r = .95$), as well θ_{TOI} estimates under Model 2a ($r = .99$).

The θ_{NST} estimates ranged from -1.01 to 3.25 with a mean of -0.06 and standard deviation of 0.59. As depicted in Figure 14, the distribution of θ_{NST} was slightly positively skewed. The positive skew of the θ_{NST} distribution in Figure 12 aligns with the positive skew in the number of times an examinee selected NOTA (see Figure 7). When considering their relationship, it was positive; as θ_{NST} increases, the number of times an examinee selected NOTA increased. Examinees with lower θ_{NST} estimates represented examinees who did not select NOTA (i.e., lower total NOTA selection in Figure 7), and their estimate was being influenced by the prior and the covariation between θ_{NST} and θ_{TOI}. The variation among the θ_{NST} estimates was likely due to differences in θ_{TOI}.

The θ_{NST} estimates were positively related to total NOTA selection ($r = .86$). As seen in Figure 14, there is one extremely high θ_{NST} estimate compared to the others. The examinee's θ_{NST} was 3.25, and they selected NOTA on nine out of the 10 NOTA items, the most among all examinees. In addition to being related to total NOTA selection, θ_{NST} was related to attitudes toward NOTA. θ_{NST} was positively related to being drawn toward selecting NOTA ($r = .17$), negatively related to avoidance toward selecting NOTA ($r = -.22$), but not related to indifference toward selecting NOTA ($r = -.09$).

Recall that under Model 2b, I specified estimation of the correlation between θ_{TOI} and θ_{NST} ($r = -.44$). In addition to the EAP estimate of the correlation between θ_{TOI} and θ_{NST}, I calculated the correlation between Model 2b EAP θ_{TOI} estimates and EAP θ_{NST} estimates, which was also negative ($r = -.65$). Given that NOTA was always the incorrect option, the negative correlation between θ_{NST} and θ_{TOI} estimates was expected. A similar pattern was found between the θ_{NST} estimates and Model 2a θ_{TOI} estimates ($r = -.68$). Further, the number of times an examinee selected NOTA was negatively related to Model 2b θ_{TOI} estimates ($r = -.30$) and Model 2a θ_{TOI} estimates ($r = -.35$). The difference in correlations is an indication that the Model 2a θ_{TOI} estimates were confounded by the inclusion of NOTA. By modeling a NOTA selection tendency, the correlation between Model 2b θ_{TOI} estimates and θ_{NST} estimates and the correlation between the Model 2b θ_{TOI} estimates and total NOTA selection were less negative due to taking out the variance that could be attributed to the NOTA selection tendency.

Discussion

The purpose of the current study is to explore whether the presence of NOTA as a distractor in multiple-choice items influences examinees item responses. Specifically, NOTA is hypothesized to be a nuisance construct and result in construct irrelevant variance. Previous NOTA research has alluded to a potential, separate construct due to NOTA (Caldwell & Pate, 2013; Hughes & Trimble, 1965; Mullins, 1963) where there is often an increase in NOTA selection when it is a distractor (Frary, 1991; Garcia-Perez, 1993). This may manifest as an irrelevant construct in the form of a NOTA selection tendency. Whereas previous NOTA research focuses on NOTA as it influences item statistics, the current research extends the focus on not only the influence of NOTA on item statistics, but also on NOTA as it may negatively impact the interpretation of test scores through construct irrelevant variance. To explore the influence of NOTA on examinee item responses, I sought to answer two research questions:

1. Is there a difference in item difficulty for examinees who receive a test with NOTA compared to examinees who receive a test without NOTA, after controlling for ability?
2. Do examinees exhibit a tendency to select NOTA when present as a distractor in multiple-choice items?

Research Question 1

With regards to the first research question, model comparisons across Model 1a, Model 1b, and Model 1c indicated little support for the inclusion of DIF parameters. Specifically, the AIC and BIC associated with Model 1a were smaller than those for Model 1c, the LRT for Model 1a and Model 1c was not statistically significantly different

than 0, and the variation in examinee abilities did not change when including DIF parameters. While model comparisons did not support the inclusion of the DIF parameters, three of the 10 NOTA items were flagged for DIF (i.e., DIF parameter estimates statistically significantly different than 0). Item 6 was easier for NOTA group examinees to correctly answer, after controlling for information literacy ability differences between the NOTA and non-NOTA groups. In contrast, item 1 and item 11 were harder for NOTA group examinees to correctly answer, after controlling for information literacy ability differences between the NOTA and non-NOTA groups.

The three items flagged for DIF also had meaningful differences in item difficulty between the NOTA and non-NOTA groups after controlling for impact. This alignment was expected as, under the Rasch model, total score is a sufficient statistic for examinee ability (Anderson, 1977). For item 6, the NOTA group more frequently selected the correct option compared to the non-NOTA group. For item 1 and item 11, the NOTA group less frequently selected the correct option compared to the non-NOTA group. Further, the three flagged DIF items also exhibited meaningful differences in fourth option selection frequency between the NOTA group and non-NOTA groups. For item 6, NOTA selection was less frequent for the NOTA group than fourth option selection for the non-NOTA group. This pattern suggests that NOTA group examinees were more drawn to select the correct answer rather than the NOTA option as compared to the non-NOTA group. In other words, the fourth option was more attractive as a possible correct answer than the NOTA option. For item 1 and item 11, NOTA selection was greater than fourth option selection. Additionally, NOTA option discrimination was closer to 0 than fourth option discrimination, indicating a weaker relationship between NOTA option

selection and total score. This pattern of results suggests that NOTA group examinees were drawn to select NOTA over the correct option. In other words, the NOTA option was more attractive as a possible correct answer than the fourth option.

Although identified DIF does not attribute a cause of the differences, it is likely due to the presence of NOTA, as it is the largest difference between the randomly assigned groups. However, when paired with the model comparison indices, the most likely explanation for DIF is not simply due to the presence of NOTA. Rather, it is likely the interaction of examinees with NOTA. Previous NOTA work has alluded to a separate NOTA construct (Caldwell & Pate, 2013; Hughes & Trimble, 1965; Mullins, 1963). Butler (2018) describes that some examinees may be drawn to select NOTA, making items where NOTA is a distractor seemingly harder. Further, Garcia-Perez (1993) and Frary (1991) reported an increase in NOTA selection when it was a distractor, and both Caldwell and Pate (2013) and Butler (2018) explain this increased selection as a result of it just being present and not necessarily due to an examinee's ability. In this study, as exhibited with item 1 and item 11, there was an association between higher NOTA selection by NOTA examinees who have information literacy high total scores, as compared to a replaced distractor. This mimics the pattern described by Caldwell and Pate (2013), and it provides support to Butler's (2018) hypothesis that there is a selection tendency to select NOTA for some examinees. However, due to the model comparison results and inconsistencies in the direction of DIF, there is not strong evidence to suggest that NOTA was the source of DIF.

To answer this question, I examined whether item difficulty varies across NOTA items for examinees who completed a test with NOTA and a test without NOTA, after

controlling for impact. However, future studies should examine the influence of NOTA on item discrimination. Impacts of NOTA on item discrimination tend to vary in prior NOTA studies as discussed in meta-analyses conducted by Rodriguez (1997) and Knowles and Welch (1992). Specifically, NOTA has resulted in an increase in item discrimination in half of the studies examined but a decrease in item discrimination in the other half of the studies examined. Exploring whether item discrimination varies across NOTA items after controlling for ability in a multilevel framework may provide additional evidence of NOTA influences on item properties.

Research Question 2

To address my second research question, and to formally model a NOTA selection tendency, I used an item response tree (IRTree). There was supportive validity evidence to support NST and TOI interpretations under the model. However, there was a lack of evidence to support a selection tendency toward NOTA based on the model-data fit comparisons of the Non-NOTA Selection Tendency Model (i.e., Model 2a) and the NOTA Selection Tendency Model (i.e., Model 2b). The lack of support for Model 2b over Model 2a may be explained by low rates of examinees selecting NOTA. The majority of examinees never selected NOTA, and of those who did, few did so more than three times. In other words, there may simply not be a NOTA selection tendency trait. If there is, it could be rare in the population.

Additionally, while NOTA option selection by the NOTA group was greater than fourth option selection by the non-NOTA group, there was no sufficient evidence to support the NOTA Selection Tendency Model. There was evidence to support the validity of the θ_{iNST} (i.e., positive relationship with total NOTA selection, positive

relationship with examinees reported attitudes of being drawn toward NOTA, and decreased correlation between the information literacy TOI from Model 2a to Model 2b), but based on the DIC, the NOTA Selection Tendency Model did not fit better than the Non-NOTA Selection Tendency Model. Researchers should consider using other measures to evaluate model-data fit, such as posterior predictive model checks, as the DIC makes certain assumptions about the posterior distribution. Stone and Zhu (2015) describe that the DIC may not be a good estimate of model-data fit in situations where the posterior distribution is not normally distributed.

The NOTA Selection Tendency Model I developed was inspired by previous IRTree work, such as Böckenholt's (2012a) Cognitive-Miser Response (CMR) Model and Deng and Bolt's (2016) Sequential Response Model for Multiple-Choice Items (SRM-MC). It is important to recognize that under the NST Model, I assume that an examinee uses a sequential process when answering a multiple-choice item. Specifically, I assume that an examinee's decision to select NOTA is independent of their decision to select the correct or incorrect options. Rather, an examinee may simultaneously evaluate all response options rather than independently decide to select NOTA. Even if examinees use a simultaneous process to answer multiple-choice items when NOTA is present, it may be possible that a secondary NST is influencing response selection. Therefore, in the future, researchers should compare model-data fit of the current NST Model with a model that assumes a simultaneous response process. For example, Deng and Bolt (2016) compare the SRM-MC to a two-dimensional nominal response model, which assumes a simultaneous evaluation of the response options and involves estimating a separate, latent trait.

Aside from the number of examinees who selected NOTA per item, the attitudes toward NOTA responses provide another potential reason for the lack of support for the NST model. There were significant differences in attitudes toward NOTA between the NOTA and non-NOTA groups, which, by chance, may have influenced the results. The NOTA group reported lower average endorsement to being drawn to select NOTA compared to the non-NOTA group. Additionally, the NOTA group reported higher average endorsement to avoiding NOTA compared to the non-NOTA group. While the groups did not differ in average information literacy ability (i.e., the impact estimate from Model 1b was not statistically significant), there is evidence to suggest that the groups differed in attitudes toward NOTA. It is important to note that examinees completed the information literacy test prior to completing the NOTA attitude items. Examinees' responses to the attitudinal NOTA items may have influenced completing the information literacy test with NOTA. Additionally, the results may have been different with regards to the NOTA Selection Tendency Model had the non-NOTA group completed a test with NOTA as their attitudes toward NOTA were different than the NOTA group.

Limitations and Future Directions

The number of items that contained NOTA and the procedures used to replace the fourth option with NOTA likely played a role in the results of both research questions. Examinees in the NOTA group completed a test which included 10 NOTA items. If more items included NOTA, then examinees might have noticed NOTA more, which could result in a greater influence on item responses. Additionally, because NOTA was replaced with an implausible distractor, the presence of NOTA may not have been as influential had it been if NOTA was replaced with a well-functioning distractor. In other

words, the presence of NOTA and the feasibility of other distractors being correct are likely impacting results.

For future studies, it would be useful to collect response process data from examinees as they answer multiple-choice items with NOTA as a distractor. Researchers could use think aloud protocols to ask examinees to describe their response processes as they answer items with and without NOTA (Leighton & Lehman, 2020). These responses could be compared to see whether there are differences in the amount of thinking and effort when answering an item with and without NOTA. Not only is it imperative to collect and evaluate validity evidence for item-writing guidelines, but test developers are often tasked with describing examinees' proposed response processes when developing items (AERA et al., 2014; Embretson, 2016). For example, test developers should state the depth of knowledge (e.g., recall, recognition, etc.) associated with specific items (Embretson, 2016). To support use of an item-writing guideline when developing multiple-choice items, test developers are required to cite evidence for its use (Kane, 2016; Rodriguez, 2016).

