Adaptive Profile Difference Analysis with Applications to Personality Assessment

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Matthew Allen Snodgress

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Adviser: David J. Weiss, Ph.D.

December 2024

Acknowledgements

I have many people to acknowledge for their support during my time in graduate school. First, I have immense gratitude for my cohort and for the many days and evenings spent working through challenging ideas together: I am grateful for Allison Cooperman, King Yiu Suen, Ziming Clark, and Lauren Berry. I am also grateful for my family. Their continued love and support have kept me going when times were challenging. I am grateful for my four grandparents; and I am especially grateful for my two caring parents, Terry Snodgress and Patricia Snodgress, as well as for my younger sister, Elizabeth Snodgress. Moreover, I am grateful for my fiancée, Hannah Asis, who believed in me and supported me as I wrote this dissertation. I have gratitude for the many nights that you spent listening to my ideas, and for the many hours you spent away from your own work to help me as I struggled. Thank you for all of your support and love.

I am grateful for the faculty of the Quantitative/Psychometric Methods (QPM) Program. Although I have experienced numerous setbacks during my time in graduate school, I have received support from QPM faculty members, helping me grow into the person I am today. Thank you, Dr. Chun Wang, Dr. Niels Waller, Dr. Katerina Marcoulides, Dr. David Weiss, and Dr. Nathaniel Helwig. I am especially grateful to Dr. Weiss for his support and guidance. Finally, I am grateful for the thoughtful feedback provided by Dr. Colin DeYoung and Dr. Lu Yang: thank you for serving on my preliminary and final committees.

Dedication

To my fiancée, Hannah Tamar Asis—we might have started at different places, but I fall in love with you more every day. You are so much more than I could ever want in a partner. I will not forget Miami and what you told me after the film.

Abstract

The adaptive measurement of change (AMC) framework uses item response theory (IRT) and computerized adaptive testing (CAT) to detect psychometrically significant change between two or more occasions for a single individual. In recent years, AMC has been extended to include novel omnibus hypothesis tests for detecting change; multiple measurement occasions; polytomous IRT models; and multidimensional IRT models. In addition, numerous AMC studies support AMC's ability to detect change with high accuracy. One unexplored application of AMC is to the detection of intra-individual, psychometrically significant differences among multiple traits for measurements obtained at a single occasion. Rather than administering one CAT on each occasion for a single trait, a CAT would instead be administered for each trait on one occasion with AMC's hypothesis tests applied to detect significant differences among traits using IRT-based trait estimates. For example, if one individual's score on a measure of Extraversion differs significantly from the same individual's score on a measure of Agreeableness, knowing whether these two personality traits differ significantly could provide useful information about an individual's personality tendencies. Extending the concept to all Big Five personality traits, understanding how such traits differ within a single person could be used to tailor job training or educational interventions. More generally, this procedure, denoted *adaptive profile difference analysis* (APDA), could improve the objective interpretation of multiscale assessments.

In this study, AMC omnibus hypothesis tests were applied to detect intra-individual differences across multiple traits. A Monte Carlo simulation study was conducted using synthetic data based on three real personality datasets. Nine design factors were varied to examine APDA under various realistic conditions. Two primary outcome measures included the true positive rate (i.e., the proportion of true differences over the total number of detected differences) and the false positive rate (i.e., the proportion of detected differences that are significant under conditions where there is no true difference). Findings indicate that APDA is viable under certain conditions, particularly for personality multiscale assessments. Based on these results, recommendations for assessment design and future research are provided.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Multiscale Assessments

In many areas of psychology, multiscale assessments provide useful information about psychological phenomena. In clinical psychology, for example, clinicians use the Minnesota Multiphasic Personality Inventory (MMPI) to inform diagnoses (Ben-Porath & Tellegen, 2020). In personnel selection, industrial psychologists assess potential applicants by administering measures of the Big Five personality traits (McCrae, 2003), which can be predictive of job performance (Tisu et al., 2020). In education, students often complete comprehensive state assessments during specific grade levels to identify areas of strength or weakness. For example, the Minnesota Comprehensive Assessment (MCA) measures students' achievement in science, reading, and mathematics at the eighth-grade level (Lombard, 2006).

In all such cases, a multiscale assessment measures multiple traits, producing a set of scores for each examinee. An educator, psychologist, or practitioner interprets such scores with respect to substantive knowledge about their domain of interest. For instance, a high score on an MMPI clinical scale could indicate mild to severe levels of psychopathology in a particular clinical domain. The presence of multiple scale elevations could be indicative of a particular mental disorder—such as elevations of Demoralization, Low Positive Emotions, and Dysfunctional Negative Emotions suggesting high levels of depression (Ben-Porath, 2012). In the personality domain, the presence of low scores in Extraversion and Openness to Experience could suggest an individual is low in the metatrait of Plasticity (Digman, 1997; DeYoung, 2006), indicating a low tendency toward exploration and toward the creation of new goals or strategies (DeYoung, 2015, p. 47). Such a person might perform poorly in a job that demands high exploration capabilities, such as market research. Across various domains, patterns of scores can constitute score profiles indicative of specific populations. Score profiles can be used in psychological research to connect such populations to outcomes of interest, such as to measures of academic performance.

