COMPUTERIZED ADAPTIVE TESTING FOR IDEAL POINT PERSONALITY ASSESSMENT: A COMPARISON OF TEST CHARACTERISTICS

by

JEREMIAH T. MCMILLAN

(Under the Direction of Nathan Carter)

ABSTRACT

Computerized adaptive testing (CAT) is a popular method for boosting efficiency, reducing costs, and improving examinee reactions in employee selection. Extant research has primarily established that CAT provides utility over static personality testing when the response model is monotonic (i.e., higher standing on the trait results in participant endorsement of a higher response option). Given the recent emergence of the use of ideal point item response theory (IRT) models for personality testing—which assume that higher response probability is inversely related to an individual's distance from the item—it is important that research examine whether these models support effective CAT, and the test characteristics that may play a role. This is because ideal point models require more response data than monotonic models to accurately estimate θ, potentially hindering the utility of CAT. The present study used real-data simulations to examine the performance of different CAT conditions using a pool of conscientiousness items calibrated under the generalized graded unfolding model (GGUM) on a sample of 1,724 Amazon Mechanical Turk workers. General measurement accuracy/precision and the accuracy of dichotomous employee selection decisions based upon theta estimates were examined while manipulating the cut-score adopted for employee selection, total test length,

number of pre-adaptive items presented (i.e., an initial testlet), and the use of a sequential versus multistage testing design. Results indicate that adaptive tests outperform ideal point static tests on general measures of accuracy but not on employee selection decision accuracy. The most critical test characteristic for successful adaptive testing is the presence of an initial testlet. Implications for testing theory, CAT design considerations, and future research directions are discussed.

INDEX WORDS: personality, computerized adaptive testing, multistage testing, ideal point modeling

COMPUTERIZED ADAPTIVE TESTING FOR IDEAL POINT PERSONALITY ASSESSMENT: A COMPARISON OF TEST CHARACTERISTICS

by

JEREMIAH T. MCMILLAN

B.A., Azusa Pacific University, 2008

M.S., University of Georgia, 2018

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

COMPUTERIZED ADAPTIVE TESTING FOR IDEAL POINT PERSONALITY ASSESSMENT: A COMPARISON OF TEST CHARACTERISTICS

by

JEREMIAH T. MCMILLAN

Major Professor:	Nathan Carter
Committee:	Dorothy Carter
	Joshua Miller

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
December 2019

DEDICATION

This work is dedicated to my sister and best friend, Sarah.

ACKNOWLEDGEMENTS

I would like to thank my advisor Nathan Carter for all his insight and hard work on this project. He has taught me a great deal not only about latent response modeling but also how to communicate complex information in down-to-earth, practical ways. I also gratefully thank my committee, Dorothy Carter and Josh Miller, for their keen insights and encouragement in this process. I would like to thank Kristen Shockley for playing a central role in honing my scientific writing skills throughout my graduate education. She has also taught me a great deal about persevering and overcoming obstacles in the research process. I thank Rachel Williamson for her assistance and advice on statistical approaches used in this project. Finally, I would like to thank Michael Chajewski and Charles Scherbaum for instilling in me early in my graduate career a passion for statistical analysis.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

Study Overview 5

Item Response Theory and the Generalized Graded Unfolding Model 8

CAT and Personality Assessment 12

Utility of Ideal Point Personality CAT 16

Employee Selection Cut-Score 17

Initial Item Selection 19

Multistage versus Sequential Testing 24

Test Length 25

2 METHOD 26

Simulation Data 26

Procedure 30

Study Outcomes 32

3 RESULTS 35

General Measurement Outcomes 35

Employee Selection Accuracy 39

4 DISCUSSION 42

Implications for Measurement Theory and Practice 42

Limitations 48

Future Directions 51

REFERENCES 55

APPENDICES

A R CODE FOR CAT FUNCTION 64

LIST OF TABLES

Page

Table 1: All 36 Study Conditions 7

Table 2: Summary of Dominance-Based Studies of Personality CAT.....13

Table 3: Participant Demographic Information 27

Table 4: Final 111-Item Pool Parameters and Item-Data Fit.....29

Table 5: General Measurement Outcomes for each Study Condition 36

Table 6: General Measurement Outcomes Across Study Conditions.....37

Table 7: Normalized Employee Selection Accuracy ($\mathcal{A}_{\text{norm}}$) for each Study Condition.....40

Table 8: Normalized Employee Selection Accuracy ($\mathcal{A}_{\text{norm}}$) Across Study Conditions 41

LIST OF FIGURES

Page

Figure 1: General form of the item response functions and item information functions in dominance versus ideal point models for an item on a six-point Likert scale with $\delta = 0$....9

Figure 2: Possible IIF overlay variants for a two-item ideal point testlet.....23

Figure 3: Test information and standard error of measurement for the final 111-item pool.....31

Figure 4: Test information and standard error of measurement for the final (a) 6-item static test and (b) 12-item static test.....44

Figure 5: True theta plotted against bias ($\hat{\theta}$ minus θ) for the 6-item—2 item start—multistage adaptive condition.....50

CHAPTER 1

INTRODUCTION

Computerized adaptive testing (CAT)[1] has gained popularity among large organizations as a means to increase efficiency and reduce costs in personnel assessment for both cognitive and noncognitive predictors (Kantrowitz, Dawson, & Fetzer, 2011). However, little research has systematically examined optimal CAT characteristics for personality assessment, particularly when utilizing ideal point item response theory (IRT) measurement models, which recent empirical evidence suggests more accurately model the way persons respond to personality tests than monotonic, dominance-based models (Drasgow, Chernyshenko, & Stark, 2010a, 2010b; LaPalme, Tay, & Wang, 2018; Stark, Chernyshenko, & Drasgow, 2005). Organizations may invest millions of dollars into the development of personality CATs based upon the assumption that the reduced length of such assessments compared with static assessments will result in more positive applicant attitudes, test security, and reduced scoring efforts (Fetzer, Dainis, Lambert, & Meade, 2011). However, to the author's knowledge, no evidence exists regarding how to best leverage the efficiency of CAT for ideal point personality assessment (i.e., which test features are most effective).

The purpose of the present study is to simulate multiple CAT variations using existing examinee data from a static test of conscientiousness by manipulating the total length of the test, the percentile rank used as a cut-score for a hypothetical employee selection decision (i.e., 10th, 20th, and 30th percentile), the number of items presented prior to initializing adaptation, and the

[1] Throughout, I use the acronym CAT to refer to both "computerized adaptive testing" and "computerized adaptive test."

use of stages or sequential adaptation throughout the CAT. Mirroring the diversity of real-world applications of assessments in organizational contexts, two interrelated sets of outcomes are examined: (1) general precision/accuracy of measurement, and (2) the ability of the CAT to accurately guide dichotomous employee selection decisions based on a predetermined cut-score meant to reflect the "select-out" screening typical of personality test use in organizations (Mueller-Hanson, Heggestad, & Thornton, 2003). The performance of all CAT conditions are compared relative to each other and to a static test of equal length to provide evidence regarding the utility of CAT for organizations using ideal point personality assessment for employee selection.

Extant research on CAT-based personality assessment is predominantly based upon dominance IRT models (Forbey & Ben-Porath, 2007; Hol, Vorst, & Mellenbergh, 2008; Makransky, Mortensen, & Glas, 2013; Reise & Henson, 2000; Simms & Clark, 2005). In such models, respondents are assumed to have a cumulatively higher probability of endorsing an item as their underlying trait level increases (denoted as the Greek symbol theta, θ). On the other hand, ideal point models suggest that as one's distance from an item increases in either direction, the probability of endorsing that item decreases. Ideal point models are believed to be a better fit to personality data than dominance-based models due to the manner in which individuals consider their relative distance from the location of an item, regardless of whether they are higher or lower on the underlying trait (LaPalme et al., 2018). Thus, as a primary contribution, the present study aims to determine how well short ideal point personality CATs perform in terms of accuracy and precision.

It is noteworthy that a small body of research has emerged supporting the utilization of complex ideal point models for CAT in large-scale testing environments such as the Army (e.g.,

Drasgow, Stark, Chernyshenko, Nye, & Hulin, 2012) and the Navy (Houston, Borman, Farmer, & Bearden, 2006). All such applications have utilized a forced-choice item response format (discussed in more detail later). However, to the author's knowledge, evidence regarding the adequacy of CATs using single-stimulus ideal point items (i.e., traditional Likert scale response options) remains nonexistent. This represents a significant gap considering that the single-stimulus format is used much more in practical applications than the forced-choice format due to lower cost and ease of implementation (N. Carter, personal communication, February 23, 2019).

A second contribution of the present study is the examination of multiple uses of CAT scores. Personality testing may be used for a variety of purposes in organizational settings, but is most appropriately used to either make "select-out" decisions based upon a predetermined cut-score or for general assessment of applicants/incumbents across the entire trait range for validation research (Kantrowitz et al., 2011; Mueller-Hanson et al., 2003). The latter purpose requires precise trait estimates at all levels and has been the emphasis of previous examinations of optimal CAT characteristics (e.g., Forbey & Ben-Porath, 2007; Hol et al., 2008; Houston et al., 2006; Makransky et al., 2013). Use of a cut-score simply requires certainty that an examinee's true trait level is above or below the cutoff criterion. Thus, identification in the present study of how a CAT operates against both outcomes provides information about the types of organizational decisions it can reliably inform. Furthermore, by manipulating severity of the cut-score, the present study provides valuable information to practitioners regarding potential boundary conditions for the accuracy of the CAT in informing employee selection decisions.

The present study also examines specific algorithmic characteristics that may optimize ideal point CAT performance. It is well established that longer tests provide more stable estimates of psychological constructs than shorter tests (Crocker & Algina, 1986). However, this

principle is frequently at odds with practical considerations around cost and testing time in employee selection contexts (Ployhart, Schmitt, & Tippins, 2017). Thus, it would be very useful for organizations to have guidelines regarding the shortest possible CAT length that still provides reliable scores. The need for such information is especially cogent when using ideal point modeling, considering that θ estimation generally may require more items than dominance-based modeling to make sense of different types of response patterns (Dalal, Withrow, Gibby, & Zickar, 2010). Hence, the present study examines a "very short" (6-item) CAT and a "short" (12-item) CAT to provide evidence regarding what constitutes an acceptable test length.

Because ideal point items have item response functions that differ in functional form compared with dominance items (described in further detail in a later section), research is needed to determine CAT characteristics that handle these assumptions effectively. To start, this study examines whether the traditional practice in dominance-based CAT of initially presenting a single item to begin estimating θ and adaptively selecting items is effective when an ideal point response model is used. I posit that such an approach is not effective due to the ambiguity around the substantive meaning of a response to any single ideal point item. Thus, I test the effectiveness of short initial testlets assembled to explicitly sample items from across the entire trait spectrum, prior to beginning adaptation, as a means of increasing early reliability of the interim estimate of θ (denoted as theta-hat, or $\hat{\theta}$). The term *testlet* is sometimes used in the literature to refer to a group of items linked to the same stimulus and thus expected to lack conditional independence (e.g., Wainer, Bradlow, & Du, 2000). For clarity, the present usage of the term merely indicates a group of items that are presented simultaneously to an examinee, irrespective of item content.

Beyond the starting point of the CAT, I also examine the effectiveness of sequentially presenting a single item at a time versus presenting item testlets in stages throughout the

remainder of the test. Comparing tests with differing levels of adaptiveness but equal lengths allows for identification of a potential ideal balance between maximizing the rapidity of homing in on θ (more adaptiveness) and maximizing the reliability of the estimate of $\hat{\theta}$ prior to potentially presenting localized items that are not informative (less adaptiveness). It is also possible that a unique combination of initial item presentation strategy and ongoing item presentation strategy will result in optimal precision. Indeed, past research has suggested that differing adaptive test designs, such as purely sequential/traditional CAT (Stark, Chernyshenko, Drasgow, & White, 2012), multistage tests (MSTs; Stark & Chernyshenko, 2006), or a hybrid approach (Wang, Lin, Chang, & Douglas, 2016) all have value. But research has not compared how these varying approaches impact estimates in an ideal point CAT. Identifying relative performance of these designs has practical relevance, considering that MSTs may be favored to traditional CATs in practice due to higher levels of content control, easier expert review of non-statistical considerations prior to test administration, and examinee reactions (Luecht, Brumfield, & Breithaupt, 2006; Stark & Chernyshenko, 2006).

Study Overview

The present study uses a series of real data simulations (actual examinee responses to a complete static test are used to simulate results of differing CAT conditions) to examine the efficiency of CAT for assessing the personality trait of conscientiousness under ideal point measurement conditions. Conscientiousness serves as the construct of interest in the present study due to its common usage in selection batteries and high predictive validity (Judge, Rodell, Klinger, Simon, & Crawford, 2013); however, results are expected to generalize to other personality traits.

Table 1 displays all study conditions. The present study is a partially-crossed design with a total of 36 conditions. Six static test conditions are examined: 3 (Employee selection percentile cut-score: 10^{th}, 20^{th}, 30^{th}) x 2 (Test Length: 6 items, 12 items). Thirty adaptive test conditions are examined: 3 (Employee selection percentile cut-score: 10^{th}, 20^{th}, 30^{th}) x 3 (Initial Item Selection: single item, two-item testlet, three-item testlet) x 2 (Test Design: fully adaptive sequential, multistage) x 2 (Test Length: 6 items, 12 items), minus six non-examined cells. These six non-examined cells are those in which a single item is presented initially followed by multistage testing. Although these cells represent valid levels in a fully-crossed design, there is no theoretical rationale for their inclusion. The presentation of a single item prior to adaptation suggests confidence in the reliability of the estimate derived from the item response. In reality, the beginning of the CAT is when estimation is most unstable. To present a single item followed by testlets is precisely the opposite of what theory and empirical evidence would suggest is logical (Wang et al., 2016). If testlets do provide value, it is most likely to be at the beginning of the CAT.