Similar to the mixed effects NOTA has on item statistics in prior research (e.g., Knowles & Welch, 2001; Rodriguez, 1991), in this study, NOTA had mixed effects on examinee item responses. While some NOTA items were flagged for DIF, and thus a separate NOTA construct may impact examinee item responses, not all NOTA items were flagged for DIF. Additionally, based on results of the current study, modeling a NOTA selection tendency has mixed support. Instead, perhaps the presence of NOTA may just increase cognitive complexity of the items (e.g., higher item difficulty). This presence of an increase in cognitive complexity may be due to how other options function

or features of the item (e.g., stem word length, stem readability, etc.). For example, there could be an interaction between stem word length and presence of NOTA. A longer stem may increase cognitive processing by the examinee, and if NOTA is present as an option, that may also increase cognitive processing, which causes the examinee to give up and guess. This situation could be troublesome as construct irrelevant variance is introduced through increased cognitive processing.

Given the results of the current study, there is inconclusive evidence that the presence of NOTA as a distractor in multiple-choice items results in a nuisance construct. While results from the first research question suggest that NOTA may influence examinee item responses simply due to its presence, results from the second research question suggest that there is not a selection tendency toward NOTA. The mixed results of the current study align with previous NOTA meta-analyses (e.g., Knowles & Welch, 2001; Rodriguez, 1991), but the current study differs due to the focus on trying to explain construct irrelevant variance with NOTA. In the future, researchers should shift their focus from the influence of NOTA on item statistics to the influence of NOTA on examinee processing of multiple-choice items to provide a more comprehensive argument for or against the use of NOTA in multiple-choice items.

Tables

Table 1

Item and distractor analysis results for information literacy test

	Options							
	A		B		C		D	
Item	*p*	*PB*	*p*	*PB*	*p*	*PB*	*p*	*PB*
1*	.83	.32	.10	-.25	.06	-.16	**.01**	**-.09**
2	.22	-.04	.33	.06	.20	-.07	.25	.03
3*	.12	-.12	.07	-.20	.81	.29	**.01**	**-.22**
4	.48	.06	.08	-.25	.41	.09	.03	-.04
5	.02	-.20	.85	.41	.09	-.32	.03	-.11
6*	.76	.20	.14	-.08	**.03**	**-.27**	.07	-.04
7	.76	.32	.06	-.16	.15	-.16	.02	-.27
8	.47	.08	.09	-.20	.42	.09	.02	-.19
9	.03	-.20	.42	.12	.53	-.01	.02	-.14
10*	.04	-.31	.23	-.08	.69	.25	**.03**	**-.09**
11*	**.03**	**-.19**	.62	.24	.03	-.36	.32	-.04
12*	**.02**	**-.16**	.27	-.12	.56	.19	.15	-.05
13	.04	-.41	.94	.50	.02	-.24	.01	-.16
14*	**.01**	**-.21**	.87	.53	.05	-.37	.06	-.28
15*	.84	.33	.06	-.21	.07	-.18	**.02**	**-.17**
16	.03	-.18	.05	-.23	.90	.44	.02	-.34
17	.20	-.10	.73	.30	.06	-.31	.01	-.20
18	.36	.12	.15	-.12	.21	-.11	.28	.07
19	.04	-.14	.04	-.21	.88	.41	.04	-.35
20	.04	-.38	.79	.33	.10	-.17	.07	-.05
21	.14	-.10	.04	-.21	.08	-.28	.74	.34
22	.04	-.10	.83	.36	.09	-.24	.03	-.25
23*	.07	-.11	.78	.34	.12	-.21	**.02**	**-.31**
24	.08	-.18	.87	.37	.02	-.30	.02	-.19
25	.24	-.01	.63	.26	.05	-.36	.08	-.36
26	.02	-.23	.30	-.06	.63	.22	.05	-.23
27	.61	.12	.15	-.11	.19	-.04	.05	-.01
28	.40	.23	.05	-.16	.52	-.01	.03	-.39
29	.14	-.09	.05	-.22	.32	-.08	.48	.24
30*	.85	.45	.07	-.22	.04	-.33	**.03**	**-.21**

Note. *p* = item and option difficulty; *PB* = point-biserial (item and option discrimination). Implausible distractor item statistics in bold. * = item with NOTA distractor replacement.

Table 2

Average attitudes toward NOTA scores by NOTA and non-NOTA groups

	Drawn		Avoid		Indifferent	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*
Overall	2.80	1.11	3.07	0.99	3.00	0.93
Non-NOTA	3.08	1.10	2.88	0.97	2.99	0.93
NOTA	2.56	1.10	3.21	1.02	3.01	0.93

Table 3

Model Parameter Estimates of Models 1a – Models 1c

	Model 1a			Model 1b			Model 1c			
							Non-NOTA Group			*NOTA Group*
Parameter	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate*
Fixed Effects										
Item Difficulty										
$\gamma_{1,0}$	-1.23	0.13	**-9.30**	-1.18	0.14	**-8.53**	-1.49	0.19	**-7.76**	-1.14
$\gamma_{2,0}$	0.94	0.12	**7.55**	0.98	0.13	**7.48**	0.94	0.12	**7.55**	0.94
$\gamma_{3,0}$	-1.48	0.14	**-10.61**	-1.44	0.15	**-9.85**	-1.49	0.19	**-7.76**	-1.38
$\gamma_{4,0}$	0.62	0.12	**5.17**	0.66	0.13	**5.23**	0.62	0.12	**5.17**	0.62
$\gamma_{5,0}$	-2.08	0.17	**-12.60**	-2.04	0.17	**-11.95**	-2.08	0.17	**-12.60**	-2.08
$\gamma_{6,0}$	-1.40	0.14	**-10.18**	-1.35	0.14	**-9.42**	-1.08	0.18	**-6.13**	-1.71
$\gamma_{7,0}$	-1.66	0.15	**-11.33**	-1.61	0.15	**-10.59**	-1.65	0.15	**-11.33**	-1.65
$\gamma_{8,0}$	0.48	0.12	**4.05**	0.52	0.13	**4.18**	0.48	0.12	**4.06**	0.48
$\gamma_{9,0}$	0.03	0.12	0.23	0.07	0.12	0.59	0.03	0.12	0.24	0.03
$\gamma_{10,0}$	-1.28	0.13	**-9.57**	-1.23	0.14	**-8.80**	-1.29	0.18	**-7.04**	-1.17
$\gamma_{11,0}$	-0.24	0.12	-2.05	-0.19	0.12	-1.56	-0.24	0.12	-2.05	0.11
$\gamma_{12,0}$	0.20	0.12	1.69	0.24	0.12	1.95	0.20	0.12	1.69	0.30
$\gamma_{13,0}$	-3.41	0.27	**-12.57**	-3.36	0.27	**-12.26**	-3.41	0.37	**-9.20**	-3.41
$\gamma_{14,0}$	-1.82	0.15	**-11.92**	-1.78	0.16	**-11.20**	-1.83	0.21	**-8.73**	-1.71
$\gamma_{15,0}$	-1.78	0.15	**-11.78**	-1.73	0.16	**-11.05**	-1.60	0.20	**-8.11**	-1.91
$\gamma_{16,0}$	-2.33	0.18	**-13.01**	-2.29	0.18	**-12.42**	-2.33	0.18	**-13.01**	-2.33
$\gamma_{17,0}$	-1.18	0.13	**-9.02**	-1.14	0.14	**-8.26**	-1.18	0.13	**-9.02**	-1.18
$\gamma_{18,0}$	0.64	0.12	**5.37**	0.69	0.13	**5.42**	0.64	0.12	**5.38**	0.64
$\gamma_{19,0}$	-2.36	0.18	**-13.05**	-2.32	0.19	**-12.47**	-2.36	0.18	**-13.04**	-2.36
$\gamma_{20,0}$	-1.56	0.14	**-10.94**	-1.51	0.15	**-10.18**	-1.56	0.14	**-10.93**	-1.56
$\gamma_{21,0}$	-1.45	0.14	**-10.44**	-1.40	0.15	**-9.67**	-1.45	0.14	**-10.44**	-1.45

Table 3 (cont.)										
$\gamma_{22,0}$	-1.96	0.16	**-12.31**	-1.91	0.16	**-11.63**	-1.96	0.16	**-12.31**	-1.96
$\gamma_{23,0}$	-1.60	0.14	**-11.10**	-1.55	0.15	**-10.35**	-1.39	0.19	**-7.41**	-1.76
$\gamma_{24,0}$	-1.72	0.15	**-11.56**	-1.67	0.15	**-10.82**	-1.72	0.15	**-11.55**	-1.72
$\gamma_{25,0}$	-0.89	0.12	**-7.12**	-0.84	0.13	**-6.39**	-0.88	0.12	**-7.12**	-0.88
$\gamma_{26,0}$	-0.39	0.12	**-3.29**	-0.34	0.13	**-2.73**	-0.39	0.12	**-3.29**	-0.39
$\gamma_{27,0}$	1.69	0.15	**11.57**	1.74	0.15	**11.41**	1.69	0.15	**11.57**	1.69
$\gamma_{28,0}$	0.76	0.12	**6.27**	0.81	0.13	**6.27**	0.76	0.12	**6.27**	0.76
$\gamma_{29,0}$	0.01	0.12	0.13	0.06	0.12	0.49	0.02	0.12	0.13	0.02
$\gamma_{30,0}$	-2.16	0.17	**-12.76**	-2.12	0.17	**-12.12**	-1.96	0.22	**-9.01**	-2.32
Impact										
$\gamma_{01}{}^{*}$	-	-	-	0.09	0.09	1.07	-	-	-	-
DIF										
$\gamma_{1,1}$	-	-	-	-	-	-	0.35	0.13	**2.71**	-
$\gamma_{3,1}$	-	-	-	-	-	-	0.11	0.27	0.39	-
$\gamma_{6,1}$	-	-	-	-	-	-	-0.64	0.28	**-2.30**	-
$\gamma_{10,1}$	-	-	-	-	-	-	0.12	0.26	0.46	-
$\gamma_{11,1}$	-	-	-	-	-	-	0.35	0.13	**2.71**	-
$\gamma_{13,1}$	-	-	-	-	-	-	0.10	0.54	0.19	-
$\gamma_{14,1}$	-	-	-	-	-	-	0.12	0.30	0.39	-
$\gamma_{15,1}$	-	-	-	-	-	-	-0.31	0.30	-1.04	-
$\gamma_{23.1}$	-	-	-	-	-	-	-0.37	0.29	-1.30	-
$\gamma_{30,1}$	-	-	-	-	-	-	-0.36	0.34	-1.05	-
Random Effects										
$var(u_{oj})$	0.51	-	-	0.50	-	-	0.51	-	-	-

Note. *NOTA group item difficulties were not estimated but calculated using the DIF estimates. Bolded values indicate p-values less than .05.

Table 4

Comparison of NOTA item difficulty estimates under Model 1a and Model 1c

	Model 1a		Model 1c	
NOTA Items	All	DIF	Non-NOTA group	NOTA group
1*	-1.23	0.35	-1.49	-1.14
3	-1.48	0.11	-1.49	-1.38
6*	-1.40	-0.64	-1.08	-1.71
10	-1.28	0.12	-1.29	-1.17
11*	-.24	0.35	-0.24	0.11
12	.20	0.10	0.20	0.30
14	-1.82	0.12	-1.83	-1.71
15	-1.78	-0.31	-1.60	-1.91
23	-1.60	-0.37	-1.39	-1.76
30	-2.16	-0.36	-1.96	-2.32

Note. Items with an asterisk indicate statistically significant DIF parameter estimates. DIF parameter estimates from Table 3 were added to the non-NOTA group item difficulty estimates to obtain the NOTA group item difficulty estimates.

Table 5

Model 1a, Model 1b, and Model 1c model-data fit information criteria and likelihood ratio test results

	Information Criteria		Likelihood Ratio Test			
Model	AIC	BIC	Deviance	DF	LRT	*p*
Model 1a	11287	11408	11225	337		
Model 1b	11288	11413	11224	336	1	.317
Model 1c	11289	11450	11207	327	18	.055

Note. $N = 368$. DF = degrees of freedom.