In addition to indicating populations, score profiles provide an opportunity to rank or prioritize traits. For the MMPI, a score of 65 or greater often indicates an important elevation on a clinical scale (Ben-Porath, 2012). As shown graphically in Figure 1, a client might receive scores of 71, 73, and 80 on Demoralization (RCd); Low Positive Emotions (RC2); and Dysfunctional Negative Emotions (RC7), respectively; and as a result, a clinician might prioritize all three traits but give slightly greater weight to Dysfunctional Negative Emotions, modifying a diagnosis and treatment plan accordingly. Alternatively, in education, a state might establish that a fourth-grade student who scores at or above a 70 out of 100 on science, reading, and mathematics scales demonstrates an acceptable level of performance for that student's grade level. A student receiving scores of 55, 60, and 70, with respect to these scales, might be identified as needing additional learning resources, first in science and second in reading. Science would take priority as the subject with the lowest score below the standard. Finally, in personality research, the Big Five personality traits include Openness to Experience, Contentiousness, Extraversion, Agreeableness, and Neuroticism. At the population level, each of the Big Five scales is normally distributed with scores that can be placed onto a *z*-score metric (Costa & McCrae, 2009). As in Figure 2, an individual might have Big Five scores of 0, 2, 0, 2, and -2, respectively, and would be elevated in both Contentiousness and

Agreeableness while being low in Neuroticism. This pattern indicates high levels of the metatrait Stability, which is related to the control of emotional impulses and the maintenance of long-term goals (DeYoung, 2015, p. 47). Such an individual might be highly capable in certain managerial jobs that require maintaining social bonds among team members.

Figure 1

Example MMPI-3 Profile for a Single Client—Restructured Clinical Scales

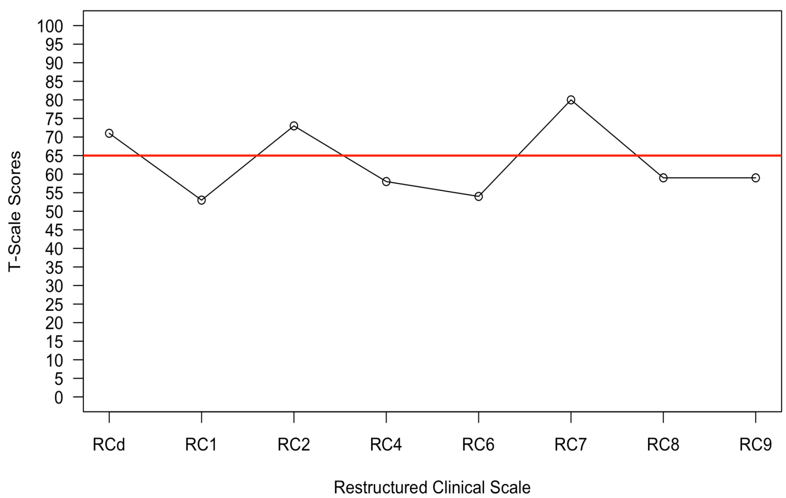
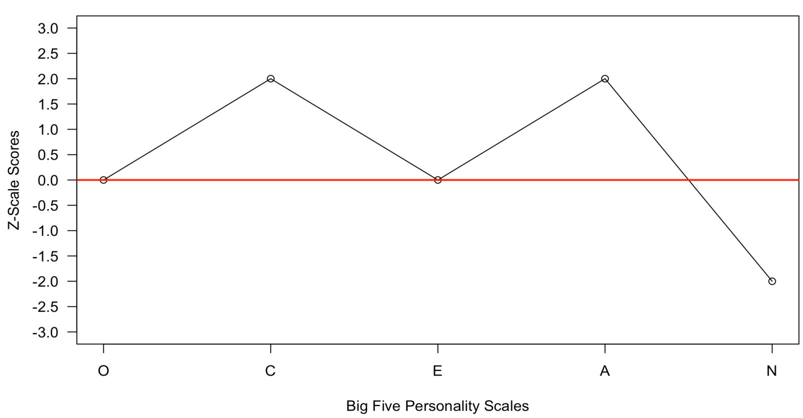


Figure 2

Example Profile for Big Five Personality Traits



While prioritizing certain traits, someone interpreting score profiles typically examines the differences among scores to facilitate the ranking. For instance, assuming scores are on the same scale, a student with scores of 55, 60, and 70 on science, reading, and mathematics scales has a set of score differences of $|55 - 60| = 5$; $|60 - 70| = 10$; and $|55 - 70| = 15$, respectively, in absolute value. Science and reading (the first two scales) are close to one another, but both are more distant from mathematics. However, the student's performance on test day might have been impacted by factors unrelated to science or reading knowledge, such as sickness or sleep quality. The student could also have been fortunate, having received several easier items in reading than in science for the administered form of the assessment. As such, the observed score differences could be due to measurement error. When developing a learning plan, an educator would want to know the degree to which science and reading scores should be interpreted as equivalent with respect to such error.