Each CAT condition is evaluated on both general measurement effectiveness and effectiveness for making employee selection decisions. First, conditions are evaluated across and conditional on θ based upon root mean square error (RMSE; precision), the mean signed difference between $\hat{\theta}$ and θ (systematic bias), and the correlation between $\hat{\theta}$ and θ ($R_{\theta\hat{\theta}}$; accuracy). Second, accuracy of the dichotomous employee selection decision is determined by calculating the proportion of individuals correctly classified above the cut-score by comparing $\hat{\theta}$ with θ after normalizing based on the cut-score used.

In the following sections, because of the critical bearing that the choice of response model has on CAT, I first provide a brief background on ideal point IRT models, focusing on the

Table 1

All 36 Study Conditions

(a) Selection Cut-Score = 10th percentile

		Total Items	
		6	12
Start	Static	Static	Static
	Single-Item	Sequential	Sequential
	Two-Item	Sequential	Sequential
		Multi-Stage	Multi-Stage
	Three-Item	Sequential	Sequential
		Multi-Stage	Multi-Stage

(b) Selection Cut-Score = 20th percentile

		Total Items	
		6	12
Start	Static	Static	Static
	Single-Item	Sequential	Sequential
	Two-Item	Sequential	Sequential
		Multi-Stage	Multi-Stage
	Three-Item	Sequential	Sequential
		Multi-Stage	Multi-Stage

(c) Selection Cut-Score = 30th percentile

		Total Items	
		6	12
Start	Static	Static	Static
	Single-Item	Sequential	Sequential
	Two-Item	Sequential	Sequential
		Multi-Stage	Multi-Stage
	Three-Item	Sequential	Sequential
		Multi-Stage	Multi-Stage

Note. "Sequential" and "Multi-Stage" refer to item presentation strategy *after* the start (e.g., the Two-Item Start, Sequential condition with 6 total items would present two initial items simultaneously, followed by four individual items in sequence).

popular generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) that is used in the present study. I then discuss extant empirical evidence regarding the value of CAT compared with static testing in personality assessment. Finally, I discuss the rationale for examining each of the various manipulated CAT characteristics within the present study.

Item Response Theory and the Generalized Graded Unfolding Model

IRT encompasses a broad class of models in which the probability of endorsement of an item is related non-linearly to one's standing on the latent trait of interest. Because IRT allows for scaling individuals and scale items on the same metric, it is an incredibly useful basis for CAT to efficiently capture θ, iteratively presenting items tailored to an individual's current $\hat{\theta}$ estimate and then updating $\hat{\theta}$ (Embretson & Reise, 2000). Two broad classes of IRT models, dominance (Likert, 1932) and ideal point (Thurstone, 1927), may be used for CAT but make different assumptions about the psychological process of responding to an item. These differing assumptions have implications for how the item response function (IRF) is modeled, the resultant shape of item information, and practical considerations around creating scale items.

As seen in the left-hand side of Figure 1, dominance-based IRF's predict that the probability of endorsing an item increases monotonically as θ increases (i.e., individuals with higher θ will "dominate" items lower on the trait continuum), whereas ideal point IRF's predict that the probability of endorsing an item increases as one's distance from the item decreases. Complete endorsement of the item (i.e., selecting the highest possible response option) in dominance models suggests that an individual's θ lies at some point above the item's location (specifically, anywhere after the ogival curve levels out). On the other hand, complete endorsement of the item in ideal point models suggests that the individual's θ is approximately equal to the item location. Anything less than complete endorsement in dominance models

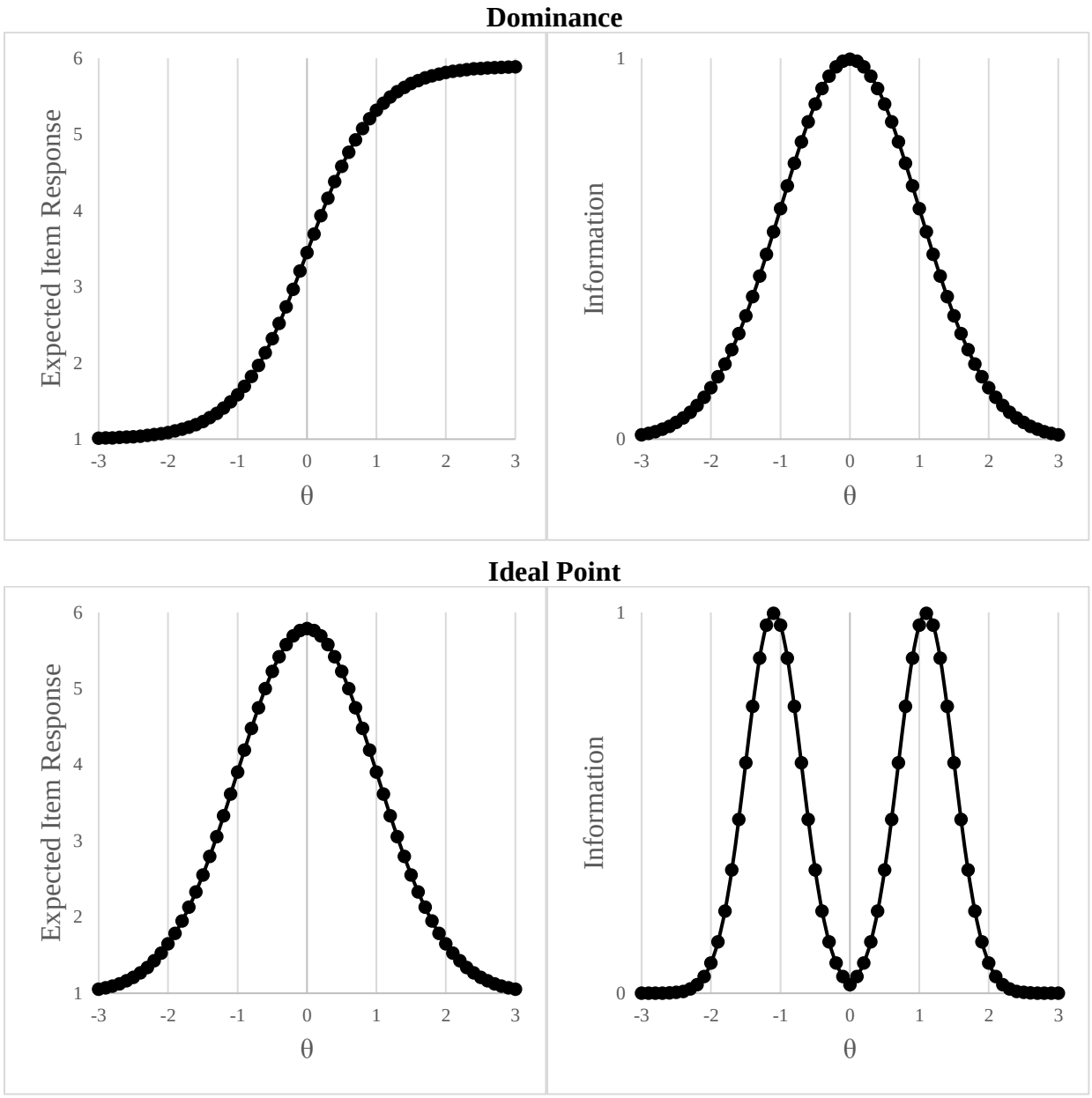


Figure 1. General form of the item response functions and item information functions in dominance versus ideal point models for an item on a six-point Likert scale with $\delta = 0$.

suggests progressively lower trait standing, whereas less than complete endorsement in ideal point models suggests distance from the item location in either direction. As a result, any objective response category to an ideal point item (e.g., disagree) has two possible subjective response categories (i.e., disagree from above or disagree from below the item; Roberts et al., 2000). Only through triangulation based on multiple item responses or prior information can the correct subjective response category be ascertained.

Item information in IRT is analogous to reliability in classical test theory and is closely tied to the IRF. Information is plotted conditional on θ and is inversely related to the standard error of measurement (SEM), providing evidence regarding where an item is reliable on the trait continuum and to what degree (Embretson & Reise, 2000). Item information serves a highly useful role in CAT as it allows for items to be selected based upon maximizing information at $\hat{\theta}$. Because items display maximum information where discrimination is steepest, the right-hand side of Figure 1 illustrates that information for dominance items peaks at the location of the item. However, for ideal point items, information is double-peaked about the item location, providing no information when θ is exactly equal to the item location.

As a result of their differing statistical properties, dominance and ideal point modeling have differing implications for generating scale items (Chernyshenko, Stark, Drasgow, & Roberts, 2007). Item writing in a dominance-based paradigm focuses on generating only items with extreme locations. For instance, to assess punctuality, a researcher might create the item “I am never on time” to tap the low end of the trait and the item “I am always on time” to tap the high end of the trait. All negatively worded items would then need to be reversed-scored prior to modeling so that they are in the same direction as positively worded items. Dominance models are not able to effectively accommodate items tapping the mid-range of the trait spectrum, such

as “I am sometimes on time.” To the degree that individuals very high on punctuality disagree with such a moderate item, the ogive-shaped curve in the first panel of Figure 1 will not fit the data. Ideal point item generation, however, allows for generating items designed to tap the entire range of the trait spectrum. An ideal point model would effectively capture the trait standing of low, moderate, and high punctuality individuals based on strong endorsement of the low, moderate, and high location items, respectively.

The present study uses ideal point items calibrated under the generalized graded unfolding model (GGUM; Roberts et al., 2000). This model is represented mathematically for each option response curve as follows:

$$P[Z_i = z|\theta_j] = \frac{\exp\left(\alpha_i\left[z(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-z)(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}\right]\right)}{\sum_{w=0}^{C}\exp\left(\alpha_i\left[w(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right)}, \quad (1)$$

where α represents item discrimination (slope), δ represents item location (difficulty) of item *i*, τ indicates the *k*th item boundary for item *i*, C represents the highest response option, and M = 2C + 1. The probability of endorsing a specific option is a function of one’s distance from the overall item location, the spread of the item options, and the item’s discrimination. Whereas each individual option response curve (other than complete endorsement) will resemble a symmetrical double-peaked function, the aggregate of all response curves for a single item will produce a bell-shaped item response curve, as seen in Figure 1. The relevance of the differing functional forms of dominance models and ideal point models, such as GGUM, for CAT functioning are discussed next.

CAT and Personality Assessment

Evidence abounds for the general utility of CAT for personality assessment. Overall estimates suggest that CAT allows for approximately one-third to one-half as many items as a static test to obtain comparable precision because a CAT can strategically maximize the information gained from each item presented (van der Linden & Glas, 2010). However, based upon the nature of the construct assessed and practical considerations, CAT may not be uniformly useful across all contexts (Reise & Henson, 2003; van der Linden & Glas, 2010). I first discuss applications of dominance-based personality CAT—which constitutes the majority of research to date. I then discuss both conceptual considerations around and empirical evidence regarding ideal point personality CAT.

Dominance-based CAT. Within the dominance response paradigm, evidence has borne out the utility of CAT across a wide variety of contexts. Table 2 summarizes extant research in this area. All studies in this table reached qualitatively similar conclusions that CAT provided efficiency gains over a static test. Universally, there is also a direct tradeoff between the efficiency of the CAT and the level of precision, such that presentation of fewer items results in less accurate estimates (e.g., Hol et al., 2008). Alternatively stated, CATs terminated based upon less stringent SEM are shorter than those with more stringent requirements.

Although all of the above studies found an efficiency gain from CAT over static testing, it is important to distinguish between efficiency and utility per se. Reise and Henson (2000) concluded that a CAT of personality does not provide appreciable utility over and above an equally long, well-designed static test. Their study examined the creation of a 4-item test from an original pool of 8 items. CAT is unlikely to add value with such a limited item pool because choosing four optimal items on-the-fly using CAT versus a priori based on favorable properties

Table 2

Summary of Dominance-Based Studies of Personality CAT

Study	Measure	Response Model	Item Pool Size	Termination Criteria(on)	Item Pool Usage[a]
Ben-Porath et al., 1989	MMPI	Non-IRT	383	(1) Sufficient sum score for classification; (2) Full scales presented only to those above sum score threshold for classification	Across 32 conditions, range from 68.7%-92.2%
Kamakura & Balasubramanian, 1989	CPI-socialization	2PL	44	(1) Fixed-length; (2) SEM-based; (3) Combination of SEM and length-based	34.1%; 38.6%; 39.8%
Waller & Reise, 1989	MPQ-absorption	2PL	34	(1) Confidence interval does not contain cut-score; (2) Fixed length	26.9%; 50.0%
Reise & Henson, 2000	NEO PI-R	Samejima's GRM	240	Fixed-length	12.5%; 25.0%; 37.5%; 50.0%
Simms & Clark, 2005	SNAP	2PL	297	Combination of minimum length, SEM-based, and minimum information of remaining items	63.5%
Forbey & Ben-Porath, 2007	MMPI-2	Non-IRT	557	Sufficient sum score for classification	79.4%; 78.4%; 82.3%; 82.5%
Hol et al., 2008	Adjective Check List-dominance	Samejima's GRM	36	SEM-based	7.5%; 10.0%; 13.3%; 19.0%; 32.6%; 78.1%
Makransky et al., 2013	NEO PI-R	Multi-dimensional GPCM	240	Fixed-length	25.0%; 37.5%; 50.0%; 75.0%

Note. MMPI = Minnesota Multiphasic Personality Inventory. CPI = California Psychological Inventory. MPQ = Multidimensional Personality Questionnaire. NEO PI-R = NEO Personality Inventory-Revised. SNAP = Schedule for Nonadaptive and Adaptive Personality. 2PL = 2-parameter logistic model. GRM = Graded response model. GPCM = Generalized partial credit model.