Table 6

Model 2a item parameters (# of iterations = 20000)

Discrimination					Difficulty				
Parameter	*EAP*	*SD*	95% HPD		Parameter	*EAP*	*SD*	95% HPD	
$\boldsymbol{a_1}$	0.78	0.24	0.33	1.25	$\boldsymbol{b_1}$	-1.36	0.48	-2.34	-0.61
a_2	0.46	0.16	0.17	0.78	b_2	2.18	0.79	0.86	3.75
$\boldsymbol{a_3}$	1.47	0.39	0.71	2.20	$\boldsymbol{b_3}$	-1.26	0.29	-1.83	-0.73
a_4	0.29	0.12	0.10	0.53	b_4	2.09	0.88	0.65	3.95
a_5	1.34	0.36	0.65	2.04	b_5	-1.73	0.41	-2.54	-1.06
$\boldsymbol{a_6}$	0.65	0.19	0.31	1.02	$\boldsymbol{b_6}$	-2.89	0.81	-4.54	-1.58
a_7	0.77	0.23	0.37	1.24	b_7	-2.59	0.75	-4.09	-1.37
a_8	0.26	0.12	0.06	0.51	b_8	1.38	0.84	-0.02	3.20
a_9	0.34	0.16	0.06	0.65	b_9	-0.13	0.62	-1.46	1.07
$\boldsymbol{a_{10}}$	0.86	0.26	0.37	1.36	$\boldsymbol{b_{10}}$	-1.63	0.51	-2.65	-0.81
$\boldsymbol{a_{11}}$	0.31	0.15	0.05	0.61	$\boldsymbol{b_{11}}$	-0.20	0.69	-1.61	1.25
$\boldsymbol{a_{12}}$	0.28	0.14	0.05	0.55	$\boldsymbol{b_{12}}$	0.89	0.77	-0.44	2.58
a_{13}	1.45	0.44	0.67	2.32	b_{13}	-2.93	0.66	-4.24	-1.80
$\boldsymbol{a_{14}}$	1.31	0.36	0.66	2.04	$\boldsymbol{b_{14}}$	-1.67	0.39	-2.48	-1.01
$\boldsymbol{a_{15}}$	1.22	0.33	0.60	1.86	$\boldsymbol{b_{15}}$	-1.95	0.47	-2.87	-1.15
a_{16}	1.50	0.42	0.75	2.34	b_{16}	-1.99	0.44	-2.87	-1.26
a_{17}	0.76	0.23	0.32	1.21	b_{17}	-1.66	0.56	-2.79	-0.77
a_{18}	0.32	0.12	0.11	0.57	b_{18}	2.10	0.86	0.67	3.81
a_{19}	3.47	2.18	1.40	5.88	b_{19}	-1.42	0.22	-1.87	-1.01
a_{20}	0.66	0.19	0.32	1.04	b_{20}	-2.98	0.83	-4.64	-1.56
a_{21}	1.41	0.38	0.71	2.15	b_{21}	-1.33	0.31	-1.94	-0.77
a_{22}	0.67	0.19	0.33	1.05	b_{22}	-3.10	0.84	-4.83	-1.72
$\boldsymbol{a_{23}}$	0.98	0.28	0.47	1.52	$\boldsymbol{b_{23}}$	-2.13	0.57	-3.26	-1.18
a_{24}	0.90	0.26	0.42	1.41	b_{24}	-2.01	0.56	-3.17	-1.09
a_{25}	0.63	0.21	0.26	1.04	b_{25}	-1.55	0.58	-2.68	-0.61
a_{26}	0.33	0.14	0.09	0.60	b_{26}	-1.41	0.76	-2.97	-0.12

Table 6 (cont.)									
a_{27}	0.46	0.14	0.22	0.74	b_{27}	3.21	0.90	1.65	5.00
a_{28}	0.56	0.20	0.19	0.95	b_{28}	1.21	0.55	0.37	2.34
a_{29}	0.87	0.27	0.37	1.40	b_{29}	0.10	0.23	-0.33	0.58
$\boldsymbol{a_{30}}$	2.26	0.67	1.09	3.60	$\boldsymbol{b_{30}}$	-1.63	0.29	-2.21	-1.10

Note. Bolded parameters indicate NOTA items.

Table 7

Model 2b item parameters (# of iterations = 20000)

Discrimination					Difficulty				
Parameter	*EAP*	*SD*	95% HPD		Parameter	*EAP*	*SD*	95% HPD	
$\boldsymbol{a_{1NST}}$	0.79	0.28	0.32	1.34	$\boldsymbol{b_{1NST}}$	2.46	0.79	1.17	4.01
$\boldsymbol{a_{1TOI}}$	1.20	0.39	0.52	1.96	$\boldsymbol{b_{1TOI}}$	-1.80	0.54	-2.91	-0.98
a_{2TOI}	0.48	0.17	0.19	0.81	b_{2TOI}	2.15	0.77	0.88	3.69
$\boldsymbol{a_{3NST}}$	1.39	0.48	0.58	2.34	$\boldsymbol{b_{3NST}}$	3.69	0.91	2.09	5.48
$\boldsymbol{a_{3TOI}}$	1.43	0.40	0.69	2.23	$\boldsymbol{b_{3TOI}}$	-1.32	0.34	-2.00	-0.76
a_{4TOI}	0.30	0.12	0.10	0.54	b_{4TOI}	2.06	0.86	0.62	3.82
a_{5TOI}	1.36	0.36	0.70	2.09	b_{5TOI}	-1.69	0.40	-2.49	-1.03
$\boldsymbol{a_{6NST}}$	2.88	1.67	0.83	5.79	$\boldsymbol{b_{6NST}}$	2.92	0.64	1.90	4.25
$\boldsymbol{a_{6TOI}}$	0.60	0.17	0.29	0.94	$\boldsymbol{b_{6TOI}}$	-3.18	0.90	-5.00	-1.68
a_{7TOI}	0.81	0.24	0.38	1.29	b_{7TOI}	-2.45	0.69	-3.87	-1.31
a_{8TOI}	0.26	0.12	0.06	0.49	b_{8TOI}	1.41	0.84	-0.05	3.15
a_{9TOI}	0.37	0.17	0.07	0.71	b_{9TOI}	-0.08	0.56	-1.29	1.01
$\boldsymbol{a_{10NST}}$	4.95	4.14	1.04	11.37	$\boldsymbol{b_{10NST}}$	2.57	0.51	1.70	3.62
$\boldsymbol{a_{10TOI}}$	0.83	0.25	0.38	1.33	$\boldsymbol{b_{10TOI}}$	-1.70	0.55	-2.82	-0.82
$\boldsymbol{a_{11NST}}$	2.52	1.36	0.66	5.17	$\boldsymbol{b_{11NST}}$	2.12	0.52	1.32	3.17
$\boldsymbol{a_{11TOI}}$	0.28	0.14	0.04	0.55	$\boldsymbol{b_{11TOI}}$	-0.62	0.79	-2.27	0.90
$\boldsymbol{a_{12NST}}$	0.96	0.34	0.41	1.65	$\boldsymbol{b_{12NST}}$	2.70	0.80	1.39	4.31
$\boldsymbol{a_{12TOI}}$	0.28	0.15	0.05	0.57	$\boldsymbol{b_{12TOI}}$	0.11	0.75	-1.49	1.65
a_{13TOI}	1.49	0.46	0.69	2.38	b_{13TOI}	-2.84	0.64	-4.14	-1.74
$\boldsymbol{a_{14NST}}$	1.92	0.85	0.65	3.49	$\boldsymbol{b_{14NST}}$	2.80	0.71	1.68	4.24
$\boldsymbol{a_{14TOI}}$	1.20	0.34	0.58	1.90	$\boldsymbol{b_{14TOI}}$	-1.93	0.50	-2.95	-1.12
$\boldsymbol{a_{15NST}}$	4.46	3.01	0.80	10.21	$\boldsymbol{b_{15NST}}$	2.39	0.50	1.62	3.41
$\boldsymbol{a_{15TOI}}$	1.13	0.33	0.53	1.78	$\boldsymbol{b_{15TOI}}$	-2.15	0.57	-3.32	-1.24
a_{16TOI}	1.56	0.44	0.78	2.45	b_{16TOI}	-1.92	0.43	-2.78	-1.21
a_{17TOI}	0.75	0.23	0.33	1.21	b_{17TOI}	-1.66	0.58	-2.81	-0.75
a_{18TOI}	0.32	0.12	0.11	0.56	b_{18TOI}	2.14	0.86	0.74	3.94

Table 7 (cont.)									
a_{19TOI}	3.56	1.47	1.47	6.41	b_{19TOI}	-1.37	0.23	-1.82	-0.98
a_{20TOI}	0.66	0.19	0.33	1.06	b_{20TOI}	-2.95	0.82	-4.59	-1.55
a_{21TOI}	1.41	0.38	0.72	2.17	b_{21TOI}	-1.31	0.31	-1.94	-0.76
a_{22TOI}	0.66	0.19	0.32	1.03	b_{22TOI}	-3.13	0.85	-4.86	-1.66
$\boldsymbol{a_{23NST}}$	1.48	0.55	0.62	2.59	$\boldsymbol{b_{23NST}}$	3.05	0.78	1.78	4.61
$\boldsymbol{a_{23TOI}}$	1.06	0.31	0.50	1.69	$\boldsymbol{b_{23TOI}}$	-2.24	0.62	-3.48	-1.24
a_{24TOI}	0.89	0.26	0.42	1.41	b_{24TOI}	-2.00	0.58	-3.22	-1.10
a_{25TOI}	0.65	0.21	0.27	1.08	b_{25TOI}	-1.49	0.57	-2.60	-0.57
a_{26TOI}	0.32	0.14	0.09	0.60	b_{26TOI}	-1.43	0.80	-3.15	-0.15
a_{27TOI}	0.46	0.14	0.22	0.74	b_{27TOI}	3.25	0.92	1.63	5.06
a_{28TOI}	0.56	0.21	0.19	0.96	b_{28TOI}	1.23	0.55	0.40	2.39
a_{29TOI}	0.89	0.27	0.36	1.42	b_{29TOI}	0.12	0.22	-0.30	0.57
$\boldsymbol{a_{30NST}}$	2.32	1.20	0.73	4.49	$\boldsymbol{b_{30NST}}$	2.91	0.70	1.73	4.28
$\boldsymbol{a_{30TOI}}$	2.45	0.82	1.06	4.04	$\boldsymbol{b_{30TOI}}$	-1.64	0.32	-2.29	-1.08

Note. NOTA items are bolded. a_{jTOI} and b_{jTOI} are stage-level trait of interest estimates. a_{jNST} and b_{jNST} are stage-level NOTA selection tendency estimates.

Table 8

Model-data fit indices for Model 2a and Model 2b

Model	$\overline{D}$	$\widehat{D}$	p_D	DIC
Model 2a	5036.141	4879.446	156.694	5192.835
Model 2b	5305.157	5121.954	183.203	5488.360

Figures

Figure 1

Example of a multiple-choice item that violates an item-writing guideline

Scholarly sources are written for an A: expert. B: general audience. C: college student. D: middle school student.