Classical Test Theory Approaches

Score differences have been examined by researchers in psychological testing, who have identified interpretation limitations and have offered alternatives in a classical test theory (CTT) framework. Researchers highlight that the reliability of a difference score decreases toward zero as the correlation between two traits approaches 1.0 (Cronbach & Furby, 1970). This result follows from the following equation for the reliability of difference scores:

$$\rho_{X-Y} = \frac{\sigma^2_{\tau_X - \tau_Y}(1-\rho_{\tau_X\tau_Y})}{\sigma^2_{\tau_X - \tau_Y}(1-\rho_{\tau_X\tau_Y}) + \sigma^2_{\epsilon_X - \epsilon_Y}}, \tag{1}$$

where ρ_{X-Y} is the reliability of the difference between score sets X and Y, $\sigma^2_{\tau_X-\tau_Y}$ is the variance of the true score difference between the two traits, $\sigma^2_{\epsilon_X-\epsilon_Y}$ is the variance of the error difference between the two traits, and $\rho_{\tau_X\tau_Y}$ is the correlation between the trait scores (Thomas & Zumbo, 2012). A correlation approaching 1.0 sends this equation's numerator toward zero, sending the fraction toward zero. In a multiscale assessment, multiple traits often have positive correlations, suggesting their differences have a reduced reliability. Moreover, researchers have questioned the construct validity of simple difference scores (e.g., the construct of science knowledge minus reading knowledge) as a basis for making judgments about the relative strengths or weaknesses of an examinee, especially given potential scaling differences between two traits measured using a test (Reckase, 1984).

In the context of scores from a multiscale assessment, Anastasi and Urbina (1997, p. 109–110) state, for a single examinee, the following: "It is particularly important to consider test reliability and errors of measurement when evaluating the differences between two scores. Thinking in terms of the range within which each score may fluctuate serves as a check against overemphasizing small differences between scores." They further emphasize how such a question is "especially relevant to the proper interpretation of scores on multiscore batteries in both abilities and personality traits" (Anastasi & Urbina, 1997, p. 109–110). Instead of examining score differences, Anastasi and Urbina suggest including confidence bands around scores, which are based on the standard error of measurement (SEM) in the CTT framework (Anastasi & Urbina, 1997, p. 109–110). A CTT-based SEM can be expressed as

$$\text{SEM} = \sigma_X\sqrt{1-\rho_{XX}}, \quad (2)$$

where σ_X is the standard deviation among a set of observed scores, and ρ_{XX} is the reliability of that set (Crocker & Algina, 2008, p. 122-123). For a single score, a confidence band would be defined as $\text{CI} = x \pm (z)(\text{SEM})$ where x is a single examinee's score, and z refers to a z-score used to define a degree of confidence. When constructing a CI, measurement error follows a standard normal distribution, by assumption. Traits with overlapping confidence bands could be considered equivalent with respect to measurement error, assuming traits are on the same scale. For the educator developing a learning plan, overlapping bands around two scores (e.g., science and reading scores) would suggest that the two subjects could be prioritized equally for the student, even if the student's scores in these subjects differ. Moreover, for an industrial psychologist, a job applicant with overlapping bands between two traits might receive less or more consideration than an applicant whose traits differ beyond measurement error, depending on the directions of the differences and on the job's requirements. However, a remaining question is whether two traits can be compared directly to one another conceptually.

In the CTT framework, alternatives to difference scores and confidence bands have been proposed for interpreting score differences. The reliable change index (RCI) is one such alternative, defined as

$$\text{RCI} = \frac{X-Y}{\sqrt{2(\text{SEM}_{X-Y})^2}}, \quad (3)$$

where $X - Y$ represents the difference between two scores, and where SEM_{X-Y} is the standard error of measurement of difference scores (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1992). The RCI was developed in the context of studying change between scores at two occasions for a single individual on a single trait. Moreover, with the simplifying assumption that the SEM is equal for both occasions, $\mathrm{SEM}_{X-Y} = \sqrt{2}(\mathrm{SEM})$ (Maassen, 2004). The RCI statistic follows a standard normal distribution that defines the difference values expected to occur due to measurement error. From a null hypothesis significance testing (NHST) perspective, an RCI value greater than approximately 1.96 in absolute value indicates an observed difference that would not be expected due to measurement error alone with a Type I error probability of 0.05. Using simulation research, Kruyen, Emons and Sijtsma (2014) examined the ability of the RCI to detect true changes. Their findings suggest that the RCI provides detection rates near 90% accuracy for scales with 40 items and for change magnitudes of 1.5 standard deviations, but performance declined as the change magnitude or the number of items decreased. In practice, the RCI has been used to detect individual change in pre-treatment, post-treatment designs successfully (Speer & Greenbaum, 1995). Applied to detecting differences between two traits, the RCI would have a similar form with the SEM_{X-Y} modified to no longer assume equal SEM: instead, in this case,