[a]Values are separated by different CAT conditions, if tested, but collapsed across different traits.

will tend to result in the same presentation of the four items with the most favorable discrimination parameters. This study does not definitively indicate that CAT is not useful in personality assessment. Rather, it highlights that there is a limit to how useful a CAT may be with very short test lengths and small item banks. This problem may be exacerbated when moving from dominance-based to ideal point modeling, as discussed in the next section.

Ideal point CAT. All IRT-based studies on CAT for personality assessment discussed above were conducted using a dominance-based response model. Traditional CATs using dominance-based models are relatively straightforward. The assumption of monotonic item response functions allows one to conclude that lack of endorsement of an item indicates probabilistically that the individual's latent trait standing is lower than the item location. It would logically follow that the next item presented should be at a location lower than said item. However, assessing personality using one or few ideal point items presents a psychometric challenge (Williamson, Castille, & Harris, 2017). Because of the aforementioned feature of ideal point that an objective response can be associated with one of two possible subjective responses, certain response patterns may prove more difficult for ideal point than dominance to pinpoint. Bayesian θ estimation methods technically allow for an estimate to be derived based upon any response pattern. But variance around that estimate may be quite large, rendering the aim of CAT to produce precise, efficient estimates difficult. In fact, poor item selection made by a CAT in response to inaccurate $\hat{\theta}$ estimates could potentially result in less efficiency than intentionally presenting a variety of item locations regardless of examinee θ. Such an approach ensures that the full trait spectrum is covered (as would be done in a static test).

Despite these conceptual challenges, limited empirical evidence generally supports the utility of CAT for ideal point measurement. The seminal study on the topic demonstrated that an

ideal point CAT of attitudes toward abortion was able to effectively capture the trait with as few as 7 or 8 items, drawn from a 20-item pool (Roberts, Lin, & Laughlin, 2001). However, the nature of the trait under study may matter. Attitudes may cover a smaller, more homogenous construct space than a multifaceted personality trait such as conscientiousness (Judge et al., 2013). Thus, ideal point CAT may have an easier time capturing an attitudinal trait than personality.

The remainder of extant ideal point personality CAT studies all utilize a forced-choice response format. In this format, rather than indicating level of agreement to a single stimulus, respondents are asked to choose between two or more stimuli the one that is most similar to them. Paired stimuli may be drawn from the same or different traits (i.e., a unidimensional versus multidimensional test) and are drawn from different points on the trait continuum to optimize information for a given respondent (Stark et al., 2005). Examining the unidimensional case, past evidence suggests that these CATs are able to accurately measure multiple personality traits, demonstrating acceptable reliability and predictive validity (Houston et al., 2006) using as few as six item pairs per trait (Schneider, McLellan, Kantrowitz, Houston, & Borman, 2009). Results are similar when examining multidimensional ideal point CATs, demonstrating that five items per trait may be sufficient for accurate θ recovery. Mirroring general CAT research, adaptive ideal point forced-choice tests are able to effectively halve the length of the test compared with static testing (Drasgow et al., 2012; Stark et al., 2012).

Despite evidence that ideal point CAT has been found to yield benefits in the above operational settings using highly complex response models, scant evidence is available as to how an ideal point CAT operates using commonly used single-stimulus response formats. Although full treatment of the psychometric assumptions underlying forced-choice response models are

beyond the scope of the present paper, these models are different from single-stimulus in the handling of θ estimation and the selection of pairs of items to be presented (Brown & Maydeu-Olivares, 2012; Stark et al., 2005). As a result, the efficiency of CATs using forced-choice formats may be fundamentally different than that of CATs using single-stimulus formats. As stated previously, this represents a significant gap in the literature considering the widespread use of ideal point single-stimulus items for personality assessment.

Utility of Ideal Point Personality CAT

The overarching aim of the present study is to identify the utility of CAT for ideal point personality assessment. CAT is costly and, as discussed previously, may not be worth the cost in all contexts (Reise & Henson, 2000). Comparing different versions of an ideal point personality CAT is useful for identifying which of these versions function better in a relative sense. However, it is also useful to compare CAT as a whole against a static test to serve as a baseline. To the extent that a static test captures θ as accurately as a CAT with the same number of items, there is little justification for developing the CAT. This issue is likely to be particularly salient when an inordinately short test length is used. As an extreme example, a test constrained to a total of two items is liable to operate most effectively—albeit not effectively per se—across all examinees by statically presenting two highly discriminating items near the center of θ. Utilizing a CAT in which the second item is presented adaptively as a function of the unreliable response to the first item is liable to yield highly unstable θ estimates. In sum, there is likely a lower bound of item length beneath which CAT algorithms are not able to gain sufficient momentum. As seen in Table 1, the present study includes two conditions that are static (i.e., non-adaptive) and serve as a frame of comparison for all CAT conditions of the same length. These static tests are generated by using best practices for scale creation. That is, items are chosen from the same

pool used in the CAT conditions with the aim of tapping the entire range of θ and maximizing discrimination parameters.

I do not offer a specific prediction regarding the overall, absolute value of CAT over static testing in ideal point personality assessment. Rather, I posit that the proportion of candidates selected (i.e., the cut-score adopted) and certain test characteristics will be differentially associated with efficiency and accuracy of employee selection decisions based on the CAT. Discussion of each of these characteristics in turn comprises the remainder of this paper. First, the impact of setting the cut-score for employee selection at different percentile ranks is explored. Second, the number of items presented in an initial testlet prior to $\hat{\theta}$ estimation is discussed as a potential solution for the ambiguity surrounding the meaning of the response to any single ideal point item. Third, the use of fully adaptive (sequential) designs versus multistage designs is discussed in terms of relative advantages to CAT efficiency. Finally, length-based CAT termination criteria of 6 items versus 12 items are presented as a means to explore the relative tradeoff between item savings and precision.

Employee Selection Cut-Score

Organizations may score and utilize assessments for employee selection in a variety of ways. In large-scale testing contexts, it is common to determine a minimal cut-score that applicants must surpass to be considered for hire. Decisions around where to set this cut-score vary based on a number of considerations, including the score anticipated to produce minimum acceptable on-the-job performance, anticipated selection ratio, and adverse impact (Cascio & Aguinis, 2011). More broadly, organizations may utilize an assessment from a "select-in" perspective—eliminating as many applicants as possible and retaining only those with the most favorable scores—or a "select-out" perspective—removing the bottom of the distribution where

scores are particularly low. Due to rampant faking on personality assessments, the top end of the distribution is liable to include both those who are genuinely high on the desired trait as well as those who have artificially inflated their responses. As a result, some researchers have suggested that organizations only utilize personality scores for selecting out the bottom of the distribution (Mueller-Hanson et al., 2003). Such an approach serves not only to eschew unfairness for "true" high scores but also removes those most undesirable for selection (i.e., those who either cannot or will not identify and enact acceptable on-the-job behaviors).

In addition to the conceptual issues above, the specific choice of a cut-score has important implications for measurement outcomes. Social validity or a highly competitive job market may lead organizational decision-makers to set low cut-scores. However, more extreme scores (i.e., those particularly low or high on the score distribution) are generally measured with less precision than scores in the middle of the distribution. This may become problematic for CATs that are fixed length. Namely, choosing a more extreme cut-score will likely result in more measurement error and greater misclassification of those whose true θ surpasses/does not surpass the cut-score. The present study explores the impact of this issue by examining outcomes for CAT conditions with cut-scores set at the 10^{th}, 20^{th}, and 30^{th} percentile of the θ distribution. It is common for large organizations with high human capital demands to use such low cut-offs (e.g., the military; Drasgow et al., 2012; Stark et al., 2014). But the pattern of findings uncovered from manipulating the extremity of the cut-score should be equally useful for informing select-in assessment. Assuming a symmetric distribution, the measurement outcomes of the 10^{th}, 20^{th}, and 30^{th} percentile conditions should be mirrored in the 90^{th}, 80^{th}, and 70^{th} percentiles, respectively. The common link between these opposing approaches to selection is that the extremeness of scores deemed acceptable/unacceptable has important measurement implications.

Initial Item Selection

From the start of a CAT, the aim is to capture θ as efficiently as possible. As will become evident in the discussion below, the best method for doing so in ideal point may be different than in traditional dominance-based tests. I propose the potential utility of a two-item or three-item testlet, grounding optimal assembly of these testlets in the idiosyncrasies of ideal point modeling.

The theta indeterminacy challenge. The typical starting strategy for CAT is to assume a normal prior distribution for θ and present an initial item that maximizes Fisher's information at θ = 0 (van der Linden & Glas, 2010). When using a dominance model, this effectively presents an item with a location very close to θ = 0. This correspondence between location and peak information is not true of ideal point items. Therefore, maximizing information using a single ideal point item at θ = 0 would require presenting an item intentionally above or below 0 (as any item location at θ would provide no information). The issue is that any level of examinee (dis)agreement to such an item, apart from selecting the highest response option, leaves the algorithm with no information on which direction to head for presentation of subsequent items. Practically speaking, this "theta indeterminacy" based on initial item responses will result in very large variance around early estimates and poses a challenge for selecting an item that will maximize information for a given examinee. In fact, an interesting conundrum arises: If an individual strongly disagrees to an item near the center of the trait spectrum, we can be confident that they are not near the center. However, because this response is not informative for $\hat{\theta}$ estimation (Roberts et al., 2001), we are left with (a) no $\hat{\theta}$ estimate in the case of maximum likelihood estimation or (b) a Bayesian estimate deriving information solely from the prior distribution. Assuming a normal prior, the Bayesian estimate would lead us to present the next item once again near the center of the trait spectrum, even though we have already ruled that

region out! In theory, then, such an approach would require an inefficient number of items be presented for θ estimation to build momentum in eliciting variance and triangulating any level of θ not in the center of the distribution.

One solution might be to elicit variance in a given examinee's responses using a testlet consisting of two or more items selected from across the range of θ, prior to the first $\hat{\theta}$ estimate (Williamson et al., 2017). Such an approach has been shown to be incrementally useful in dominance item testing to increase $\hat{\theta}$ stability prior to test adaptation (Kamakura & Balasubramanian, 1989; Wang et al., 2016). However, it may be much more useful for ideal point item testing compared with dominance if it effectively addresses the theta indeterminacy challenge. Several decision points in the assembly of initial testlets need to be addressed. These include how many items to include in the testlet, the location of these items, and/or the θ range(s) for which to maximize information. I address such decisions in the following sections.

Two-item testlet. Theoretically speaking, the theta indeterminacy issue could be addressed with as few as two items. Given any two partially but not entirely overlapping item response curves (IRCs), an individual can be theoretically triangulated based on relative agreement to both items (albeit this represents an idealized scenario in which measurement error does not play a role, and the response scale is granular enough to pick up on differential agreement to the two items). Put another way, the subjective meaning of the level of agreement to one item is anchored based on level of agreement to the other. There is, however, one case that still represents a blind spot using this methodology. Given that the IRCs overlap, complete lack of agreement to both suggests that the individual is either well below the pair of items or well above. Admittedly, this leads to the same theta indeterminacy as described after a single item response. The crucial difference, from a practical perspective, is the small percentage of

examinees for whom this case is anticipated to arise with the use of a testlet. Indeed, invariant response patterns (i.e., all strongly agree or all strongly disagree) are a pernicious issue that extend to static ideal point tests as well (Williamson et al., 2017). Thus, the present solution is not aimed at eliminating all uncertainty, but rather with greatly lessening the issue on the aggregate.

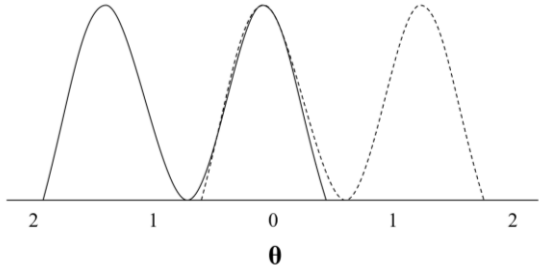
In addition to determining the size of the testlet, one must determine optimal parameters of the items selected. In the two-item testlet case, simply selecting the two items that maximize information for a given point estimate—as one might do in a traditional dominance test—yields the unsavory possibility that the two item locations/response functions will be too close to overlapping. This does little to assuage the initial theta indeterminacy issue, as similar responses to both items by an examinee accomplishes little in triangulating where that individual lies on the trait spectrum. What is really needed then are two items with optimal spacing on the spectrum such that estimation can triangulate where an individual falls in relation to those items.