Figure 2

Perceptions of None-of-the-Above survey

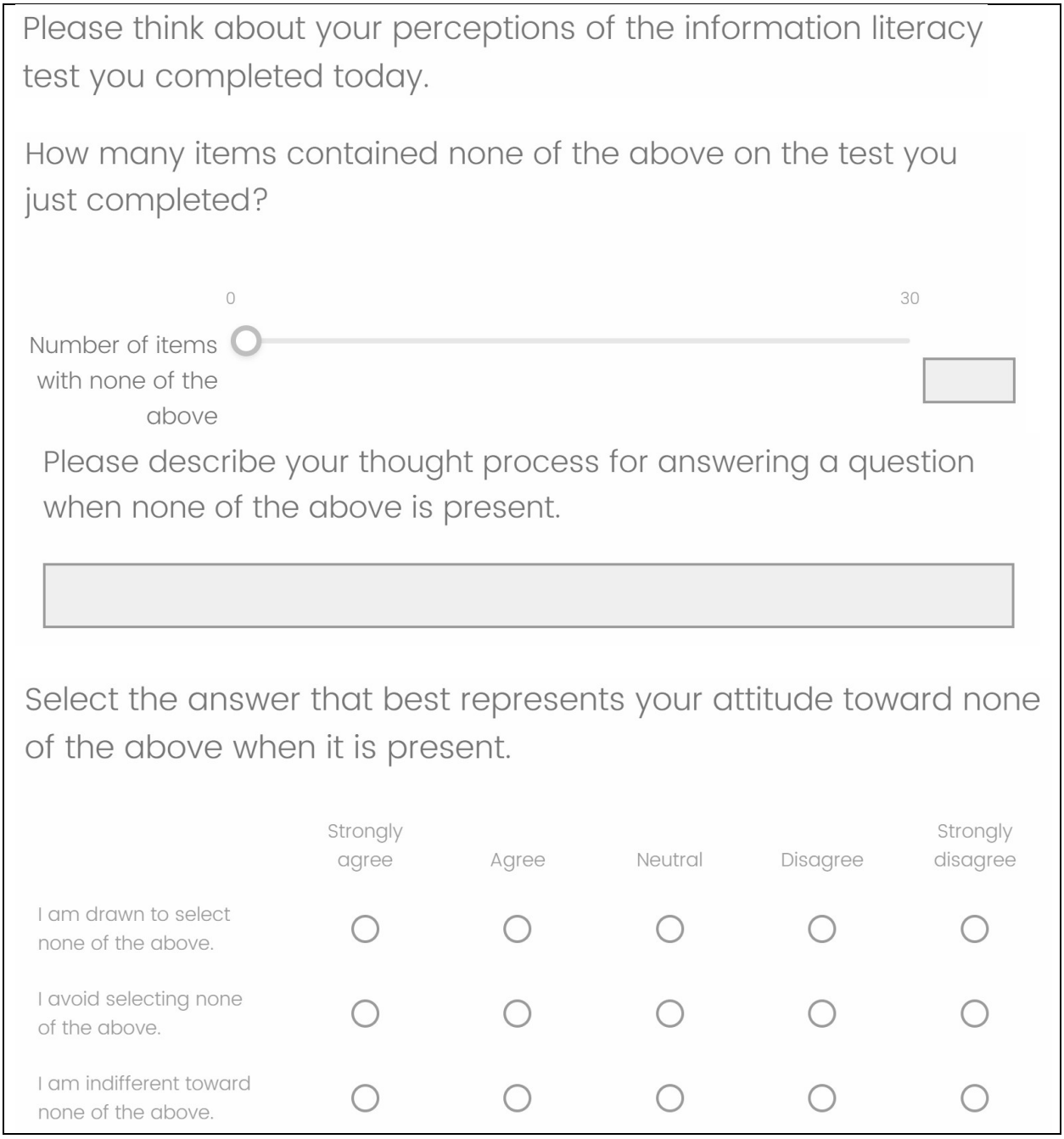


Figure 3

Selection tendency toward NOTA IRTree model

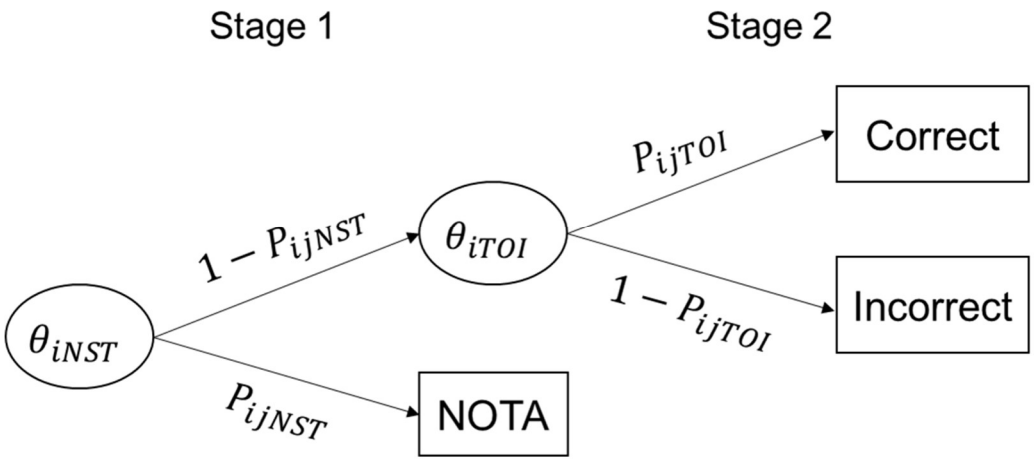
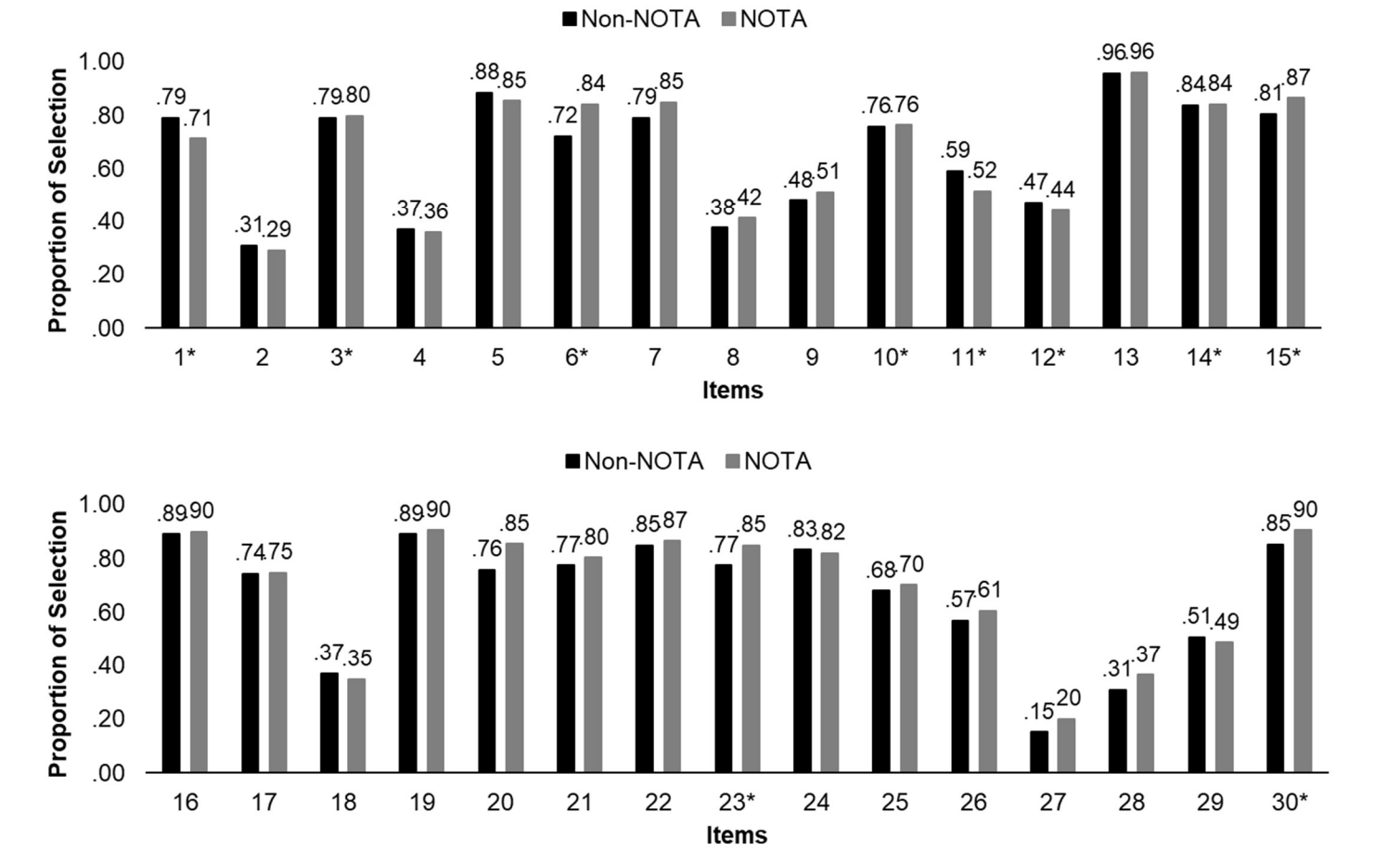


Figure 4

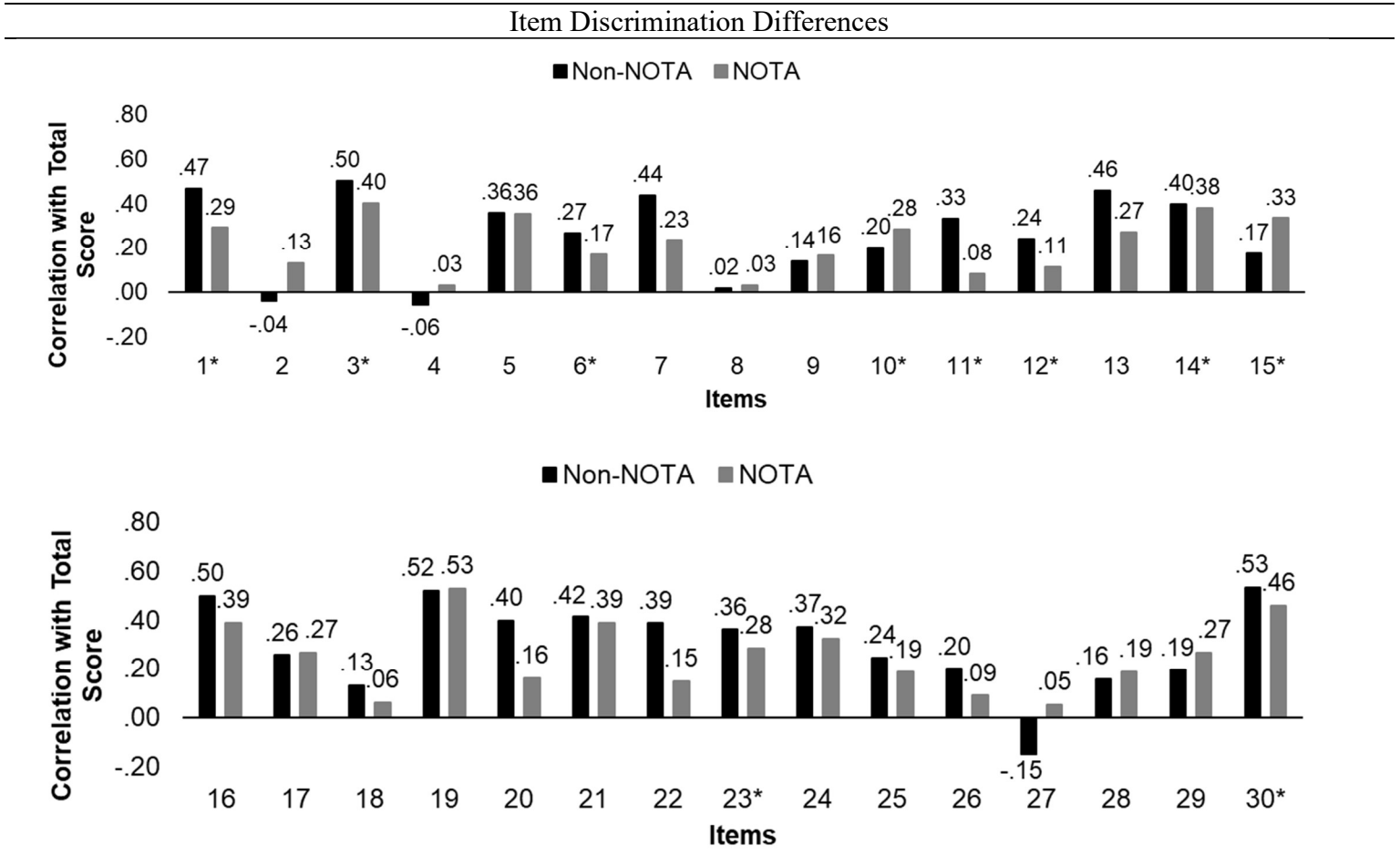
Item difficulty statistics by NOTA and non-NOTA groups



Note. Items with an * indicate items that contained NOTA for the NOTA group.