$$\mathrm{SEM}_{X-Y} = \sqrt{\sigma_{\tau_X}^2(1-\rho_{XX}) + \sigma_{\tau_Y}^2(1-\rho_{YY})}. \tag{4}$$

Note that no research has examined the use of the RCI at a single occasion between two different traits.

While the RCI is conceptually simple and provides utility, researchers have identified limitations. Critically, because the RCI depends on CTT-based SEM, the RCI assumes an equal SEM for all measured individuals, which is a basic assumption in CTT (Lord & Novick 1968). However, different students, job applicants, or clients would be expected to have different levels of measurement error for each trait measured. Moreover, individuals at the extreme ends of a trait continuum would be expected to have larger SEM values when using CTT-based SEM, which is not optimal. For a multiscale assessment with T scales, a total of $T(T-1)/2$ RCI values would be used for pairwise comparisons; for example, for a Big Five assessment, 10 RCI values would be used to compare the $T=5$ scores from the assessment. Using the RCI could lead to more inaccurate comparisons as the value T increases because individual differences in measurement error are not considered, possibly reducing the usefulness of RCI-based score profile interpretations for any one individual.

An alternative to the RCI is the minimally important difference (MID; Jaeschke, Singer, & Guyatt, 1989), used extensively in research on patient-reported outcomes (Revicki, et al., 2008). Between two measurement occasions, the MID is the smallest perceived change—positive or negative—in a measured outcome (1) that clients or clinicians consider important and (2) that indicates a need to modify the client's care. The MID can be calculated by anchor-based approaches or distribution-based approaches. The former approach uses an anchor. For example, clients might be asked to describe their perceived change; are categorized into groups based on their perceived change (e.g., "slightly better", "better", "much better"); and scores on the outcome (e.g., a depression scale) for the minimal group (e.g., "slightly better") are averaged to define the MID (Lassere, van der Heijde, & Johnson, 2001). Guyatt, Walter, and Norman (1987) have described the judgment of change to be difficult for patients, however, and have questioned whether some anchor measures are reliable and valid. In addition, averaging introduces a limitation: individuals likely have different thresholds for judging change, so the calculated average difference obscures individual starting points. Comparatively, distributional methods rely on statistical properties of change scores to calculate the MID, such as using the criterion of 0.5 standard deviations above the mean of the observed difference score distribution. More objective distributional approaches also suffer from the averaging issue (Mouelhi et al., 2020).

Like the RCI, the MID has not been applied to the analysis of difference scores in multiscale assessments administered at a single occasion. In this context, the MID would be the smallest perceived difference between two scales that an examinee or someone interpreting the assessment considers important with respect to an outcome. In the context of a clinical assessment, the MID might be the smallest difference between two clinical scales where one scale gives more practical utility in informing diagnosis over the other scale. For personality, an MID in a human resources context might indicate the smallest difference between two personality traits where one trait begins to indicate job performance better than the other. Because the MID is linked closely to outcomes, however, wide variability in MID values would be expected across many potential applications of the measure. Moreover, even in the context of score profiles, the MID still relies on group-based statistics that overlook individual-level measurement—potentially obscuring individual differences in optimal treatment or performance. The MID also remains unexplored in this context and might be difficult to interpret conceptually.