An initial item selection criterion based solely on item location, however, (i.e., one item at $\delta = -1$ and one item at $\delta = 1$) may result in the presentation of suboptimal items (low discrimination and low information), despite desirable item locations. Thus, I propose the following decision rule when assembling an initial two-item testlet in an ideal point personality test: (1) present from among all items with $\delta < 0$ the item that maximizes information at $\theta = 0$, and (2) present from among all items with $\delta > 0$ the item that maximizes information at $\theta = 0$. This strategy is equivalent to selecting one item for which the *second* item information function (IIF) peak maximizes information at $\theta = 0$ and selecting one item for which the *first* IIF peak maximizes information at $\theta = 0$, respectively. The purpose of this decision rule is to prioritize higher quality items (namely, those with high discrimination) while ensuring the items are

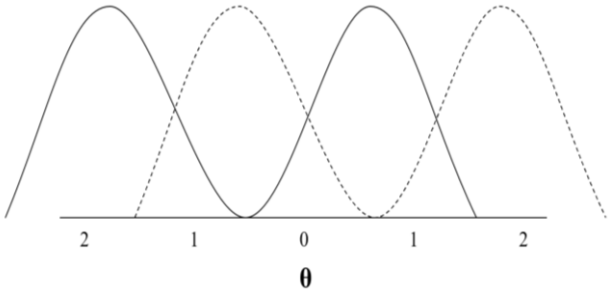
necessarily distant from each other. Theoretically possible variants of this decision rule with qualitatively different intersections of the two IIF's can be seen in Figure 2. The exact parameters of the items to be presented will be dependent upon a given CAT's available item pool. In all cases, however, the two-item testlet will include both the single item that would have been presented if solely maximum information at $\theta = 0$ were considered, plus an additional item. Due to the increased reliability around $\hat{\theta}$, I propose that the remainder of the CAT may more efficiently narrow in on θ in the two-item testlet case compared with a single item.

Three-item testlet. Despite the theoretical sufficiency of the two-item solution, measurement error is an issue such that the probability of inconsistent item responses (e.g., strongly agreeing to items in different locations) remains relatively high with only two items. All else being equal, more items lessen measurement error in θ estimation. But for any given fixed length test, the addition of more items to the initial testlet clearly comes at a cost: a point of adaptiveness within the test as a whole is lost. The optimal number of items in the initial testlet is therefore unclear, requiring empirical evidence regarding how these different conditions fare when pitted against each other. I propose adding only one additional item as a means of increasing precision over the two-item testlet while allowing as many points of adaptiveness as possible to remain in the test as a whole. Specifically, the three-item testlet should consist of the two-item testlet assembled according to the rule in the previous section, with the addition of that item remaining in the pool that maximizes information at $\theta = 0$. In summary, the present study compares three different starting conditions: a single item, a two-item testlet, and a three-item testlet.

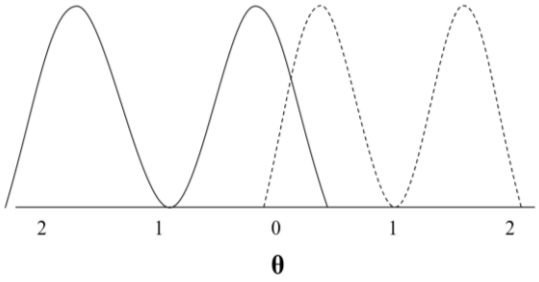
(a) Overlapping peaks



(b) Overlap on the “interior” slope



(c) Overlap on the “exterior” slope



(d) One IIF subsumed by the other

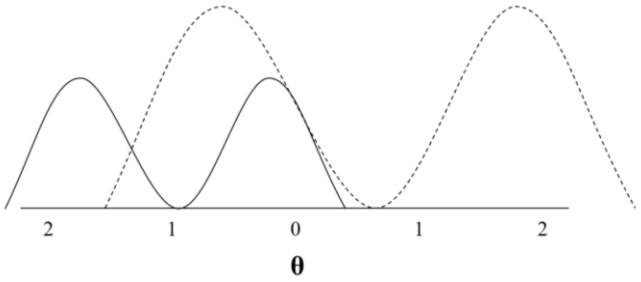


Figure 2. Possible IIF overlay variants for a two-item ideal point testlet.

Multistage versus Sequential Testing

After initial presentation of the testlet or first item, items may be presented in sequence, with adaptation after each item (i.e., a traditional sequential CAT), or testlets may continue to be presented throughout the remainder of the test (i.e., multistage testing). Efficiency in the traditional sense would seem to be maximized by adapting after each item to leverage as much information as possible in the selection of every single item presented, but efficiency is ultimately relative. To the extent that item selection "bounces around" due to unreliable estimates in sequential testing, multistage testing may be a better way to estimate θ in a slower, steadier manner (Wang et al., 2016). This issue may be particularly relevant in an ideal point modeling context. Estimation techniques, generally speaking, have a harder time with all possible response patterns to ideal point items compared with dominance (Dalal et al., 2010). Thus, although it may make sense in dominance to begin adapting as soon as possible, it may be beneficial to recognize the coarseness of measurement in ideal point and attempt to remedy that coarseness by boosting reliability in the form of more items in between points of adaptiveness.

Similar to the initial testlet, determining the number of items to be presented within each stage, given a fixed total test length, creates a challenging tradeoff: more items will increase reliability within stage but decrease efficiency overall (although, to reiterate, efficiency is greatly harmed if too few items are presented, and the next round of item selection is based upon a very inaccurate $\hat{\theta}$). Thus, how many items to include at each stage represents an empirical question to be addressed (Luecht et al., 2006). I propose testing both two item stages and three item stages because in most employee selection contexts, efficiency is highly valued (i.e., a larger number of within-stage items would start to approach the desired total test length). In summary, the present

study compares three different item presentation strategies (after initial item presentation): fully adaptive, two item stages, and three item stages.

Test Length

Termination rules for CAT may fall into one of two categories: fixed length or variable length. Variable length CATs may use a variety of indices to terminate, including a required minimum standard error of measurement, the maximum information any remaining item can provide given $\hat{\theta}$, convergence (i.e., a minimum change in consecutive θ estimates), or any combination of the above (Babcock & Weiss, 2012). Variable length CATs are advantageous from a pure measurement perspective in that they are as efficient as possible for each examinee. Nonetheless, fixed length tests may be used in operational contexts due to practical considerations, including ease of test creation/administration and examinee fairness perceptions (Babcock & Weiss, 2012; Kantrowitz et al., 2011; Stark et al., 2012). In the present study, fixed length termination rules are utilized, mimicking most real-world applications of CAT for employee selection. The two test lengths to be tested—6 items and 12 items—were chosen to represent the expected lower boundary of test length that could still be expected to provide reasonably reliable estimates of conscientiousness.

CHAPTER 2

METHOD

Simulation Data

Participants. Data for the simulations were drawn from the survey responses of 1,724 Amazon Mechanical Turk workers who were originally recruited for a separate validation study for a measure of conscientiousness. Responses were removed if they showed evidence of insufficient effort responding (Curran, 2016). Of the original 1,768 respondents, 36 were removed for completing the survey too rapidly, seven were removed for selecting the same response option for all items, and one was removed for excessive missing data. Participants were 59.18% female and 76.63% White; table 3 shows the complete demographics of the final sample.

Item pool creation. Model calibration was conducted using the *mirt* package in R (Chalmers, 2012). Based on previous recommendations for the GGUM, item parameters were calibrated using marginal maximum likelihood (MML), and person parameters were estimated using expected a posteriori (EAP) estimation (Roberts, Donoghue, & Laughlin, 2002). Because convergence is a known issue with parameter-heavy models such as GGUM (Huang & Mead, 2014), the convergence tolerance for calibration was set to .001, a criterion used in past GGUM studies (e.g., Roberts et al., 2002; Speer, Robie, & Christiansen, 2016).

The 180 conscientiousness items used in the present study were originally written from an unfolding perspective to cover the entire range of θ. An example of an item with a negative location is “I put little time and effort into my work”. An example of a neutral/moderate item is “I am comfortable with achieving the same as the average person in life”. An example of an item

Table 3

Participant Demographic Information

Variable	N	%
Gender		
Female	1019	59.18
Male	703	40.82
Ethnicity		
White (non-Hispanic/Latino)	1315	76.63
African American	94	5.48
Asian/Pacific Islander	98	5.71
Hispanic/Latino	110	6.41
Native American/American Indian	9	.52
Other	23	1.34
Multiple (non-Hispanic/Latino)	67	3.90
Employed		
Yes	1316	76.47
No	405	23.53
Education		
Less than high school	12	.70
High school/GED	170	9.88
Some college/Associate's	720	41.86
Bachelor's	574	33.37
Professional (Master's, PhD, JD, MD)	244	14.19
Age	*Mean = 34.58, SD = 12.44*	

with a positive location is “I give 100 percent effort for everything that I do”. These items were raw and not previously validated. Items that exhibit extreme location parameters or very small discrimination parameters are indicative of poor model fit, suggest a violation of unidimensionality, and may bias the estimation of other parameters in the model (Carter & Zickar, 2011). Thus, the final item pool was created in an iterative process. After calibrating the model on all 180 items, any items with extreme location parameters (generally, less than -5 or greater than +5) or very poor discrimination (i.e., less than .10) were considered for removal. After removing problematic items for a given iteration, the model was re-calibrated and all new item parameters were re-evaluated. This process continued until all items demonstrated minimally acceptable parameters (i.e., all discrimination parameters greater than .10 and location parameters between -5 to 5), suggesting that the item pool was sufficiently unidimensional and operated as expected. In total, this process resulted in paring the original 180-item pool down to 111 items over 12 iterative calibrations.

Estimated parameters and item-data fit for the final item pool are listed in Table 4. The $S\text{-}X^2$ fit index is distributed as chi-square and serves as an indication of the degree to which empirical item responses and model-predicted item responses diverge (Orlando & Thissen, 2000). Although the $S\text{-}X^2$ was statistically significant for 54 out of the 111 items, the RMSEA’s were generally small (M = .009, SD = .004), suggesting minimal deviation between empirical and model-predicted responses. In the present study, decisions around item removal due to suboptimal parameters were weighed against the need for retaining as many items as possible, particularly in the middle range of the trait spectrum. Generally speaking, the larger the item pool from which a CAT can pull, the more efficient it will be (Flaugher, 2000). Because no item

Table 4

Final 111-Item Pool Parameters and Item-Data Fit

	Parameters		Fit				Parameters		Fit				Parameters		Fit				Parameters		Fit		
Item	α	δ	$S\text{-}X^2$	df	RMSEA	Item	α	δ	$S\text{-}X^2$	df	RMSEA	Item	α	δ	$S\text{-}X^2$	df	RMSEA	Item	α	δ	$S\text{-}X^2$	df	RMSEA
C1_2	.27	-.65	512.70**	416.4	.012	C2_24	.44	1.08	488.16	450.6	.006	C4_13	.27	-.45	456.77**	371.6	.011	C6_9	.50	-.13	504.17**	423.0	.011
C1_3	1.44	.97	423.01***	322.2	.013	C2_26	.79	.51	335.11	330.0	.003	C4_14	.68	.95	466.35	416.6	.008	C6_11	.67	.94	468.62	419.2	.008
C1_5	1.42	.91	477.42***	367.2	.013	C2_28	.84	1.04	458.20	425.2	.007	C4_15	1.21	.78	353.83	357.4	.001	C6_13	.52	1.15	485.04*	428.2	.009
C1_7	2.23	1.04	449.76***	289.8	.018	C2_29	.84	.77	346.81	359.0	.001	C4_16	.70	.89	439.06	416.0	.006	C6_16	1.18	1.16	490.50**	402.4	.011
C1_9	.87	1.06	510.74***	398.4	.013	C2_30	.56	1.05	468.16	419.0	.008	C4_17	.54	.38	383.66	357.2	.006	C6_17	.47	-.66	472.15***	353.4	.014
C1_10	.80	.93	522.86**	423.0	.012	C3_1	.24	-.77	510.98	460.2	.008	C4_18	.92	.87	442.68	411.0	.007	C6_19	.48	-.76	520.91***	382.4	.015
C1_11	.56	-.02	488.49**	399.0	.011	C3_2	1.74	1.01	461.88***	353.4	.013	C4_20	.98	.53	378.09	347.2	.007	C6_20	.92	1.08	513.56**	405.8	.012
C1_12	.45	.39	397.59	379.0	.005	C3_3	.93	.80	366.57*	319.4	.009	C4_21	.99	.42	346.99	321.6	.007	C6_21	.43	-.20	500.32**	411.0	.011
C1_13	1.66	1.01	499.38***	357.6	.015	C3_5	2.07	.99	381.62***	275.8	.015	C4_23	.29	-.33	517.67	475.6	.007	C6_22	.68	.21	477.26*	416.8	.009
C1_15	.71	.39	353.19	338.4	.004	C3_6	.39	-.68	561.78**	457.6	.011	C4_24	1.28	.90	421.08	385.8	.007	C6_23	.39	-.94	506.59***	392.4	.013
C1_17	.60	.58	396.04	386.6	.004	C3_7	2.17	1.02	486.36***	337.6	.016	C4_26	.40	-.07	547.92**	440.2	.012	C6_25	.27	-.61	560.66*	472.4	.010
C1_19	1.67	.97	364.00**	287.6	.012	C3_8	.42	-.59	516.18**	423.6	.011	C4_27	.58	.31	459.35	426.2	.006	C6_26	2.07	1.08	467.35***	348.6	.014
C1_21	.46	-.92	351.53*	295.0	.010	C3_9	.16	-.53	486.46	461.8	.005	C4_30	.31	.23	448.22	416.8	.006	C6_27	.29	-.37	505.50	454.2	.008
C1_22	1.75	1.04	425.60**	335.6	.012	C3_11	.14	-1.39	540.87	494.4	.007	C5_1	1.38	1.05	367.83*	309.6	.010	C6_28	.49	-.61	491.69***	364.8	.014
C1_24	.30	-.17	477.14	421.6	.009	C3_12	.43	-.86	542.71***	425.4	.013	C5_3	.19	.12	539.52	521.6	.004	C6_29	1.37	1.04	540.92***	392.4	.015
C1_25	1.76	.99	395.79**	300.4	.014	C3_14	1.21	.91	485.98**	394.8	.012	C5_4	.31	-.11	473.75	420.8	.008						
C1_26	.63	.55	411.88	402.4	.004	C3_20	1.91	1.02	522.68***	357.2	.016	C5_8	.80	.89	418.84	370.2	.009						
C1_27	.24	-.64	527.87**	432.8	.011	C3_21	.26	-.90	570.95**	470.2	.011	C5_12	.61	.48	379.82	395.0	.001						
C2_1	.21	-1.03	466.78	442.4	.005	C3_22	.58	.24	419.09*	370.2	.009	C5_13	.39	.10	447.26*	380.6	.010						
C2_2	.27	-.05	459.35	437.2	.005	C3_24	2.37	1.08	522.70***	355.4	.017	C5_14	.28	-.15	478.97*	406.8	.010						
C2_3	.25	-1.16	580.68*	509.6	.009	C3_27	.85	.87	453.92	408.2	.008	C5_15	.74	.88	402.99	393.8	.003						
C2_5	.43	.45	393.21	382.2	.004	C3_29	.81	.85	402.88	387.0	.004	C5_19	.25	-.42	501.45	451.0	.008						
C2_6	.34	.40	436.77	431.4	.002	C3_30	2.36	1.07	512.60***	358.6	.016	C5_20	.75	.44	363.76	332.0	.007						
C2_7	1.09	.99	419.10	392.0	.006	C4_1	.51	.41	399.60	388.6	.003	C5_21	.83	.82	407.77	369.2	.008						
C2_8	.41	.09	477.64	434.6	.007	C4_2	.18	.11	510.02	481.4	.005	C5_22	.53	.41	441.38	394.2	.008						
C2_9	.83	.66	421.79	373.6	.009	C4_3	.33	-.51	516.99**	426.0	.011	C5_25	.30	.23	448.40	417.6	.005						
C2_13	1.07	.98	413.99	395.2	.005	C4_4	1.37	1.06	407.27*	353.4	.009	C5_30	.72	.81	427.43	406.6	.005						
C2_14	.91	.93	418.63	400.8	.005	C4_5	.24	-.44	453.40	438.6	.004	C6_1	.79	1.03	457.49*	385.0	.010						
C2_16	.67	1.06	494.86*	427.6	.009	C4_7	.35	-.05	465.98*	404.2	.009	C6_3	1.63	1.04	479.58***	371.2	.013						
C2_19	.51	.27	438.37	420.4	.004	C4_9	.38	.30	421.49	388.6	.006	C6_4	.81	.53	436.54	397.8	.007						
C2_21	.95	1.21	542.79***	421.8	.013	C4_11	.29	.04	479.62	417.6	.009	C6_7	.44	-.99	532.19***	376.6	.015						
C2_23	.26	-.48	515.67	477.0	.007	C4_12	.14	-1.03	389.25	390.4	.002	C6_8	.44	-.41	393.40*	327.2	.011						