Figure 5

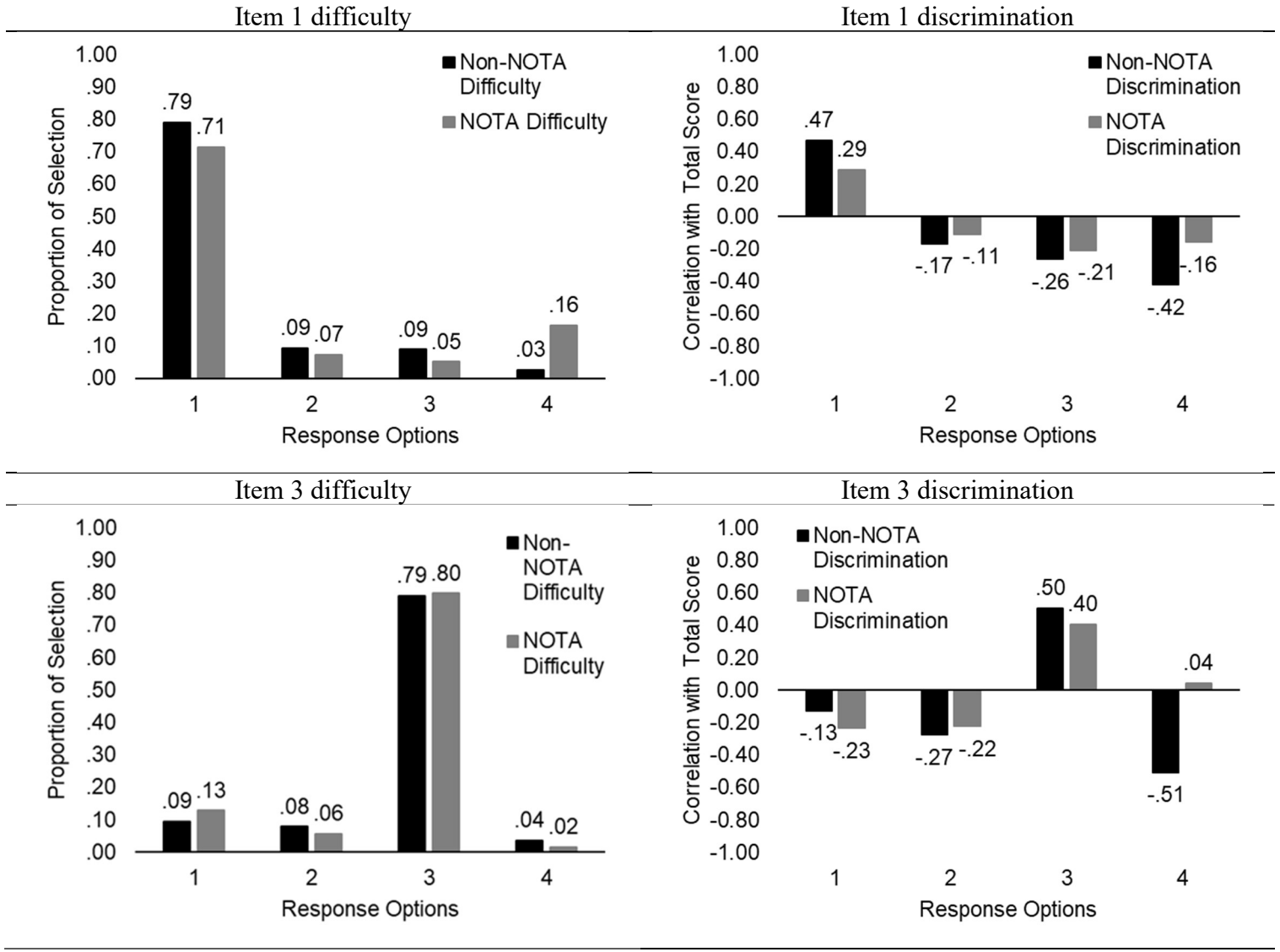
Item discrimination statistics by NOTA and non-NOTA groups



Note. Items with an * indicate items that contained NOTA for the NOTA group.

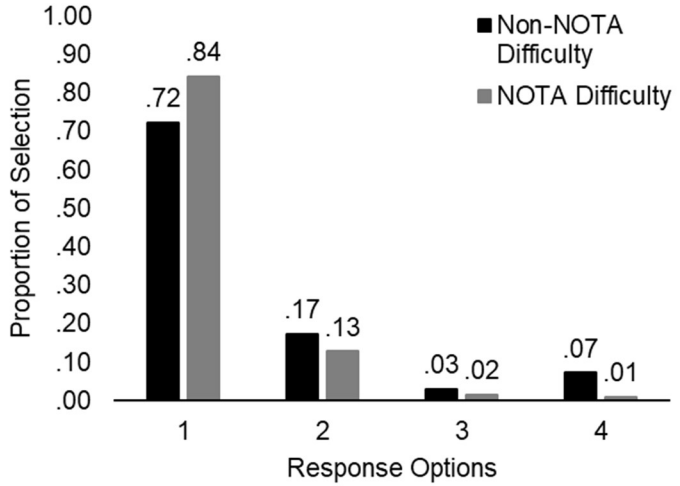
Figure 6

NOTA item differences in option difficulty and option discrimination between NOTA and non-NOTA groups

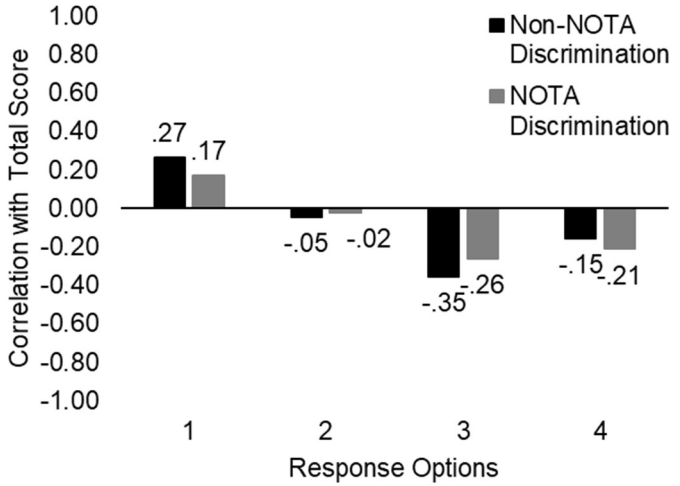


(Figure 6 cont.)

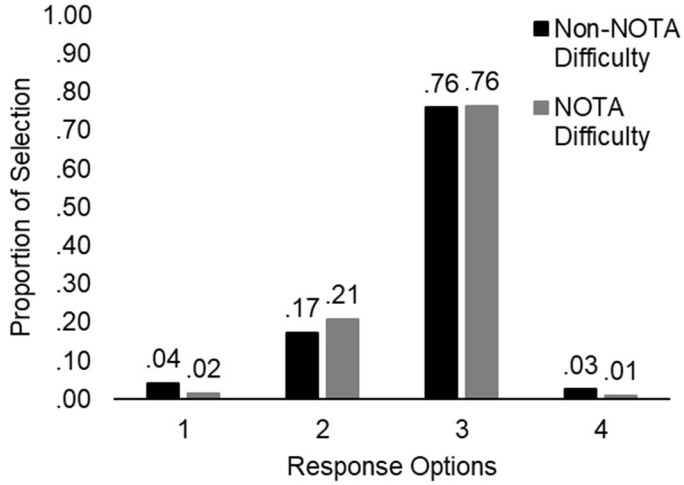
Item 6 difficulty



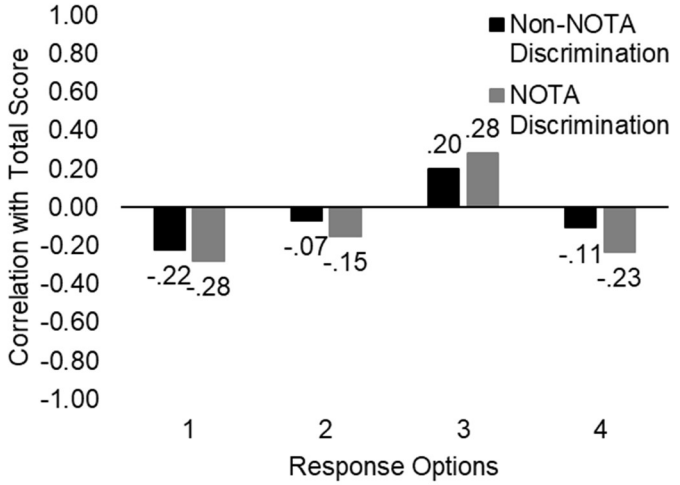
Item 6 discrimination



Item 10 difficulty

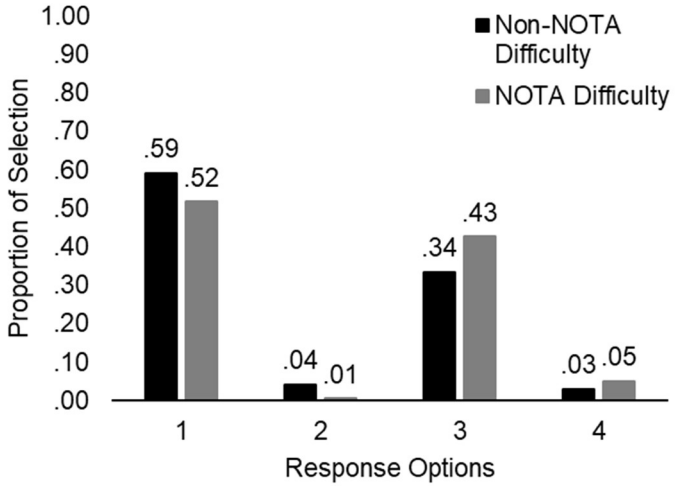


Item 10 discrimination

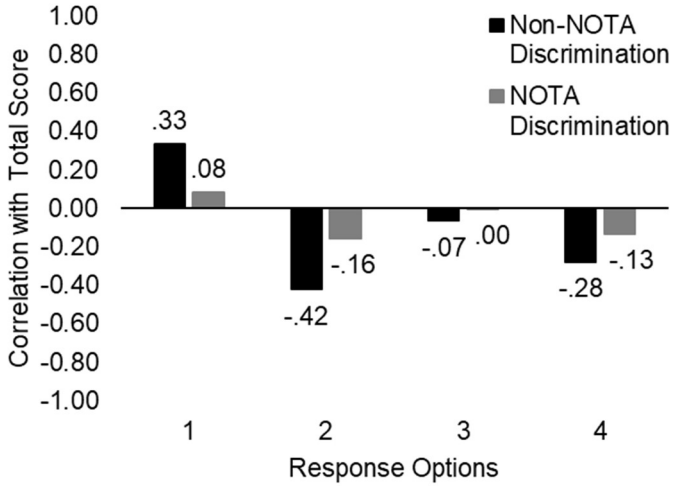


(Figure 6 cont.)