Statistical and Latent Variable Approaches

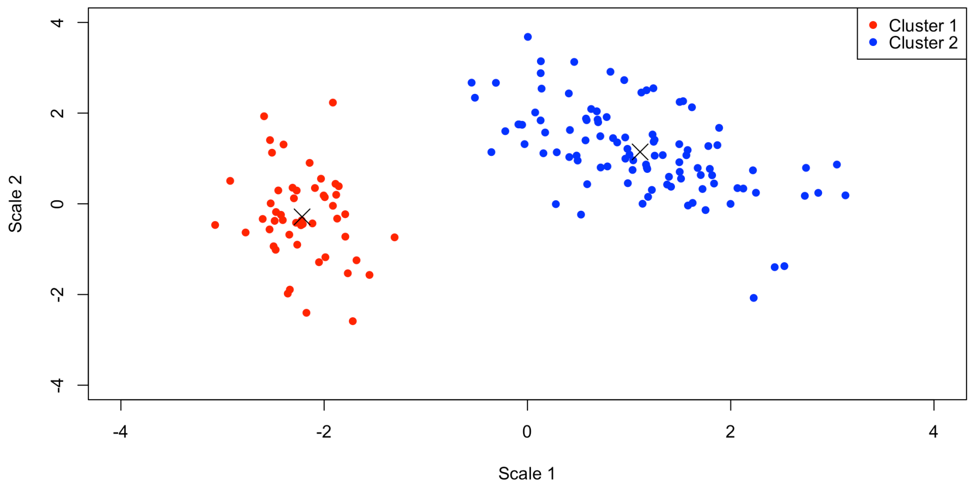
Beyond CTT-based measurement approaches, statistical methods have been developed to identify groups that demonstrate similar score profile patterns with the objective of describing individuals in terms of such patterns. Cluster analysis (CA) represents one major family of these methods. In CA, heterogeneity among a set of examinees is reduced by finding locations where many examinees can be found in the multivariate space formed by multiple scales. To this end, hierarchical methods generally build a nested tree structure of clusters, either by beginning with all examinees in unique clusters and iteratively combining them, or by starting with a single cluster and dividing it according to a rule (Lattin, Carroll, & Green, 2003a, p. 264-266). Multiple rules exist, such as measuring distance between any two examinees using Euclidian distance, written, for some k-th and j-th examinees, for $t = 1, \ldots, T$ scales, as

$$d_{kj} = \left(\sum_t^T \left(x_{kt} - x_{jt}\right)^2\right)^{1/2}, \tag{5}$$

where x_{kt} is the k-th examinee's score on the t-th scale, and where x_{jt} is the j-th examinee's score on the t-th scale. In contrast, rather than building a full tree structure, partitioning methods identify a pre-selected number of clusters and consider (1) how similar examinees are within groups and (2) how dissimilar they are from one another. In both types of clustering, the objective is to use distance, similarity, and dissimilarity rules to find subsets of examinees (Lattin, Carroll, & Green, 2003a, p. 273-279). Various methods also exist for choosing an optimal number of clusters—including rules based on information criteria or cross-validation, among others (Milligan & Cooper, 1985). A conceptual example in Figure 3 displays the results of a CA applied to simulated scales with a two-cluster solution. In the example, Cluster 1 represents a group of examinees who score low on Scale 1; and Cluster 2 consists of examinees who score high on Scale 1 and high on Scale 2. As Cluster 2 demonstrates, not all examinees follow the Cluster 2 pattern.

Figure 3

Example Results of a Cluster Analysis with Two Scales and Two Clusters



CA has been used to identify distinct groups of score profiles in multiscale assessments across various domains. Spaans et al. (2009), for example, used hierarchical CA to identify groups in a sample of pretrial criminal defendants using the MMPI-2, finding a defendant cluster that was generally elevated on the MMPI and a cluster that was not. Moreover, model-based CA—where mixtures of probability distributions are assumed to generate observed data—has been applied to discover psychopathy subtypes using the Multidimensional Personality Questionnaire (MPQ; Hicks et al., 2004). Using the Bayesian information criterion (BIC) to select an optimal number of clusters, researchers found two distinct subtypes: individuals with psychopathy who are stable emotionally and individuals with psychopathy who tend to be aggressive. In the educational domain, CA has been used to find writing strengths and weaknesses in third-grade and eighth-grade students who completed a writing assessment consisting of six writing dimensions, which trained judges used to score students' writing samples (Roid, 1994). Analyses revealed 11 unique student clusters, each demonstrating different patterns on the six dimensions: for instance, one cluster showcased strengths in ideas, organization, and voice, together with weaknesses in sentence fluency and mechanics.

While such studies demonstrate the ability of CA to uncover subpopulations in heterogeneous data, CA has several limitations with respect to interpreting individual score profiles. First, as an algorithmic and/or statistical approach, CA does not use measurement error as a basis for making score comparisons. Without accounting for measurement error, CA likely biases the assignment of individuals into specific clusters, especially for individuals on the threshold between two or more clusters. In addition, CA is inherently a group-based method, whereby individuals are evaluated with respect to clusters of other individuals. One consequence is that variation is lost in individual score profiles. That is, comparing an individual to a group entails a loss of individual nuance. As the number of scales T increases, moreover, the loss of variation can increase, making interpretations of individual score profiles less meaningful for larger multiscale assessments. Finally, while methods exist for choosing the optimal number of clusters (e.g., BIC), different numbers of clusters can often be equally viable and offer different interpretations of any single individual's score profile.