Note. RMSEA refers to the Root Mean Square Error of Approximation of the $S\text{-}X^2$ index.
* $p < .05$. ** $p < .01$. *** $p < .001$.

demonstrated notably poor fit, all 111 items were retained. The test information function and SEM for the final item pool can be seen in Figure 3.

Convergent validity with an alternate measure of conscientiousness was additionally used to corroborate the validity of the final item pool. GGUM latent trait scores correlated at $r = .68$ with examinees' sum scores on a 20-item, dominance-based measure of conscientiousness created from the freely available International Personality Item Pool (DeYoung, Quilty, & Peterson, 2007; Goldberg, 1992). This moderately high correlation aligns with expectations due to both measures tapping the same construct but utilizing different items and response models.

Procedure

Real data simulation approaches are common in the CAT literature (Hol et al., 2008; Kamakura & Balasubramanian, 1989; Makransky et al., 2013; Reise & Henson, 2000) and were used in the current study. This approach allows for higher external validity compared with fully simulated data while still leveraging a "known" θ upon which to compare CAT results (Hol et al., 2008). Prior to simulation, each individual's known θ was calculated as the EAP estimate based upon responses to all items in the finalized 111-item pool. Within the simulation, $\hat{\theta}$ is calculated only for those items presented and is then compared with θ.

Static conditions. The 6-item and 12-item static tests were comprised of the 6 and 12 items (respectively) in the item pool that had the highest α parameters.

Adaptive conditions. There are currently no widely available programs for conducting CAT simulations using ideal point models. Thus, functions from the *mirt* package were combined with new code in R (see Appendix for the R function developed in the present study). All simulated tests followed the same general process:

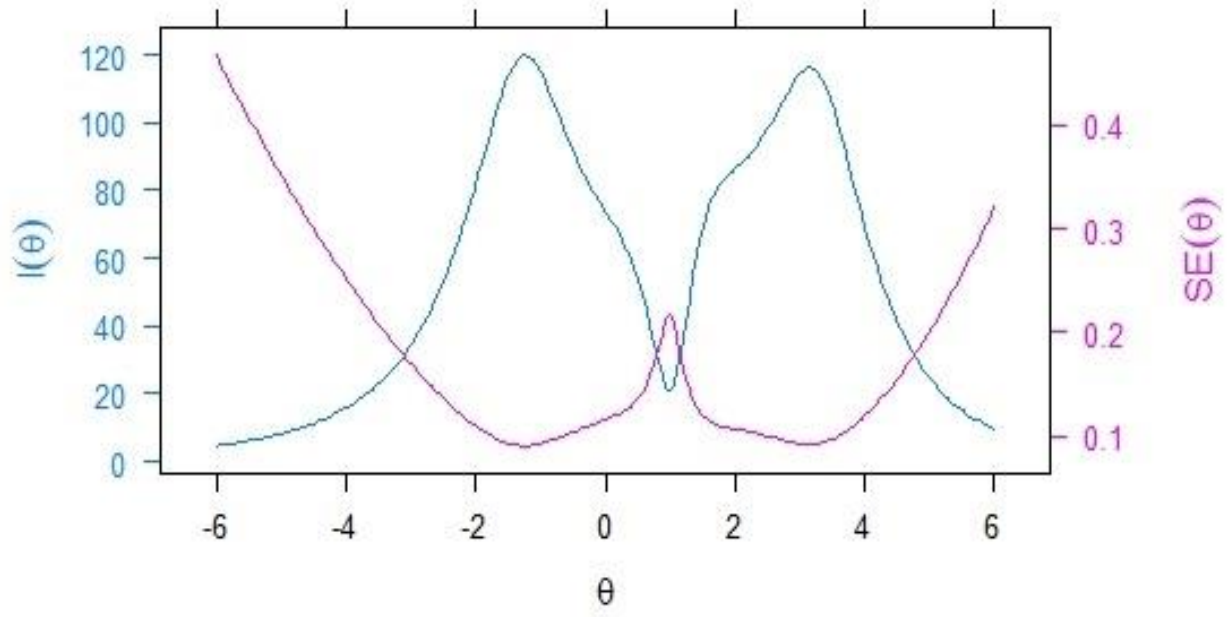


Figure 3. Test information and standard error of measurement for the final 111-item pool.

1. Present initial item/item testlet. Because all individuals are assumed to have the same θ prior to item administration (i.e., θ = 0), the same initial item(s) were presented to all examinees within a given condition.
2. Estimate $\hat{\theta}$ using EAP estimation, based upon examinee's actual responses to all item(s) presented by the CAT thus far. (In the event that an item presented by the CAT did not have a corresponding participant response, the response was simulated based on θ.)
3. Pulling from items in the pool that have not yet been presented, present the item(s) that maximize Fisher's information criterion at $\hat{\theta}$.
4. Repeat steps 2-3 until target test length has been reached.
5. Estimate final $\hat{\theta}$.

Study Outcomes

Outcomes of the present study were classified into two categories: general measurement outcomes and employee selection accuracy. All 36 conditions were compared independently on the accuracy of employee selection decisions. It should be noted here that general measurement outcomes do not vary as a function of the cut-score because they describe the entire distribution, irrespective of where this cut-score is placed. Thus, only 12 sets of general measurement outcomes are presented, collapsing across cut-scores for otherwise identical test conditions.

General measurement outcomes. General measurement outcomes upon which simulation conditions were compared include root mean square error (RMSE; see equation 2), bias (see equation 3), and accuracy of θ recovery ($R_{\theta\hat{\theta}}$). These statistics align with suggested indexes of robustness for examinations of measurement accuracy (de Ayala, 1995).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \quad (2)$$

$$Bias = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)}{n} \quad (3)$$

Employee selection accuracy. The present study adopted a select-out strategy for employee selection. From a testing fairness perspective, false negatives—those classified as failing to meet the cut-score despite being above it in reality—represent the biggest liability. Test *sensitivity* is a popular metric for evaluating the efficacy of clinical diagnostic tools (e.g., Altman & Bland, 1994). Defined as (True Positives)/(True Positives + False Negatives), a perfect score of one indicates that all individuals in the referent category were classified as such. Sensitivity is a useful metric because it is easily interpretable and readily applied to the present study. However, it is misleading to compare sensitivity across different cut-score conditions. As the cut-score is placed progressively farther away from the midpoint of the distribution, the ratio of true positives to false negatives grows larger. Were the cut-score placed in an extreme enough location (for instance, $\theta = -10$), sensitivity would be perfect, but only because the density of the distribution at that point is so low. Thus, the metric used in the present study is a normalized version of sensitivity ($\mathcal{A}_{\text{norm}}$), which corrects for the proportion of individuals above the cut-score (see Carter et al., in press; Haslbeck & Waldorp, 2017). Specifically, the $\mathcal{A}_{\text{norm}}$ statistic was calculated using the following formula:

$$\mathcal{A}_{norm} = \frac{sensitivity - \pi}{1 - \pi}$$

$$= \frac{[true\ positives/(true\ positives + false\ negatives)] - \pi}{1 - \pi}, \quad (4)$$

where π = the known proportion of individuals falling above the cut-score.

$\mathcal{A}_{\text{norm}}$ simplifies to the maximal value of one if and only if all individuals who surpassed the cut-score were classified as such. Holding all else constant, as π increases, $\mathcal{A}_{\text{norm}}$ decreases. A less extreme cut-score is rewarded for yielding the same accuracy as a more extreme cut-score because the density of the distribution is greater at a less extreme cut-score. Thus, comparable accuracy is harder to achieve because there is a greater raw number of individuals available nearby to be misclassified.

CHAPTER 3

RESULTS

The CAT characteristics of interest in the present study included total test length, the number of initial pre-adaptive items presented, and the use of multistage or sequential testing. I first present results for general measurement outcomes (i.e., bias, RMSE, and $R_{\theta\hat{\theta}}$), comparing across different conditions. Then I present results of the analyses for employee selection accuracy, which incorporates the additional use of cut-score as a predictor. Within each section, comparisons are made first between static and adaptive conditions as a whole. This is followed by comparisons across different CAT conditions. By collapsing across like conditions, I examine the marginal mean differences for each CAT characteristic in turn, comparable to interpreting main effects in an ANOVA model.[2] Where appropriate, I also draw conclusions about interactions of these characteristics.

General Measurement Outcomes

Results when comparing all conditions on bias, RMSE, and $R_{\theta\hat{\theta}}$ can be seen in Table 5. Median SEM is also included in this table as an indicator of reliability within each condition. Table 6 presents the same results aggregated by each CAT characteristic in turn.

Bias. Across all conditions, there was a small degree of systematic negative bias such that all tests tended to slightly underestimate θ by approximately .11 units. Bias was slightly less

[2] The partially-crossed design used in the present study necessitated that the single-item start conditions *not* be included in the aggregate statistics for sequential and multi-stage conditions. If the single-item start conditions were included, the sequential testing aggregate statistics would be a function of all three start conditions, but the multistage testing aggregate statistics would only be a function of the two- and three-item start conditions. The single-item start, sequential conditions are a special case that are technically both sequential and multistage. Although labeled as sequential for ease of interpretation, these conditions also meet the rule for multistage as defined in the present study: the number of items presented in the initial testlet is equal to the number of items presented in each stage (i.e., one).

Table 5

General Measurement Outcomes for each Study Condition

Test Length/Condition	Median SE	Bias	RMSE	$R_{\theta\hat{\theta}}$
6 items				
Static	.604	-.124	.731	.676
Single item start, sequential	.600	-.113	.728	.676
Two-item start, sequential	.370	-.102	.654	.750
Two-item start, multistage	.369	-.093	.651	.752
Three-item start, sequential	.368	-.101	.653	.751
Three-item start, multistage	.366	-.086	.651	.750
12 items				
Static	.586	-.134	.725	.687
Single item start, sequential	.578	-.120	.710	.699
Two-item start, sequential	.323	-.113	.628	.776
Two-item start, multistage	.322	-.109	.615	.786
Three-item start, sequential	.324	-.114	.631	.773
Three-item start, multistage	.318	-.106	.619	.783

Note. Median SE = median standard error of measurement of $\hat{\theta}$.