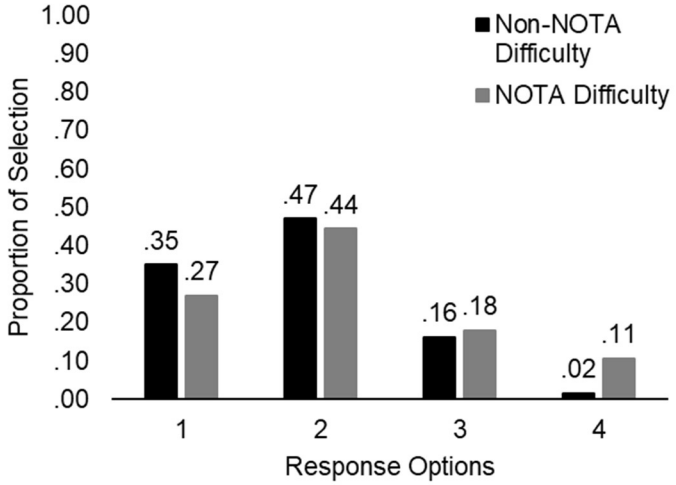
Item 11 difficulty



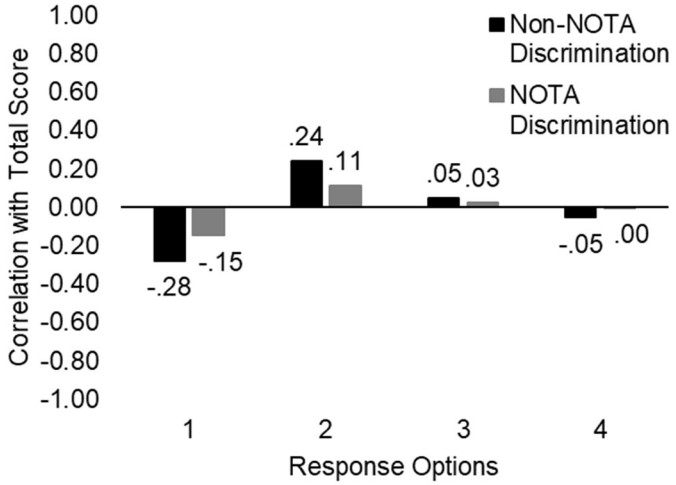
Item 11 discrimination



Item 12 difficulty

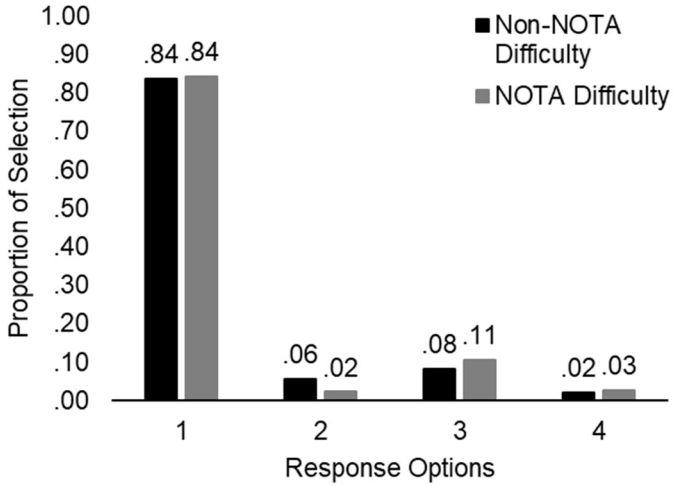


Item 12 discrimination

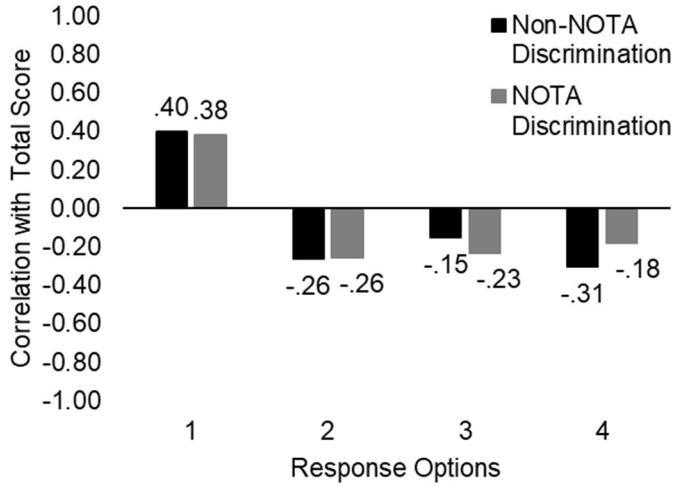


(Figure 6 cont.)

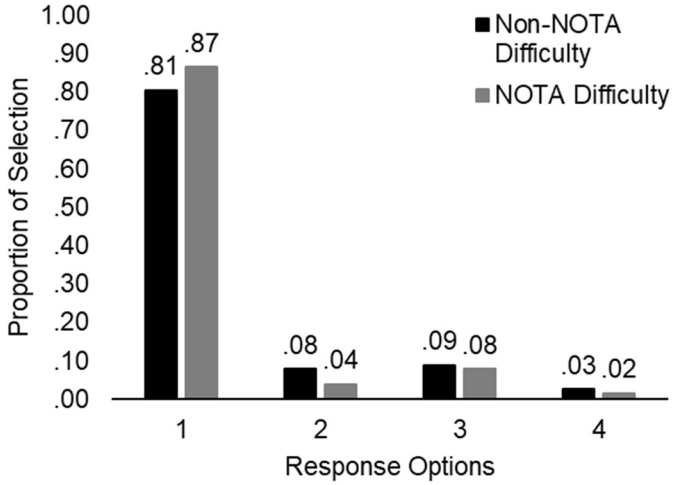
Item 14 difficulty



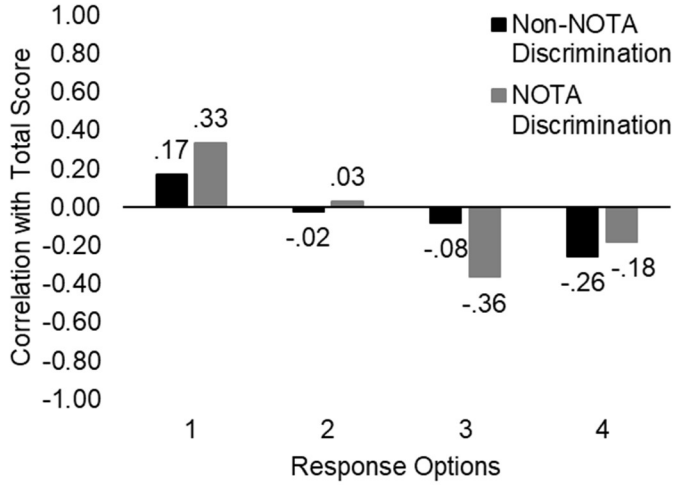
Item 14 discrimination



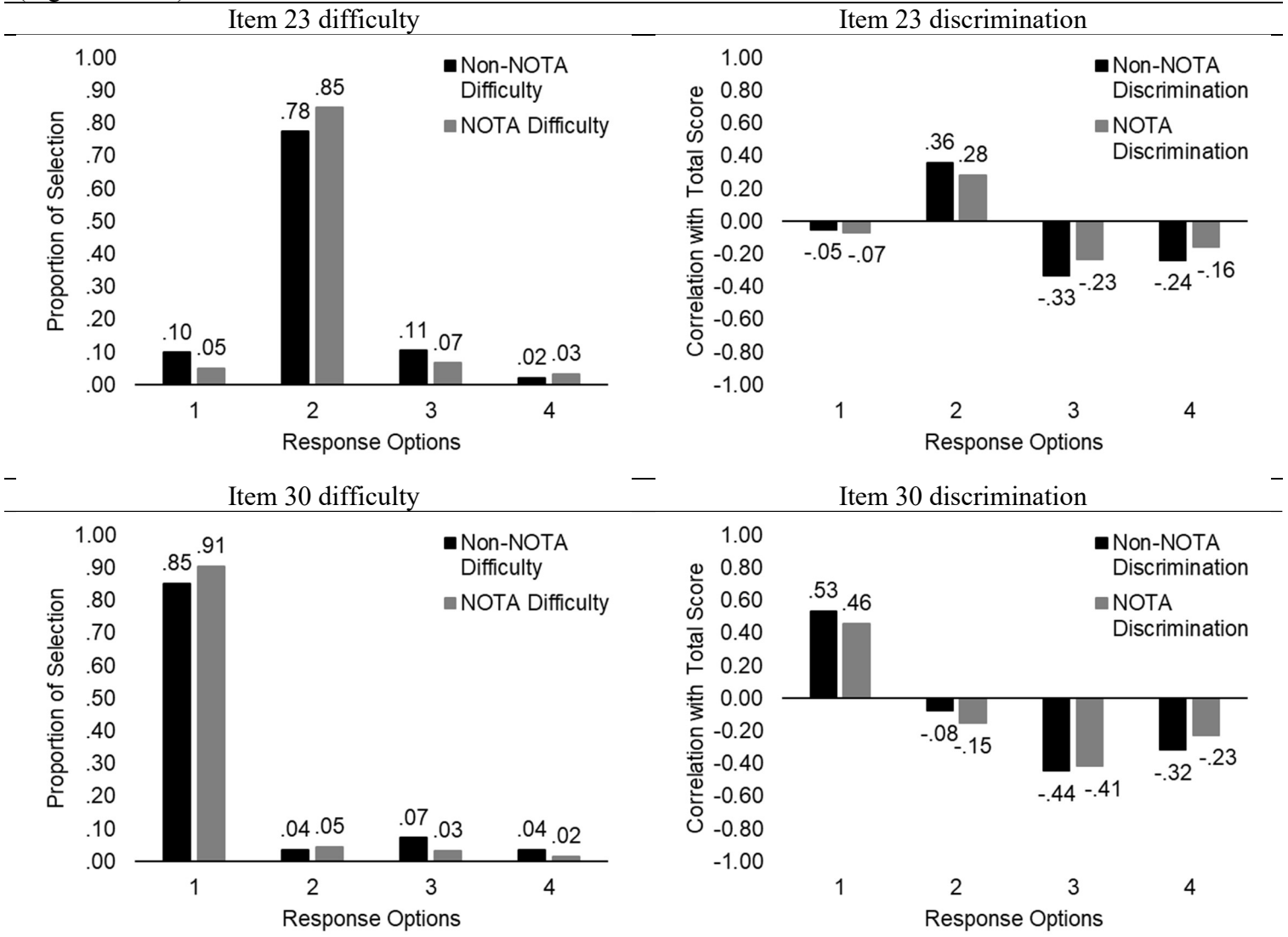
Item 15 difficulty



Item 15 discrimination



(Figure 6 cont.)



Note. Option 4 is the NOTA option (NOTA group) and fourth option (non-NOTA group).