An alternative approach to CA is multidimensional scaling (MDS). While CA uses the scores themselves to find distances among examinees to form clusters, MDS uses the distances between examinees to place them onto a low-dimensional map of the data. MDS represents a set of methods with different assumptions—including metric MDS (Shepard, 1962a, 1962b); nonmetric MDS (Kruskal, 1964a, 1964b); and individual differences scaling (INDSCAL; Carroll & Chang, 1970). Historically, MDS has been applied to translate the distances among objects into common spatial coordinates, such as using distance values among cities to construct a map of those cities in space (Carroll & Arabie, 1998). More recently, Davison, Kim and Ding (2001) have applied MDS to identify profiles among examinees in multiscale assessments. Called profile analysis via multidimensional scaling (PAMS), the PAMS model is a type of linear latent variable model that can be written

$$x_{jt} = c_j + \sum_{r=1}^{R} w_{jr}\, m_{tr} + e_{jt}, \tag{6}$$

where x_{jt} is the j-th examinee's score on the t-th scale, c_j is the j-th examinee's unweighted mean score across all scales, $r = 1, \ldots, R$ latent dimensions are assumed, w_{jr} is the j-th examinee's weight on the r-th dimension, m_{tr} is the t-th scale's weight on the r-th dimension, and e_{jt} is an error term associated with x_{jt}. PAMS makes several restrictions and assumptions to specify the model, particularly on the error term. For instance, PAMS assumes scale weights sum to zero. PAMS can be considered an unstandardized version of the Q-factor analysis model (Cattell, 1952); but unlike Q-factor analysis, PAMS allows for an exploration of the shapes and patterns of individual score profiles while accounting for overall profile level (i.e., the average score among all scales). In the context of cognitive testing, Kim, Frisby, and Davison (2004) applied the PAMS model to identify core profiles in the Woodcock-Johnson Psychoeducational Battery-Revised (Woodcock, 1977), examining a subset of individuals from the battery's standardization sample. Among seven cognitive traits, they identified a PAMS solution with two latent dimensions: (1) High Comprehension Knowledge with Low Long-term

Retrieval, and (2) High Processing Speed and High Visual Processing with Low Short-term Memory. Individual examinees received weights indicating the degree of correspondence between their individual score profile and either of these core profiles from the overall sample.

The PAMS model offers some advantages over CA, such as directly modeling an individual's general elevation across all scales of a multiscale assessment (e.g., a person's general elevation across all scales of the MMPI) while accounting for individual shapes and patterns. Moreover, individual score profiles can be evaluated by the degree that they correspond to latent dimensions, which is represented by the weight w_{jr}. Nevertheless, like CA, PAMS relies on group-based data to find common types of profiles, reducing variation in individual score profiles. In addition, although MDS methods exist to optimize the number of latent dimensions (Lattin, Carroll, & Green, 2003b, p. 216-228), choosing the number of latent dimensions introduces some subjectivity, similar to finding an optimal number of clusters in CA. Finally, like other linear factor analytic solutions, the PAMS solution experiences a rotational indeterminacy with no unique solution: different post hoc rotations could offer different interpretations of individual score profiles. One exception is the INDSCAL model (Carroll & Chang, 1970), which is rotationally unique but does require special types of data structures.

Together with CA and PAMS, latent profile analysis (LPA) represents a major set of profile discovery techniques. A key objective of the LPA model is to perform a probabilistic classification of individuals into latent classes using a set of indicators. A type of finite mixture model (McLachlan, Lee, & Rathnayake, 2019), LPA is mathematically equivalent to a latent class analysis (LCA) model but uses indicators that are continuous. According to Vermunt and Magidson (2002, p. 91), a LCA model has the form

$$f(\mathbf{x}_j|\boldsymbol{\theta}) = \sum_{r=1}^{R} \pi_r \, f_r(\mathbf{x}_j|\theta_r), \tag{7}$$

where $\mathbf{x}_j$ denotes the j-th individual's $T \times 1$ vector of observed, manifest variables, for $j = 1, \dots J$ individuals; R is the number of classes, for $r = 1, \dots, R$; π_r is either the prior probability of membership in the r-th latent class or is the size of the r-th class, depending on the specific formulation; $\boldsymbol{\theta}$ is an $R \times 1$ vector of model parameters, and θ_r are the parameters specific to the r-th class; and $f_r(\mathbf{x}_j|\theta_r)$ is a class-specific probability density function for the r-th class. In LPA, a mixture of these densities forms $f(\mathbf{x}_j|\boldsymbol{\theta})$.