Table 6

General Measurement Outcomes Across Study Conditions

Test Characteristic	*k*	Bias *M*	Bias *SD*	RMSE *M*	RMSE *SD*	$R_{\theta\hat{\theta}}$ *M*	$R_{\theta\hat{\theta}}$ *SD*
Static	2	-.129	.005	.728	.003	.681	.005
Adaptive	10	-.106	.010	.654	.035	.750	.034
6-item length	6	-.103	.012	.678	.036	.726	.035
12-item length	6	-.116	.009	.655	.045	.751	.041
Single-item start	2	-.116	.004	.719	.009	.687	.011
2-item start	4	-.104	.008	.637	.016	.766	.016
3-item start	4	-.102	.010	.639	.014	.764	.014
Sequential	4	-.108	.006	.642	.012	.763	.012
Multistage	4	-.099	.009	.634	.017	.768	.017

Note. *k* = number of conditions with specified characteristic. *M* = mean of all included conditions. *SD* = standard deviation of all included conditions.

severe within all ten adaptive conditions as a group compared with the two static conditions as a group (see Table 6). Aggregating results across study conditions by test length, starting condition, and multistage vs. sequential testing illustrated that all three of these test characteristics had a negligible impact on bias.

RMSE and $R_{\theta\hat{\theta}}$. Examining RMSE and $R_{\theta\hat{\theta}}$ in Table 6 led to congruent conclusions; thus, both are discussed in tandem presently.[3] As a group, the ten adaptive conditions generated more precise estimates than the two static conditions. Regarding test length, 12-item tests as a group generated slightly more precise estimates than the 6-item tests as a group. However, RMSE and $R_{\theta\hat{\theta}}$ were not meaningfully different between the 12-item static test and the 6-item static test (i.e., a difference of .006 and .011, respectively). Advantages from greater test length were thus driven by the adaptive test conditions. Examining different starting conditions within adaptive conditions only, the single-item start conditions generated less precise estimates than the two-item start conditions and the three-item start conditions. However, there was no meaningful difference in the RMSE and $R_{\theta\hat{\theta}}$ between the two-item start conditions and three-item start conditions Examining multistage conditions as a whole and sequential conditions as a whole revealed that the two testing strategies did not differ meaningfully on RMSE or $R_{\theta\hat{\theta}}$.

In sum, adaptive tests were more precise than static tests, 12-item tests were slightly more precise than 6-item tests, two- and three-item start conditions were more precise than single-item start conditions, and the multistage conditions did not differ from sequential. Notably, starting condition had a larger effect on precision than did any other CAT characteristics. Relatedly, all

[3] RMSE and $R_{\theta\hat{\theta}}$ are closely related but not perfectly so. In general, the greater the error in estimating theta-hat, the more that theta-hat rank ordering will diverge from theta and lead to lower correlations between theta and theta-hat. Theoretically, however, large RMSE could be due to systematic bias and a large $R_{\theta\hat{\theta}}$ could still be observed.

CATs outperformed their same-length static test counterpart except for the single-item start CAT.

Employee Selection Accuracy

$\mathcal{A}_{\text{norm}}$ for each of the 36 study conditions can be seen in Table 7. Table 8 presents the same results aggregated by each CAT characteristic in turn. Unexpectedly, the static conditions as a group demonstrated higher normalized sensitivity than did the adaptive conditions as a group. Regarding test length, 12-item tests as a group demonstrated higher normalized sensitivity than the 6-item tests as a group. Opposite of the findings for general measurement outcomes, the single-item start conditions demonstrated higher normalized sensitivity than the two-item start conditions and three-item start conditions. The two-item start conditions and three-item start conditions did not differ from each other meaningfully. Multistage and sequential testing conditions did not differ meaningfully. Finally, moving from the 10$^{\text{th}}$ to 20$^{\text{th}}$ to 30$^{\text{th}}$ percentile cut-scores resulted in progressively higher normalized sensitivity. There were a greater number of raw false negatives in the 30$^{\text{th}}$ percentile (M = 70.25) compared with the 20$^{\text{th}}$ percentile (M = 61.75) and the 10$^{\text{th}}$ percentile (M = 42.17) cut-score conditions. After normalization of sensitivity, however, the 30$^{\text{th}}$ percentile provided the most value in correct identifications above and beyond simple guessing.

In sum, static tests were more accurate in employee selection than static tests, 12-item tests were more accurate than 6-item tests, single-item start conditions were more accurate than two- and three-item start conditions, and the multistage conditions did not differ from sequential. Additionally, cut-score interacted with test length. As test length increased, the impact of the cut-score on normalized sensitivity weakened.

Table 7

Normalized Employee Selection Accuracy ($\mathcal{A}_{norm}$) for each Study Condition

Test Length/ Condition	10th percentile (θ = -.947)	20th percentile (θ = -.740)	30th percentile (θ = -.589)
6 items			
Static	.770	.830	.843
Single item start, sequential	.790	.775	.817
Two-item start, sequential	.530	.675	.757
Two-item start, multistage	.570	.685	.737
Three-item start, sequential	.530	.680	.753
Three-item start, multistage	.740	.705	.757
12 items			
Static	.840	.890	.873
Single item start, sequential	.850	.885	.877
Two-item start, sequential	.770	.800	.813
Two-item start, multistage	.790	.795	.803
Three-item start, sequential	.780	.805	.820
Three-item start, multistage	.770	.795	.817

Note. Percentile headings refer to the placement of the cut-score.

Table 8

Normalized Employee Selection Accuracy ($\mathcal{A}_{norm}$) Across Study Conditions

Test Characteristic	*k*	$\mathcal{A}_{\text{norm}}$ *M*	$\mathcal{A}_{\text{norm}}$ *SD*
Static	6	.841	.038
Adaptive	30	.756	.087
6-item length	18	.719	.092
12-item length	18	.821	.038
Single-item start	6	.832	.042
2-item start	12	.727	.090
3-item start	12	.746	.077
Sequential	12	.726	.098
Multistage	12	.747	.066
10th percentile	12	.728	.111
20th percentile	12	.777	.072
30th percentile	12	.806	.045

Note. *k* = number of conditions with specified characteristic. *M* = mean of all included conditions. *SD* = standard deviation of all included conditions.

CHAPTER 4

DISCUSSION

Implications for Measurement Theory and Practice

The primary aim of the present study was to examine the feasibility of creating an efficient, accurate CAT using relatively few ideal point items to measure conscientiousness. CAT has recently gained popularity as a method by which to increase testing efficiency, and ideal point models are now generally recognized as superior representations of the item response process (LaPalme et al., 2018). Thus, it is important to identify if the significant costs associated with creating ideal point CATs is justified by the value beyond traditional static tests. In addition to comparing ideal point static and adaptive tests, I sought to identify the conditions in starting the CAT, presenting items, terminating the CAT, and determining cut-score placement that result in optimal general measurement outcomes and accuracy in making dichotomous selection decisions. I review the implications from the present study for each of these test characteristics in turn, highlighting where general measurement and selection accuracy outcomes converge and where they diverge.

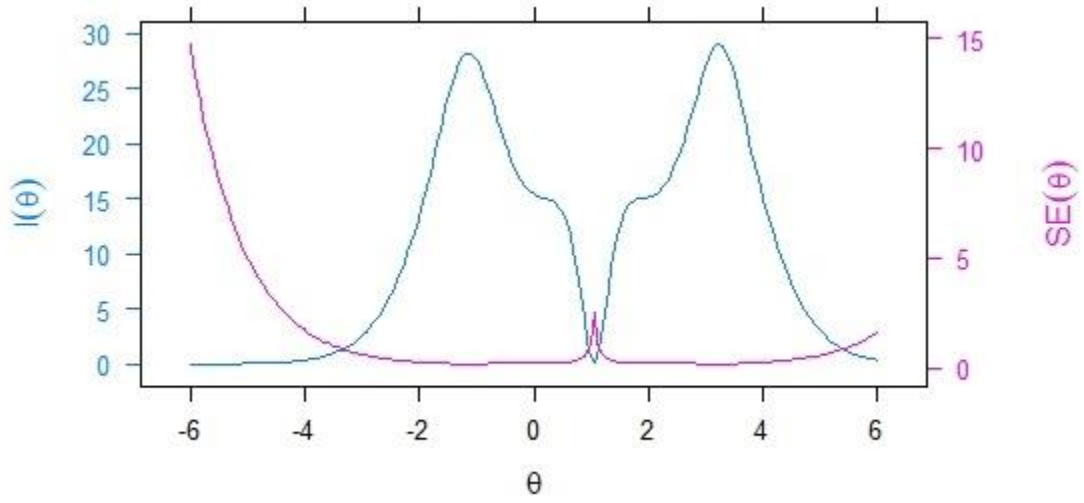
CAT versus static ideal point testing. Generally, precise measurement requires a greater number of ideal point than dominance-based items (Dalal et al., 2010; Williamson et al., 2017), which I posited might limit CAT's utility if the number of ideal point items necessary for precision approaches the length of a traditional static test. Past personality research demonstrates that CAT is useful for dominance-based, single-stimulus items (Hol et al., 2008; Makransky et al., 2013) and for ideal point, forced-choice items (Drasgow et al., 2012; Houston et al., 2006;

Schneider et al., 2009). The present study extended this evidence to the ideal point, single-stimulus item format. When compared to a same-length static test, CAT conditions in the present study performed better in terms of general measurement. As further support for the improvement in precision when utilizing adaptive instead of static tests, all but one of the 6-item CATs yielded better $R_{\theta\hat{\theta}}$ than the 12-item static test. This mirrors past research that CAT can effectively halve the number of items used in assessment (van der Linden & Glas, 2010).

However, the value of an ideal point CAT depends on the purposes of testing. When the goal of testing was to provide accurate estimates across the entire theta distribution, the value of CAT compared with static ideal point testing was relatively high. However, when the aim in the present study was to mimic a real-world selection context by dichotomously selecting individuals above a cut-score, the static conditions actually performed better than CAT. The reason for this finding can be decomposed into a study-specific component and a more generalizable component. The performance of any static test in a selection context will be highly contingent upon where the test provides statistical information and where the cut-score is placed. In the present study, it happened that the most informative items in the pool provided a great deal of information in the theta range where cut-scores were placed (i.e., $-1 < \theta < 0$). As a result, there was little for the CAT to improve upon: the static tests were already well-positioned to precisely discriminate individuals in that theta range (see Figure 4). Thus, the deck was stacked against the adaptive conditions, so to speak, as the present situation represents a best-case scenario for static test use. At a more general level, this finding illustrates that adaptive does not always provide value.

Test developers must take into account the amount of information provided by static items near the cut-score when considering if CAT is necessary for superior outcomes. Were the

(a) 6-item static test



(b) 12-item static test

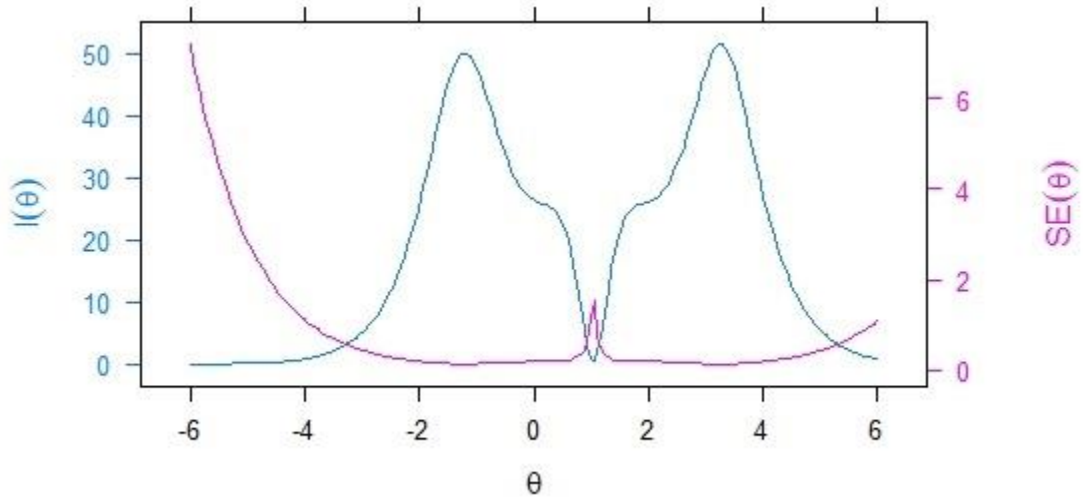


Figure 4. Test information and standard error of measurement for the final (a) 6-item static test and (b) 12-item static test.

cut-scores in the present study placed where the static tests provided minimal information (e.g., θ = 1), the adaptive conditions would have likely performed comparatively better than the static. In total, results are suggestive that CAT will provide more value when measured against general measurement outcomes than dichotomous selection decisions. Ultimately, comparative performance between static and adaptive on selection accuracy is as much a function of the sufficiency of the static test as it is about the efficiency of the adaptive test.

Initial item selection. Ultimately, the complex parameterization of ideal point items proved not to be an insurmountable obstacle in developing an efficient CAT in the present study. However, not all CAT conditions performed equally well. This was most apparent when examining the starting condition manipulation. Across all outcomes, the single-item start CAT condition generally performed similarly to the static condition, whereas all other CAT conditions hung together. In other words, when examining general measurement outcomes, evidence was found to support the theta indeterminacy issue. Beginning adaptation on only a single item response proved to be minimally useful. This is likely because ideal point models cannot differentiate lack of endorsement of an item due to being above or below that item without additional information. In turn, the value of CAT over static may be largely negated when initial $\hat{\theta}$ estimates are unstable. Notably, there was no difference in performance between the two-item start and three-item start conditions, suggesting that responses to as few as two items provide CAT sufficiently reliable information upon which to coarsely triangulate $\hat{\theta}$. Conversely, the inclusion of a 3rd item did not result in a meaningful reduction in performance. Both are viable starting conditions for CAT.