Figure 7

Count distribution of total NOTA selection

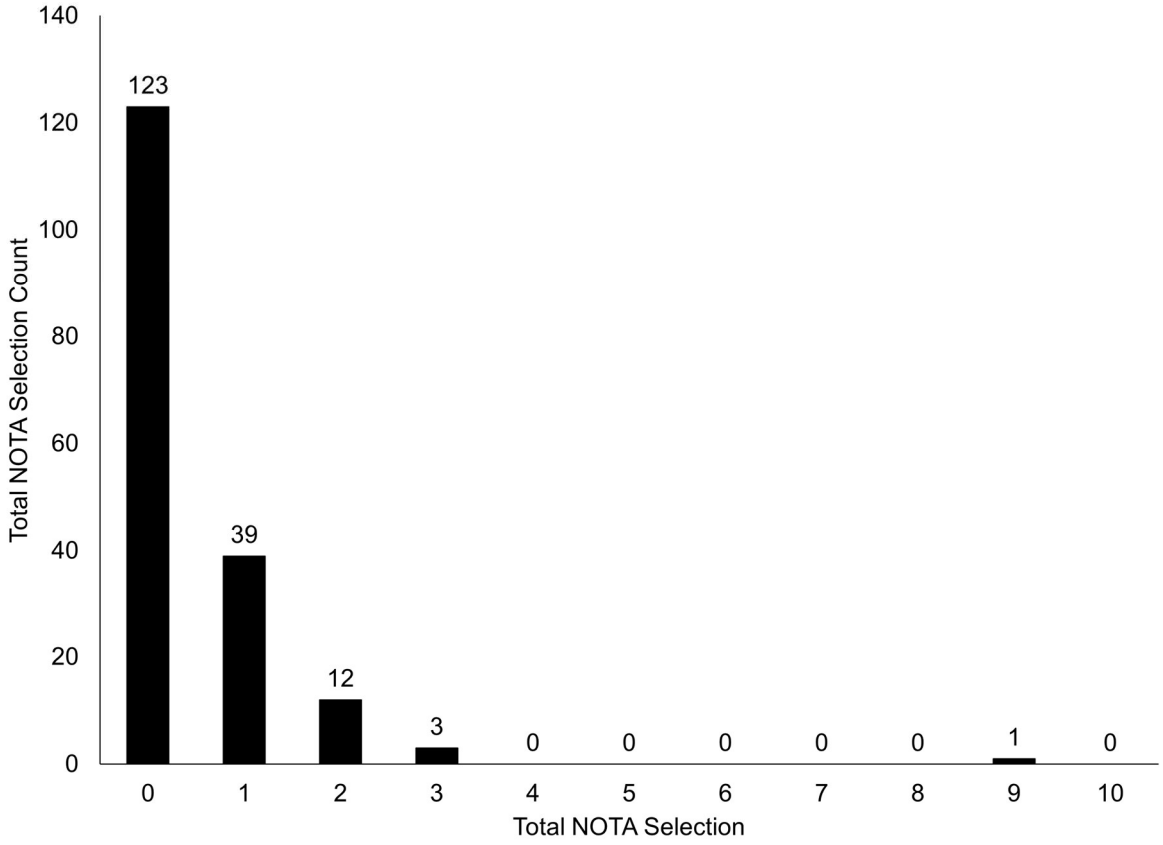


Figure 8

Perceived number of NOTA items by NOTA and non-NOTA examinees

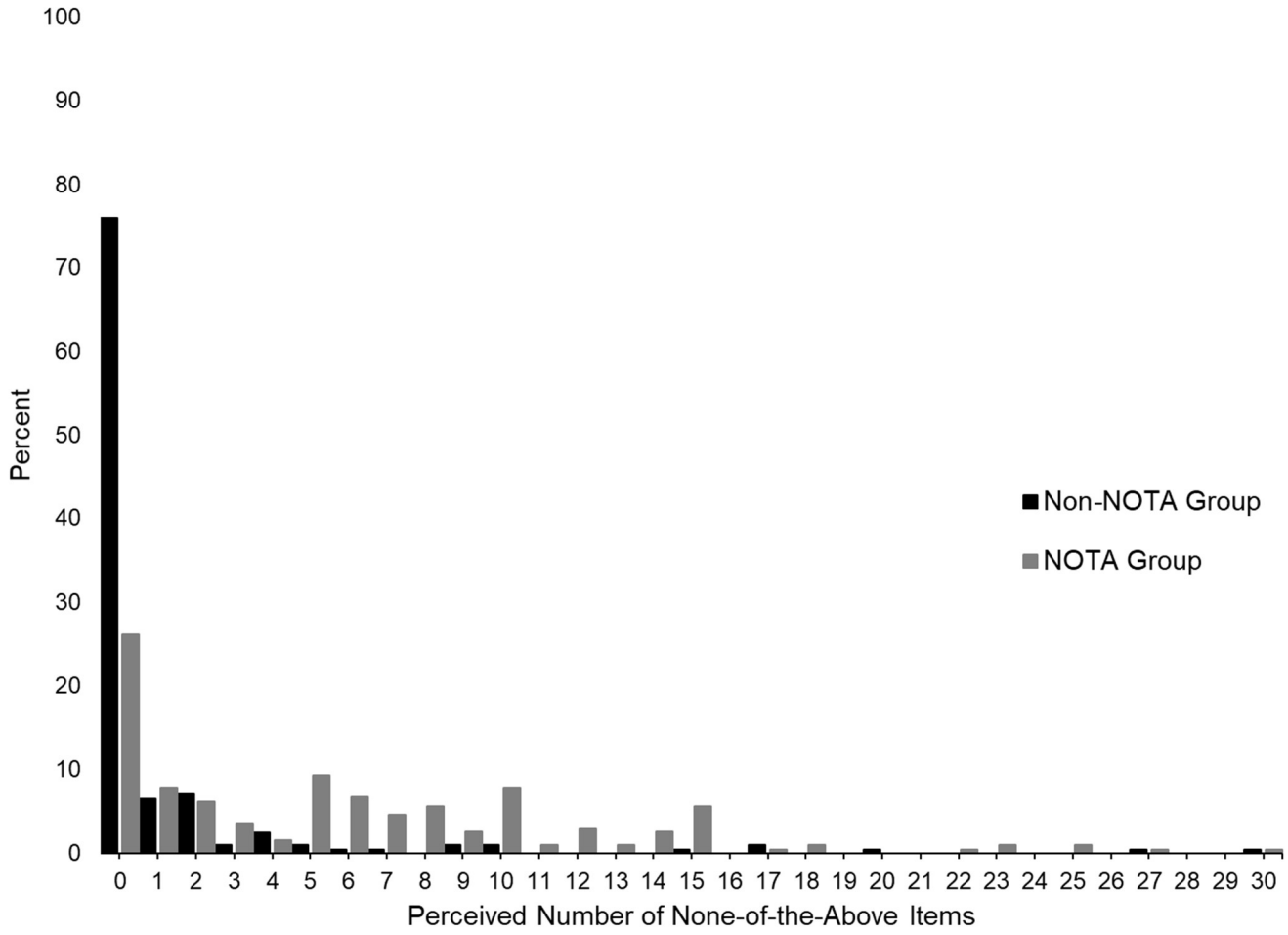


Figure 9

Comparison of item 1 response curves for non-NOTA and NOTA Groups

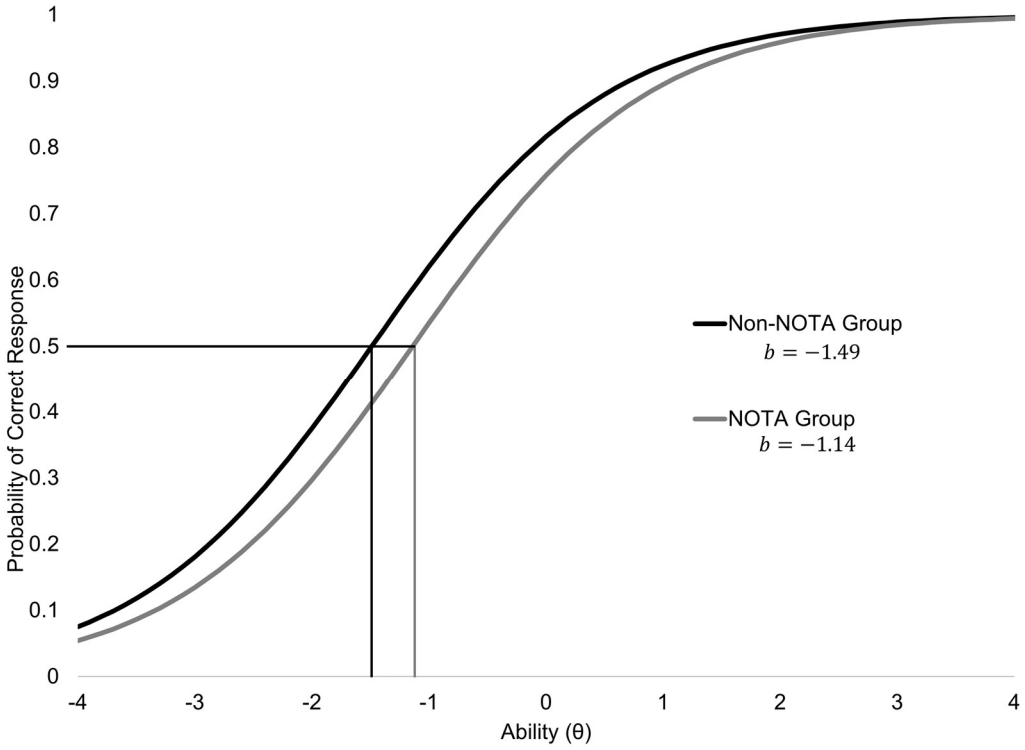


Figure 10

Comparison of item 11 response curves for non-NOTA and NOTA Groups

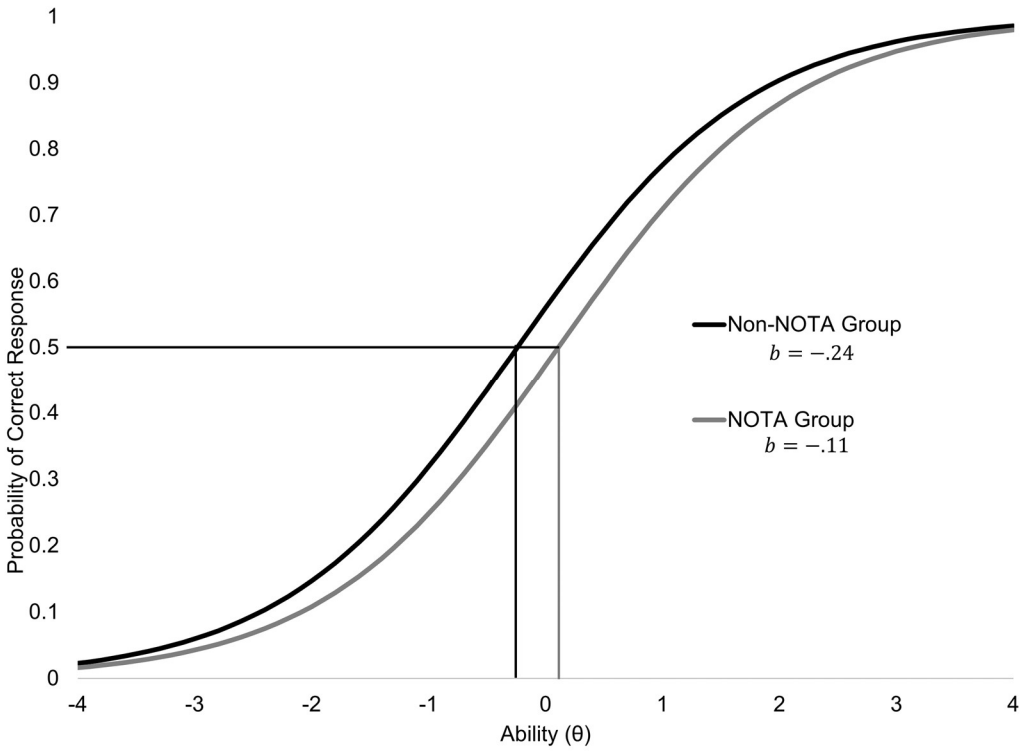


Figure 11

Comparison of item 6 response curves for non-NOTA and NOTA Groups

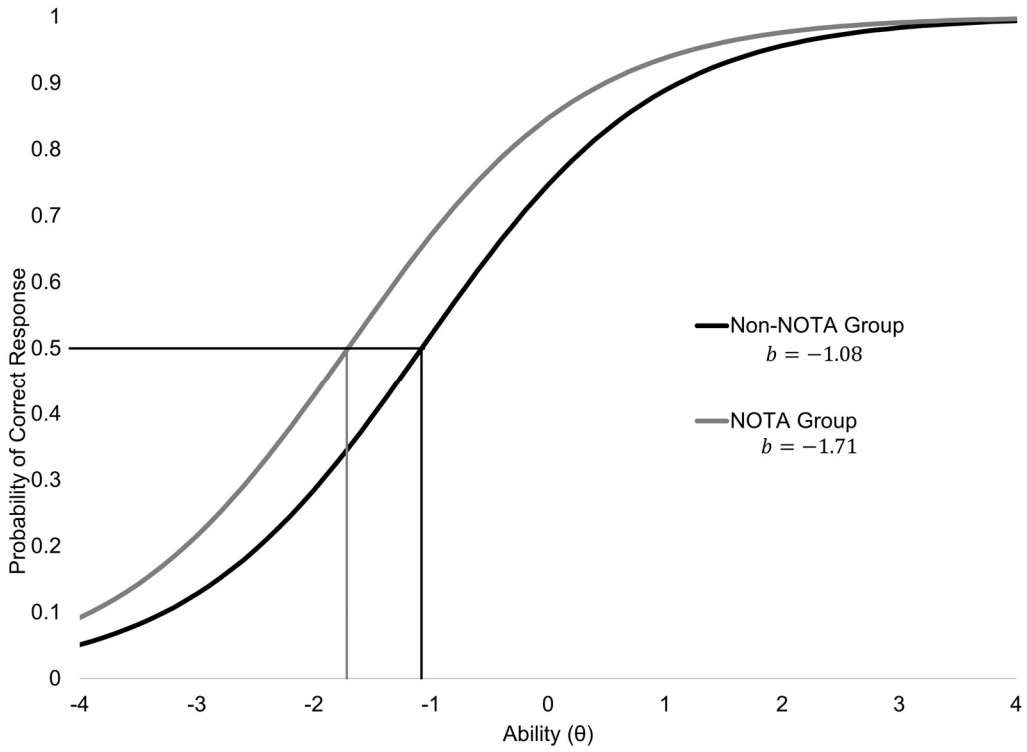


Figure 12

Model 2a TOI EAP distribution

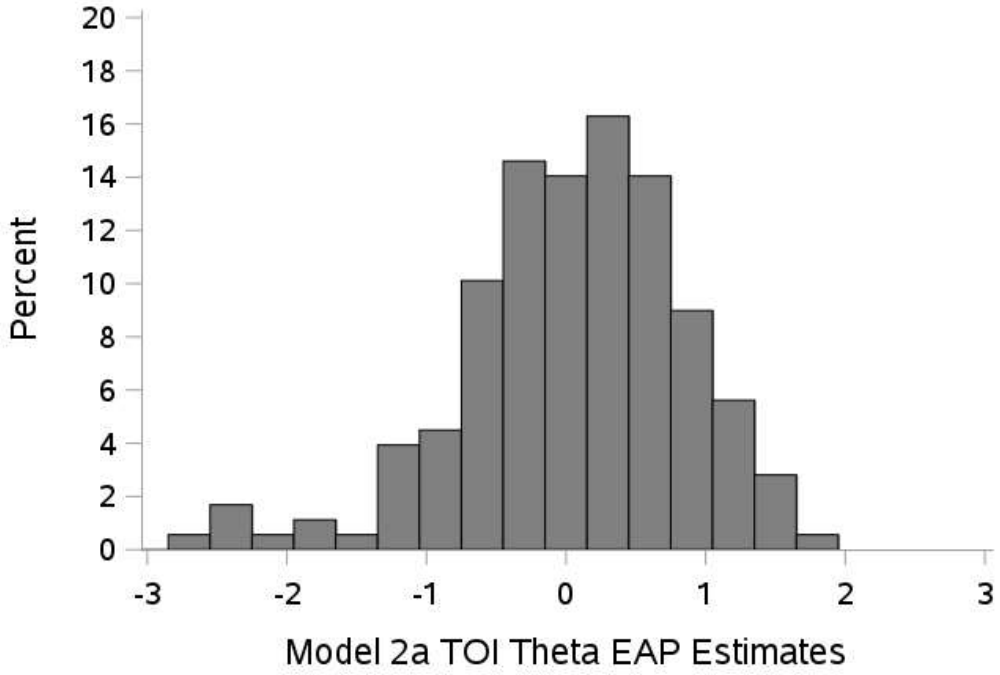


Figure 13

Model 2b TOI EAP distribution

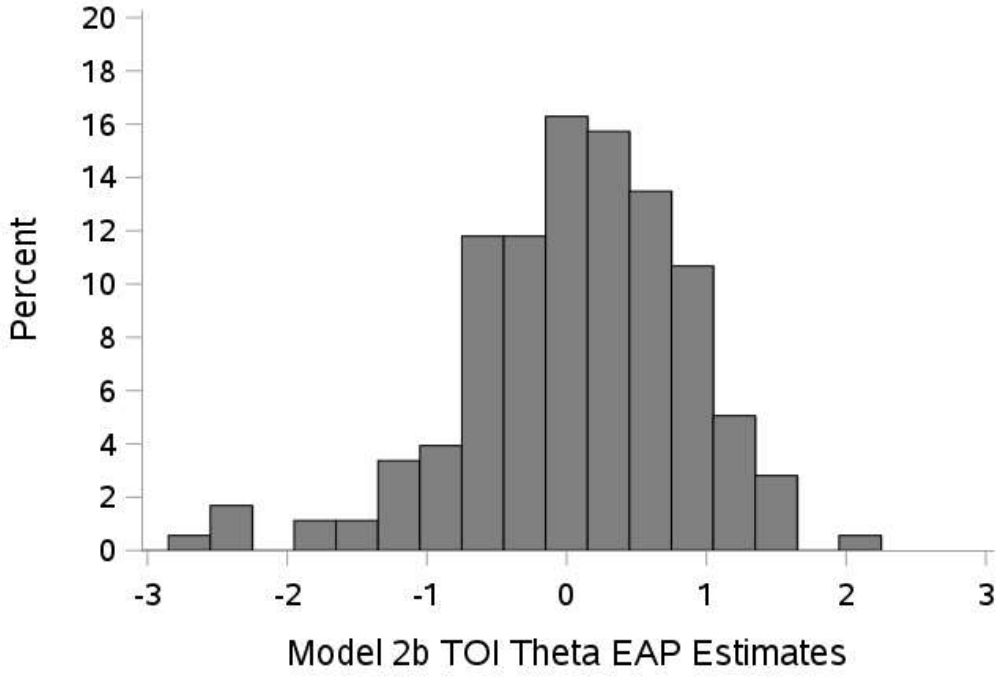
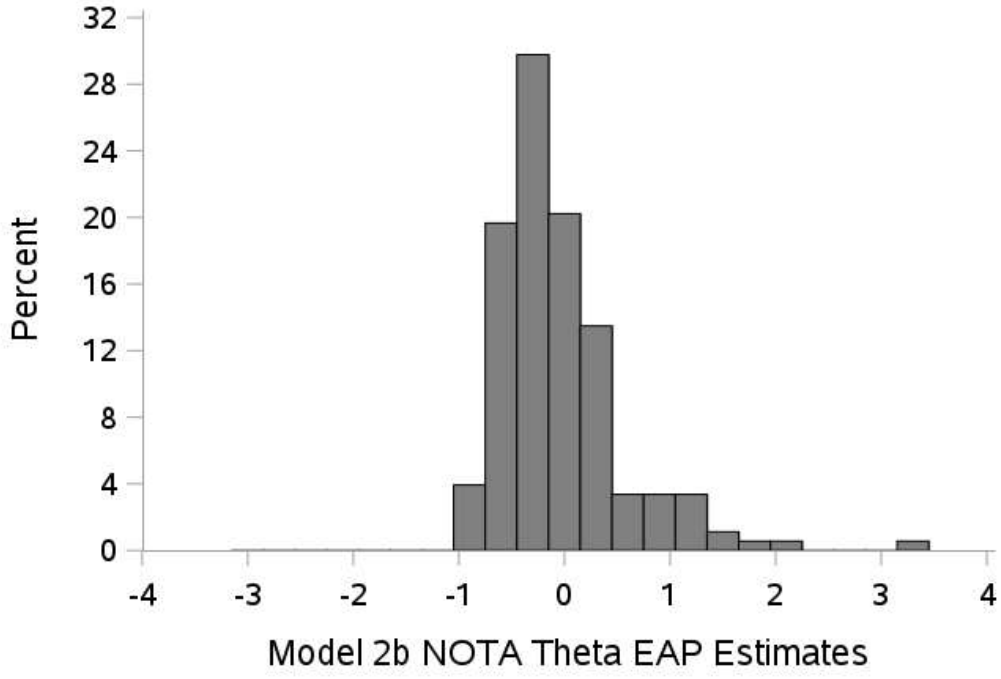


Figure 14

Model 2b NST EAP distribution



Appendix

The table below provides an example SAS dataset of the first 24 observations. This dataset was used to fit the three multilevel Rasch models presented in the first research questions. The dataset contains the first 24 observations, where six examinees (id) provide items responses to (y) four items (i1 – i4). Three of the six examinees, specifically examinee one, three, and six, received a test that contained NOTA as indicated by the 'nota_grp' column with a 1. To analyze item responses using proc nlmixed in SAS, the user must include codes for each item (i1 – i4) to indicate the associated item response (y).

Obs	id	y	i1	i2	i3	i4	nota_grp
1	1	0	-1	0	0	0	1
2	1	1	0	-1	0	0	1
3	1	0	0	0	-1	0	1
4	1	1	0	0	0	-1	1
5	2	1	-1	0	0	0	0
6	2	1	0	-1	0	0	0
7	2	1	0	0	-1	0	0
8	2	0	0	0	0	-1	0
9	3	1	-1	0	0	0	1
10	3	0	0	-1	0	0	1
11	3	1	0	0	-1	0	1
12	3	1	0	0	0	-1	1
13	4	1	-1	0	0	0	0
14	4	0	0	-1	0	0	0
15	4	0	0	0	-1	0	0
16	4	0	0	0	0	-1	0
17	5	1	-1	0	0	0	0
18	5	1	0	-1	0	0	0
19	5	1	0	0	-1	0	0
20	5	1	0	0	0	-1	0
21	6	1	-1	0	0	0	1
22	6	0	0	-1	0	0	1
23	6	1	0	0	-1	0	1
24	6	1	0	0	0	-1	1

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Blendermann, M. F., Little, J. L., & Gray, K. M. (2020). How "none of the above"(NOTA) affects the accessibility of tested and related information in multiple-choice questions. *Memory, 28*(4), 473-480.

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

Böckenholt, U. (2012a). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665.

Boynton, M. (1950). Inclusion of" none of these" makes spelling items more difficult. *Educational and Psychological Measurement*.

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.

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4).

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1-28.

Deng, S., & Bolt, D. M. (2016). A sequential IRT model for multiple-choice items and a multidimensional extension. *Applied Psychological Measurement, 40*(4), 243-257.

DeVore, S., Stewart, J., & Stewart, G. (2016). Examining the effects of testwiseness in conceptual physics evaluations. *Physical Review Physics Education Research, 12*(2), 020138.

DiBattista, D., Sinnige-Egger, J. A., & Fortuna, G. (2014). The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education, 82*(2), 168-183.

Dochy, F., Moerkerke, G., De Corte, E., & Segers, M. (2001). The assessment of quantitative problem-solving skills with "none of the above"-items (NOTA items). *European Journal of Psychology of Education, 16*, 163-177.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology, 58*(1), 116.

Embretson, S. E. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice, 35*(3), 6-22.