Like CA and PAMS, LPA has been applied to multiscale assessments, particularly in the personality domain. For instance, LPA has been used to search for personality types, which would represent interactions among personality traits. Merz and Roesch (2011) found a three-class LPA solution for an assessment of the Big Five, finding core profiles representing individuals who are reserved, well-adjusted, or excitable. Others have identified a five-class LPA solution (Kinnunen et al., 2012)—finding types that are resilient, overcontrolled, reserved, undercontrolled, or ordinary. More recently, Yin et al. (2021) examined 36 studies applying LPA to Big Five assessments between 2010 and 2020, showing that solutions typically range from two to five core profiles. Importantly, reviewers noted that the number of identified profiles increased as the number of items in each scale decreased. Likely, because measurement error increases with fewer items, artificial noise variance is added and pushes optimization rules or procedures—such as information criteria or the bootstrap likelihood-ratio test (McLachlan & Peel, 2000)—

toward higher numbers of classes to account for this variance. Conceptually, compressing personality into three or four types abstracts the meaning of personality traits, thereby eliminating variation in individual score profiles.

While LPA or similar methods are useful in identifying sets of common score profiles, all such methods reduce the degree of variation expected to occur among individuals. Moreover, any single individual must be compared to clusters, dimensions, or classes, which locates that individual with respect to other persons rather than centering the analysis on the individual themselves (i.e., rather than taking an intra-individual approach). Measurement error, moreover, does not contribute to solutions for statistical and latent variable modeling approaches with the exception of factor analysis models (e.g., factor mixture models; Lubke & Muthén, 2005). Nevertheless, factor analysis models typically emphasize latent trait factor structure more than item properties or individual scoring (Reckase, 1997a, p. 26-27). In contrast to CTT-based methods and statistical or latent variable approaches, a method is needed that (1) allows for measurement error to vary across the latent trait continuum, and (2) focuses more precisely on the measurement of individuals. Such a method could improve the interpretation of individual score profiles.

Item Response Theory Models

Compared to CTT approaches, which focus on maximizing reliability and are sample-dependent, item response theory (IRT) focuses on the probabilistic modeling of individual responses to items (Lord & Novick, 1968). In contrast to other statistical methods or latent variable models, IRT places special emphasis on item properties and individual scoring in a psychometric framework. Thus, IRT models represent one possible approach for enhancing the interpretation of individual score profiles by improving individual-level measurement.

The most common IRT models were developed for items with a binary response format, such as correct-incorrect responses to educational assessment questions. To model the probability of a correct response, the logistic three-parameter (3PL) IRT model has the form

$$p_{ij} = P\left(U_{ij} = 1 | a_i, b_i, c_i, \theta_j\right) = \left(c_i + (1 - c_i)\right)\left(1 + \exp\left(-1.7a_i(\theta_i - b_i)\right)\right)^{-1}, \quad (8)$$

for the j-th individual among $j = 1, \dots, J$ individuals, and for the i-th item among $i = 1, \dots, I$ such items. In this model, U_{ij} is a random variable specifying the j-th person's response to the i-th item; and U_{ij} follows a Bernoulli distribution with success parameter p_{ij}. Moreover, a_i is the i-th item's discrimination parameter specifying how well the item discriminates among individuals; b_i is the i-th item's difficulty parameter specifying how challenging the item is to individuals; and c_i represents the function's lower asymptote, modeling guessing in ability tests or other effects like social desirability in personality assessment. The parameter θ_j specifies the j-th individual's ability on the latent trait. Critically, the person parameter θ_j is placed onto the same scale as the difficulty parameter b_i, linking persons and items onto a common scale. This link provides an opportunity to compare person ability or trait level to item difficulty directly, allowing for the possibility of tailoring assessment items to a single individual's trait level. Finally, the model in Equation 8 is unidimensional, assuming a single latent trait is responsible for observed item responses (Lord & Novick, 1968; Embretson & Reise, 2000a).

Under the assumption of local independence between items (i.e., conditional on the latent trait θ), the joint probability of the j-th individual's $I \times 1$ response vector $\boldsymbol{U}_j$, for all I items, can be written

$$P(\boldsymbol{U}_j | \mathbf{a}, \mathbf{b}, \mathbf{c}, \theta_j) = \prod_{i=1}^{I} P\left(U_{ij} = 1\right)^{U_{ij}} \left(1 - P(U_{ij} = 1)\right)^{1-U_{ij}}, \tag{9}$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are each $I \times 1$ vectors containing discrimination, difficulty, and guessing item parameters, respectively (van der Linden, 2016). Importantly, item parameters are assumed to be invariant within a linear transformation across different samples of examinees, differentiating IRT models from sample-dependent CTT methods. By setting all c_i parameters to zero in the 3PL model, the model reduces to a two-parameter logistic model (2PL; Lord & Novick, 1968), often employed in cases when no guessing or analogous effects occur (e.g., social desirability in personality tests). By further setting all a_i parameters to one, the 2PL reduces to a one-parameter logistic model (1PL; Lord & Novick, 1968), often used in a Rasch modeling framework (Rasch, 1960, 1966). Fit indices have been developed to assess the fit of specific items within the 3PL, 2PL, and 1PL (Orlando & Thissen, 2000, 2003). Note that, in IRT, a common method for estimating person score θ_j is maximum likelihood estimation (MLE). In MLE, a likelihood function can be defined as