The finding that the single-item start conditions performed better than two- and three-item start conditions on employee selection accuracy was unanticipated. This finding is

particularly surprising considering that the single-item start conditions exhibited slightly more negative systematic bias and slightly more overall error than the others. In turn, one would expect that more individuals would drop below the cut-score and reduce normalized sensitivity in the single-start conditions. It is possible that this finding resulted from the idiosyncrasies of the item pool used, a point discussed further below.

Multistage versus sequential testing. The present study aimed to identify the ideal balance between the number of points of adaptiveness (efficiency) and the accuracy of the estimates upon which adaptiveness is based (stability). Multistage and sequential testing were tested as two competing ongoing item presentation strategies. Theoretically, sequential testing represents the more efficient testing strategy because it leverages each new piece of information—each item response—as it is provided, instantly providing feedback to CAT regarding the appropriate direction to head (van der Linden & Glas, 2010). However, to the extent that θ estimation from very few ideal point items is unstable, multistage testing provides a potential solution by collecting more information prior to each point of estimation (Wang et al., 2016). Despite the theoretical differences in the two approaches, multistage and sequential testing were equally effective strategies for both general measurement and employee selection. There seems to be a degree of equifinality between the two approaches: multistage may be slower and steadier throughout the test, whereas sequential is initially more unstable but gains precision at the end.

Because multistage and sequential testing are equally effective from a measurement perspective, practitioners may wish to take additional outcomes into consideration when deciding between the two. Multistage testing generally has fewer computational demands and elicits more

positive applicant reactions than sequential testing (Stark & Chernyshenko, 2006), although additional research in this area is warranted.

Test length. As anticipated, the 12-item test length conditions outperformed the 6-item test length conditions across all outcomes. More telling than the relative performance of the 6-item and 12-item conditions overall is the moderating effect that length exerted on other characteristics. $R_{\theta\hat{\theta}}$ did not increase meaningfully from the 6-item static test to the 12-item static test. In the present case, the 12-item test simply provided additional items with similar locations to the 6-item test, so the contribution was minimal. Adding items with locations similar to those already in a test is particularly unhelpful in ideal point testing compared with dominance-based testing due to the greater need for triangulating θ. Differences in reliability between the 6-item and 12-item adaptive tests were more marked than for static, suggesting that the addition of more items gave the CAT more time to build momentum, so to speak, in matching item location to person theta. From an absolute perspective, 6-item CATs yielded fair reliability, which may be acceptable in certain contexts, such as research. However, even the 12-item CATs did not obtain a reliability of .80, suggesting more items may be needed for applied, high-stakes testing (Nunnally, 1978).

Employee selection cut-score. When organizational decision-makers utilize personality assessments from a select-out perspective, a decision must ultimately be made regarding where the cut-score should be placed. In settings with large hiring needs or a focus on test fairness, cut-scores may be placed low so as to eliminate only those who are truly unqualified. However, evidence from the present study suggests that more extreme cut-score placement may result in poorer selection accuracy. Results in the present study relied on a normalized sensitivity metric, which does not directly address the utility of the selection criterion to an organization per se. The

placement of cut-scores must be carefully informed by desired yield ratios and job analytic data that identifies the level of the desired trait necessary for effective job performance. Nonetheless, cut-score placement should also be carefully considered as an important contributor to decision accuracy and, by extension, test fairness and legal defensibility.

Limitations

The use of real-data simulations in the present study represents a middle ground between testing using fully simulated data and testing in an operational setting. This provides the advantage of leveraging known information upon which to judge the performance of the CAT while capturing response patterns that occur in the real world. However, this design has several drawbacks, as discussed next.

Non-manipulability of item parameters. Characteristics of the item pool may impact the efficiency of CAT and may limit the generalizability of the present study. The real-data simulation approach used presently estimates item parameters rather than manipulating them—as would be done in a full simulation. Therefore, conclusions may not completely generalize to other item pools. A uniform distribution is desirable for CAT so that it can readily match items to any estimated level of θ. However, the TIF in Figure 3 demonstrates that the pool used in the present study did not yield a uniform distribution of item locations. Many items had a location near $\delta = 1$. Although items were written to tap the entire distribution, writing items that effectively tap the middle of the distribution can be a challenge (Cao, Drasgow, & Cho, 2014; Huang & Mead, 2014), which may have contributed to the non-optimal item pool.

This may have had a bearing on the unexpected presence of bias within all conditions. Bias in the present study may be reflective of the idiosyncratic nature of the item parameters as opposed to a function of test algorithms. Items with high discrimination parameters were favored

for static and adaptive conditions alike, but just so happened to hover disproportionately around a location of $\delta = 1$. Theta estimation is optimal across the entire distribution when item locations are evenly dispersed across the distribution. Thus, a fairer test of differences across conditions might entail developing a more uniform item location distribution. As an additional piece of evidence, bias did not uniformly impact θ estimates. Exploratory analyses revealed that negative bias was problematic particularly for individuals with $\theta > 1$. Figure 5 shows theta plotted against bias for all examinees on the 6 item—2 item start—multistage condition (note, however, that this pattern is representative of all conditions in the present study, static and adaptive alike). This may have occurred because the large majority of items around $\delta = 1$ did not support precise estimation for individuals near that location.

The characteristics of the item pool also may have played a role in the finding that more extreme cut-scores yield poorer employee selection accuracy. The information at any given point (e.g., a cut-score) in the theta distribution is inextricably related to the features of the underlying item pool. In this way, the finding that the 30th percentile cut-score yields optimal selection accuracy may not be generalizable to all situations. Rather, certain cut-scores will be better positioned to yield high accuracy depending upon the characteristics of the item pool used.

Ecological validity concerns. The present study draws on the assumption that item responses generalize to real-world contexts. Nested within this assumption are two related assumptions: (1) the order and number of items within test conditions would not meaningfully alter responses to the original 180-item measure and (2) the research context in which data were collected generalizes to actual applicant behavior.

Regarding the first point, there is ample evidence that context in general and item order specifically impacts interpretation and evaluation of items (Schwarz, 1999). Previous research

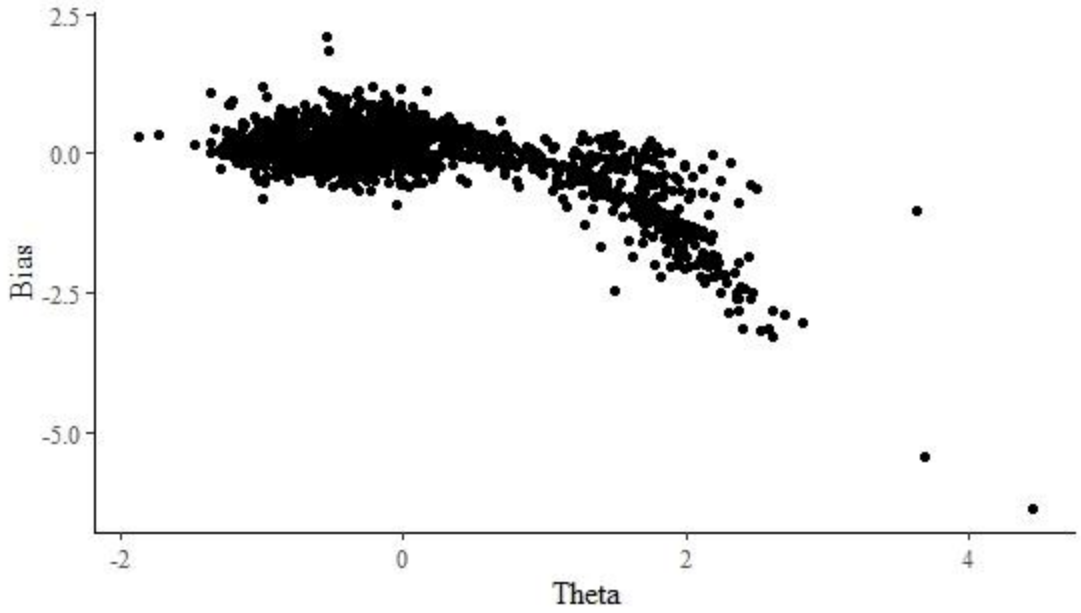


Figure 5. True theta plotted against bias ($\hat{\theta}$ minus θ) for the 6-item—2 item start—multistage adaptive condition.

suggests that examinee perceptions of how they are performing on a CAT can impact affective reactions and motivation (Tonidandel, Quiñones, & Adams, 2002). Although such research has focused on tests of cognitive ability, it is plausible that applicants may attempt to infer their performance on the assessment relative to the way in which personality items change in location as they progress through a CAT. In a related vein, specific to the present study is the use of testlets. Presentation of certain items together may have unanticipated context effects (Ortner, 2008). That is, the parameters of an item may change depending on other items in the testlet. User reactions could also potentially be impacted by presenting multiple items followed by single items, as the ability to review and edit answers in a testlet may receive better user reactions (Stark & Chernyshenko, 2006). There does not exist any evidence suggesting that such context effects threaten the present study's validity systematically, but future field research is nonetheless warranted to explicitly test for these effects.

Second, the examinee responses used in the present study were collected for research purposes. A large body of research indicates that individuals in selection settings are motivated and able to distort their responses to appear more favorable, whereas individuals in research contexts answer more honestly (Mueller-Hanson et al., 2003). Because dishonest responding may impact not only mean score but also item response functions, results of the present study should be applied tentatively to operational contexts.

Future Directions

Fully simulated data. The present study established that, given a specific item pool, adaptive testing can provide an improvement upon static testing when measuring outcomes at the level of the entire distribution but may not improve employee selection decisions based on a set cut-score. Future studies should adopt an entirely simulated approach (i.e., simulating both item

responses and item parameters) in order to exert more control over the item pool than was possible in the present study. Doing so would allow for a greater variety of situations to be explicitly tested. One particularly critical question to be addressed would be the functioning of the ideal point CAT when the item pool is completely optimal (i.e., there are highly discriminating items across the entire span of θ) and theta distribution is normal. Such a situation may yield superior CAT performance compared with static when dichotomously selecting employees. Alternative distributions of item parameters and the theta distribution could also be systematically manipulated to identify how various real-world contexts might impact CAT utility. Finally, systematic manipulation of the number of items in the pool—in an absolute sense and relative to the length of the test—would also provide valuable information, given that item-writing is a costly process but that too few items can diminish the value of CAT (Flaugher, 2000).

User reactions to ideal point CAT. Literature on user reactions to CAT in operational settings is essentially non-existent (cf. Kantrowitz et al., 2011). Evidence suggests that individuals perceive ideal point items (in a static test) as less accurate and more difficult than dominance-based items. Further, such reactions appear to be driven by the belief that test administrator interpretations of responses to middle items are liable to be inaccurate due to the possibility of disagreeing from above or below an item (Harris, McMillan, & Carter, under review). It is plausible that such items being presented in isolation, as is done in a CAT, may aggravate this concern (and rightfully so if test characteristics are not carefully constructed to account for the theta indeterminacy issue). Considering that applicant reactions to selection practices impact evaluations of the organization and acceptance of job offers (Chapman,

Uggerslev, Carroll, Piasentin, & Jones, 2005; Hausknecht, Day, & Thomas, 2004), it is important to consider this outcome in addition to measurement outcomes.

Optimizing initial testlet design. The present study suggests that an initial testlet is important for effective ideal point CAT functioning. One general future question to be addressed is the relative importance of the locations of the items in the testlet compared with the discrimination parameters. The assembly strategy for the testlets presently was relatively rudimentary: the two-item testlet was comprised of one item that maximized Fisher's information above $\theta = 0$ and one item that maximized it below $\theta = 0$. Dividing the θ distribution in half at 0 is logical considering that this represents the mean of the distribution, but this still represents an arbitrary decision. There are many possible variants of this approach that remain to be tested. Systematically manipulating the locations of the two items in the testlet relative to each other, relative to the moments of the θ distribution, and relative to information provided may provide valuable insight into optimal testlet creation.

Role of non-statistical constraints. In the present study, item presentation was based purely upon statistical considerations. In operational settings, numerous other constraints are likely to factor into test assembly, particularly item exposure rates. Item overexposure is a concern in CAT due to the relatively higher probability that highly discriminating items will be presented across many examinees. Such overexposure may threaten the security of the item pool by uniformly or non-uniformly impacting examinee prior familiarity with the items. Although this is traditionally a stronger concern for ability items that possess an objectively correct answer, it may still have implications for personality assessments. Thus, future research should examine how constraining item exposure impacts the functioning of an ideal point CAT.

The nature of ideal point modeling allows for an additional possible constraint that is not applicable in dominance-based modeling. In dominance-based modeling, presenting an item with high information necessitates an item location close to $\hat{\theta}$. With ideal point modeling, the item location will necessarily be either above or below $\hat{\theta}$. As previously discussed, evidence suggests that manipulating item locations can have important implications for how examinees believe they are performing and their overall liking of an assessment (Tonidandel et al., 2002). Thus, future research may investigate incorporation of nonstatistical constraints unique to ideal point items, such as differentially weighting for presentation items above or below interim $\hat{\theta}$.

Information criteria options. Fisher's Maximum Information Criterion (MIC) represents the most common strategy for selecting the next item in a CAT and was used presently (van der Linden & Glas, 2010). There are two other item selection algorithms that may prove particularly valuable for ideal point personality CAT: Kullback-Leibler Information (Chang & Ying, 1996) and Maximum Interval Information (Van Rijn, Eggen, Hemker, & Sanders, 2016). Both of these "global information" indices differ from MIC in that their computation prioritizes presenting an item that discriminates across an interval of θ as opposed to a point on the continuum. They essentially act as a floodlight, whereas MIC acts as a spotlight. These options remain largely unexplored as they pertain to ideal point items (Makransky et al., 2013). However, global information criteria may be very useful for early ideal point item responses considering the relatively low confidence we have around early $\hat{\theta}$ estimates. No interval, no matter how wide, is likely to resolve the theta indeterminacy issue. But in combination with initial testlets, estimation may be optimized by exploring other information criteria options.