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4*(2), 115-124.

Garcia-Perez, M. A. (1993). In defence of 'none of the above'. *British Journal of Mathematical and Statistical Psychology, 46*, 213-229.

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.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21-35.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: The case of "None of the Above". *Evaluation & the Health Professions*, *17*(1), 123-126.

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

Haladyna, T. M., & Downing, S. M. (1989b). 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. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

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*. Routledge.

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

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

Jang, Y., Pashler, H., & Huber, D. E. (2014). Manipulations of choice familiarity in multiple-choice testing support a retrieval practice account of the testing effect. *Journal of Educational Psychology, 106*(2), 435.

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

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 64 – 80). Routledge.

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.

Kyllonen, P. C. (2016). Designing tests to measure personal attributes and noncognitive skills. In Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 190 – 211). Routledge.

Leighton, J. P., & Lehman, B. (2020). Digital module 12: Think-aloud interviews and cognitive labs. *Educational Measurement: Issues and Practice, 39*(1), 96-97.

Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D. (2013). *The BUGS book: A Practical introduction to Bayesian analysis*. New York: CRC Press

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

Mullins, C. J. (1963) Self-confidence as a response set. *Journal of Applied Psychology, 47*(2), 156 – 157.

Odegard, T. N., & Koen, J. D. (2007). “None of the above” as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15*(8), 873-885.

Pachai, M. V., DiBattista, D., & Kim, J. A. (2015). A Systematic Assessment of" None of the Above" on Multiple Choice Tests in a First Year Psychology Classroom. *Canadian Journal for the Scholarship of Teaching and Learning, 6*(3), 2.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16*(3), 223-243.

Rodriguez, M. C. (1997). *The art & science of item writing: A meta-analysis of multiple-choice item format effects.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rodriguez, M. C. (2016). Selected response item development. In Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 259 – 273). Routledge.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, *49*(2), 252-279.

SAS Institute Inc. (2020). *SAS/STAT 9.4 user’s guide*. Cary, NC: Author.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (statistical methodology), 64*(4), 583-639.

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

Sundre, D. L., & Thelk, A. D. (2007). *The Student Opinion Scale (SOS): A measure of examinee motivation*. Test Manual. Harrisonburg, VA: The Center of Assessment and Research Studies.

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

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.

Wesman, A. G., & Bennett, G. K. (1946). The use of 'none of these' as an option in test construction. *Journal of Educational Psychology, 37*(9), 541.