REFERENCES

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 1: Sensitivity and specificity. *BMJ, 308*(6943), 1552. doi:10.1136/bmj.308.6943.1552

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, 1-18. doi:10.7333/1212-0101001

Ben-Porath, Y. S., Slutske, W. S., & Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the MMPI. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*(1), 18-22. doi:10.1037/1040-3590.1.1.18

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135-1147. doi:10.3758/s13428-012-0217-x

Cao, M., Drasgow, F., & Cho, S. (2014). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods, 18*(2), 252-275. doi:10.1177/1094428114555993

Carter, N. T., Harris, A. M., Listyg, B., Lowery, M. R., Williamson, R. L., Conley, K. M., . . . Carter, D. R. (in press). Understanding job satisfaction in the causal attitude network model. *Journal of Applied Psychology*.

Carter, N. T., & Zickar, M. J. (2011). The influence of dimensionality on parameter estimation accuracy in the Generalized Graded Unfolding Model. *Educational and Psychological Measurement, 71*(5), 765-788. doi:10.1177/0013164410387594

Cascio, W. F., & Aguinis, H. (2011). *Applied psychology in human resource management*. Upper Saddle River, NJ: Prentice Hall.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi:10.18637/jss.v048.i06

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229. doi:10.1177/014662169602000303

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology, 90*(5), 928-944. doi:10.1037/0021-9010.90.5.928

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88-106. doi:10.1037/1040-3590.19.1.88

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

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4-19. doi:10.1016/j.jesp.2015.07.006

Dalal, D. K., Withrow, S., Gibby, R. E., & Zickar, M. J. (2010). Six questions that practitioners (might) have about ideal point response process items. *Industrial and Organizational Psychology, 3*(04), 498-501. doi:10.1111/j.1754-9434.2010.01279.x

de Ayala, R. J. (1995). The influence of dimensionality on estimation in the partial credit model. *Educational and Psychological Measurement, 55*, 407-422.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*(5), 880-896. doi:10.1037/0022-3514.93.5.880

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010a). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology, 3*(04), 465-476. doi:10.1111/j.1754-9434.2010.01273.x

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010b). Improving the Measurement of Psychological Variables: Ideal Point Models Rock! *Industrial and Organizational Psychology, 3*(04), 515-520. doi:10.1111/j.1754-9434.2010.01284.x

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., & Hulin, C. L. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions*. Fort Belvoir, Virginia: United States Army Research Institute for the Behavioral and Social Sciences.

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

Fetzer, M., Dainis, A., Lambert, S., & Meade, A. W. (2011). *Computer adaptive testing (CAT) in an employment context [White paper]*. SHL Previsor. Retrieved from http://central.shl.com

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 37-59). Mahwah, NJ: Lawrence Earlbaum Associates.

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment, 19*(1), 14-24. doi:10.1037/1040-3590.19.1.14

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42.

Harris, A. M., McMillan, J. T., & Carter, N. T. (under review). Test-taker reactions to ideal point measures of personality. *Journal of Business and Psychology*.

Haslbeck, J., & Waldorp, L. (2017). How well do network models predict observations? On the importance of predictability in network models. Retrieved from https://arxiv.org/pdf/1610.09108.pdf

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639-683.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized adaptive testing of personality traits. *Zeitschrift für Psychologie / Journal of Psychology, 216*(1), 12-21. doi:10.1027/0044-3409.216.1.12

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS)*. Millington, TN: Navy Personnel Research, Studies, and Technology Division, Bureau of Naval Personnel

Huang, J., & Mead, A. D. (2014). Effect of personality item writing on psychometric properties of ideal-point and Likert scales. *Psychological Assessment, 26*(4), 1162-1172. doi:10.1037/a0037273

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*(6), 875-925. doi:10.1037/a0033901

Kamakura, W. A., & Balasubramanian, S. K. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment, 53*(3), 502. doi:10.1207/s15327752jpa5303_8

Kantrowitz, T. M., Dawson, C. R., & Fetzer, M. S. (2011). Computer Adaptive Testing (CAT): A faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology, 26*(2), 227-232. Retrieved from http://www.jstor.org/stable/41474872

LaPalme, M., Tay, L., & Wang, W. (2018). A within-person examination of the ideal-point response process. *Psychological Assessment, 30*(5), 567-581. doi:10.1037/pas0000499

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 55.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189-202. doi:10.1207/s15324818ame1903_2

Makransky, G., Mortensen, E. L., & Glas, C. A. (2013). Improving personality facet scores with multidimensional computer adaptive testing: an illustration with the NEO PI-R. *Assessment, 20*(1), 3-13. doi:10.1177/1073191112437756

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348-355. doi:10.1037/0021-9010.88.2.348

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection & Assessment, 16*(3), 249-257. doi:10.1111/j.1468-2389.2008.00431.x

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology, 102*(3), 291-304. doi:10.1037/apl0000081

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347-364. Retrieved from http://ejournals.ebsco.com/direct.asp?ArticleID=484287570397EA325B2B

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103. doi:10.1207/S15327752JPA8102_01

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32. doi:10.1177/01466216000241001

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*(2), 192-207. doi:10.1177/01421602026002006

Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement, 25*(2), 177-196. doi:10.1177/01466210122031993

Schneider, R. J., McLellan, R. A., Kantrowitz, T. M., Houston, J. S., & Borman, W. C. (2009). Criterion-related validity of an innovative CAT-based personality measure. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93. doi:10.1037/0003-066X.54.2.93

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17*(1), 28-43. doi:10.1037/1040-3590.17.1.28

Speer, A. B., Robie, C., & Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring. *Personality and Individual Differences, 102*, 41-45. doi:10.1016/j.paid.2016.06.058

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education, 19*(3), 257-260. doi:10.1207/s15324818ame1903_6

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-

unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184-203. doi:10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*(3), 153-164. doi:10.1037/mil0000044

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items. *Organizational Research Methods, 15*(3), 463-487. doi:10.1177/1094428112444611

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286. doi:10.1037/h0070288

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*(2), 320-332. doi:10.1037/0021-9010.87.2.320

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer.

Van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2016). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*(4), 393-411. doi:10.1177/014662102237796

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-270). New York: Kluwer.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology, 57*(6), 1051-1058. doi:10.1037/0022-3514.57.6.1051

Wang, S., Lin, H., Chang, H.-H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement, 53*(1), 45-62. doi:10.1111/jedm.12100

Williamson, R. L., Castille, C. M., & Harris, A. M. (2017). *Practical guidance for developing and implementing ideal point measurement models*. Paper presented at the Panel presented at the 32nd Annual Meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

APPENDIX A

R CODE FOR CAT FUNCTION

Function argument definitions are as follows:

`mirtobject`	A calibrated GGUM model object in the *mirt* package with class 'SingleGroupClass'.
`alliteminfo`	A matrix containing item information, crossing any number of discrete theta values (rows) with items used in above `mirtobject` (columns). First column must be the theta values. Can be generated using the mirt::iteminfo() function item-by-item, then binding the results. Column names must be of format c("theta", "name of first item",…, "name of *n* item"). A greater number of rows increases computational demands in the initial creation of the matrix but will result in higher CAT precision.
`estimationmethod`	Method for theta estimation. See mirt::fscores() documentation.
`totallength`	Termination criteria for CAT (i.e., total number of items presented).
`startlength`	Number of items presented in initial testlet.
`stagelength`	Number of items presented in each stage *except* the first stage. A value of 1 (default) represents a traditional sequential CAT.

```
CATsimGGUM <- function(mirtobject, alliteminfo, estimationmethod = "EAP",
                       totallength, startlength = 1, stagelength = 1){

  #throw error for nonsensical input combinations
  if( any((startlength + stagelength) > totallength,
          !((totallength-startlength)/stagelength)%%1==0))
    stop('combination of total test length, initial testlet length, and stage
          length is not possible')

  #create imputed responses- CAREFUL, if there were any individuals with ALL
  #missing responses, fscores() will simply remove this person without a
  #marker for where this case was located
  imputedresponses <- as.data.frame(imputeMissing(mirtobject,
                      fscores(mirtobject, method = estimationmethod)))

  #create blank table to hold final results
  tempoutput <- data.frame()

  #compute number of stages in total that will be presented
  numberofstages <- ((totallength-startlength)/stagelength)+1

  ####Identify Starting Items####
  #finds the single item (column) for which theta (row) has maximum info
  #compared with all items. Does so by finding finite theta in the table
  #closest to current theta estimate (in this case, 0)
```

```
#uses all alliteminfo columns except theta

startitem1 <- names(which.max(alliteminfo[which(abs(alliteminfo$theta-
              0)==min(abs(alliteminfo$theta-0))),-1]))

#is the item that shows maximum info above doing so on its second or first
#peak? (ie what's its location?). Extract all item parameters.

itemparmsTEMP <- coef(mirtobject)[1:length(coef(mirtobject))-1]
#*#additional two steps to remove CI's for mirtobjects that have them
if(length(unlist(itemparmsTEMP[1]))>7){
simpparms <- function(x){x[1,]}
itemparmsTEMP <- lapply(itemparmsTEMP,simpparms)
}
#*#
itemparms <- data.frame(matrix(unlist(itemparmsTEMP), ncol = 7, byrow =
             TRUE))
itemparms <- cbind(data.frame(attr(itemparmsTEMP,"names"),stringsAsFactors
             = FALSE),itemparms)
colnames(itemparms) <- c("item","a","b","t1","t2","t3","t4","t5")

#find second item for two+ item testlets that will have max info on the
#other side of theta=0 from startitem1

itemsEASY <- as.vector(itemparms[which(itemparms$b<0),"item"])
itemsHARD <- as.vector(itemparms[which(itemparms$b>0),"item"])

if (itemparms[which(itemparms$item == startitem1),"b"]>=0){
  startitem2 <- names(which.max(alliteminfo[which(abs(alliteminfo$theta-
                0)==min(abs(alliteminfo$theta-0))), itemsEASY]))
} else {
  startitem2 <- names(which.max(alliteminfo[which(abs(alliteminfo$theta-
                0)==min(abs(alliteminfo$theta-0))), itemsHARD]))
}

#find the 3rd item for the 3-item testlet (third most informative item in
#the assessment at theta = 0)

startitem3 <- names(which.max(alliteminfo[which(abs(alliteminfo$theta-
              0)==min(abs(alliteminfo$theta-0))),!(colnames(alliteminfo)
              %in% c(startitem1,startitem2, "theta"))]))

#create initial testlet
ifelse(startlength==1, assign("initialtestlet",startitem1),
ifelse(startlength==2,assign("initialtestlet",c(startitem1,startitem2)),
ifelse(startlength==3,assign("initialtestlet",c(startitem1,startitem2,
       startitem3)), stop('initial testlet longer than 3 not currently
                           supported'))))

###RUN THE CAT FOR EACH INDIVIDUAL; i = stage, j = individual###

for (j in 1:nrow(imputedresponses)){

  #reset the available item pool for each new individual
  itempool <- alliteminfo

  for (i in 1:numberofstages){
```

```
      #what item(s) presented in this stage
      if (i==1){
        itemspresented <- initialtestlet
        thetahat <- 0
      } else {
        temppool <- itempool
        for (k in 1:stagelength){
          assign(paste0("stageitemnumber",k),
                 names(which.max(temppool[which(abs(temppool$theta-
                 thetahat[1])==min(abs(temppool$theta-thetahat[1]))), -1])))
          temppool <- temppool[,!(names(temppool) %in%
                      get(paste0("stageitemnumber",k)))]
        }
        itemspresented <- unlist(mget(ls(pattern = "^stageitemnumber")))
        rm(list = ls(pattern = "^stageitemnumber"))
      }

      #dynamically create new columns and record within the master dataframe
      #for this stage which item(s) were chosen
      for (k in 1:length(itemspresented)){
        tempoutput[j,paste0("Item",k,"Stage",i)] <- itemspresented[k]
      }

      #record for current loop all items presented to j thus far
      if (i==1){
        testthusfar <- itemspresented
      } else {
        testthusfar <- c(testthusfar,itemspresented)
      }

      #pass response pattern to theta estimation, leaving all non-presented
      #items as NA

      fullresponsevector <- as.numeric(rep(NA,ncol(imputedresponses)))
      fullresponsevector[(colnames(imputedresponses) %in% testthusfar)] <-
        imputedresponses[j,colnames(imputedresponses) %in% testthusfar]
      fullresponsevector<-unlist(fullresponsevector)

      thetahat <- fscores(mirtobject, response.pattern = fullresponsevector,
                  method = estimationmethod, append_response.pattern = FALSE,
                  full.scores.SE = TRUE)

      #record estimation results
      tempoutput[j,paste0("interiminteriinterimtheta",i)] <- thetahat[1]
      tempoutput[j,paste0("interimthetaSE",i)] <- thetahat[2]

      #update item pool
      itempool <- itempool[,!(colnames(itempool) %in% testthusfar)]
    }

    #get final theta/SE's in common format for comparison across conditions
    tempoutput[j,"finaltheta"] <- thetahat[1]
    tempoutput[j,"finalthetaSE"] <- thetahat[2]
  }
  return(tempoutput)
}
```