$$L(\theta_j | \, \boldsymbol{u}_j, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{i=1}^{I} P\left(u_{ij} = 1\right)^{u_{ij}} \left(1 - P(u_{ij} = 1)\right)^{1-u_{ij}}, \tag{10}$$

where parameters are the same as defined in the joint probability, except $\boldsymbol{u}_j$ represents the j-th individual's observed $I \times 1$ response vector. Then, $\hat{\theta}_j$ is the value that maximizes log(L) = $\ell(\theta_j | \boldsymbol{u}_j, \mathbf{a}, \mathbf{b}, \mathbf{c})$—which is found by finding the first derivative of ℓ, setting it equal to zero, and solving for θ_j.

Together with IRT models for binary responses, IRT models have been developed for items with three or more response options, such as Likert-type items with three or more ordered response categories. Among such polytomous IRT models, the graded response model (GRM) is one of the most commonly applied for items with ordered response categories (Samejima, 1969). The GRM first models the boundaries between response categories: for integers $m = 0, 1, \ldots, M$, and for $k = 1, 2, \ldots, M + 1$ response categories, the GRM boundary response function has the form

$$P^*_{ijm}(\theta_j) = \left(1 + \exp\left(-1.7a_i(\theta_j - b_{im})\right)\right)^{-1}, \tag{11}$$

where a_i is the i-th item's discrimination parameter, b_{im} is the i-th item's m-th boundary location parameter, θ_j is the j-th individual's person parameter, and $P^*_{ijm}(\theta_j)$ is the probability of responding at or above the $(m + 1)$-th category. In addition, $P^*_{ij0}(\theta_j)$ = 1 because one of the responses is selected by assumption. The GRM further specifies that the probability of responding within any specific k-th category is

$$P_{ijk}(\theta_j) = P^*_{ijm}(\theta_j) - P^*_{ij(m+1)}(\theta_j), \tag{12}$$

where $P^*_{ij(M)}(\theta_j) = P_{ij(M+1)}(\theta_j)$. The GRM can be considered an extension of the 2PL IRT model to cases with more than two response options that are ordered (Embretson & Reise, 2000b). Other IRT models exist for handling polytomous responses—such as the nominal response model (Bock, 1972); the rating scale model (Andrich, 1978); or the partial credit model (Masters, 1982).

In the context of multiscale assessments, both binary and polytomous IRT models have been applied extensively. In the educational domain, both the Scholastic Aptitude Test (SAT; Angoff, 1971) and the Graduate Record Examination (GRE; Conrad, 1977) have been developed using IRT models. For instance, Educational Testing Service, responsible for the GRE, has modeled items on its GRE Verbal and GRE Quantitative scales using the 2PL model (Zwick, 2012). In addition, IRT models have been used with large-scale educational survey assessments, such as the National Assessment of Educational Progress (NAEP) and the Trends in Mathematics and Science Study (TIMSS; van Rijn, et al., 2016). In studying psychopathology, Reise and Waller (2003) applied the 2PL and 3PL models to 15 scales of the MMPI-A—the adolescent version of the MMPI—finding (1) good fit for both models on many of these scales and (2) that the 3PL could offer important interpretational advantages (e.g., modeling social desirability or faking).

In the personality domain, Reise and Waller (1990) further applied the 2PL to model binary personality items on the 11 primary scales of the Multidimensional Personality Questionnaire (MPQ) for individuals drawn from the Minnesota Twin Registry (Tellegen & Waller, 2008), demonstrating good fit across scales. Reise and Widaman (1999), moreover, converted binary items into parcels for three MPQ scales and summed them to create Likert-type items, applying the GRM to model such items and achieving (1) reasonably high discriminations and (2) good person-fit based on IRT-based individual fit indices (see Drasgow, Levine, & Williams, 1985). Comparatively, the GRM has shown poorer fit for certain attachment scales (Fraley, Waller, & Brennan, 2000) and for scales from an assessment of Cattell's 16 personality factors (16PF) and the Big Five (Cattell &

Mead, 2008; Chernyshenko, et al., 2001). However, in this last case, poor item fit might have been a function of the use of classic chi-square statistics to assess fit (see Yen, 1981). Improved statistics have since been developed for polytomous models (Kang & Chen, 2008, 2011). The 2PL and 3PL were also applied to model the same two assessments with results being somewhat better than those obtained using the GRM (Chernyshenko, et al., 2001).

Notably, multidimensional IRT (MIRT) models have been applied to multiscale assessments. MIRT models are multidimensional extensions of unidimensional IRT models to cases where multiple latent traits correlate and are measured simultaneously. In a MIRT model, an item could be associated with a single dimension (i.e., between-items multidimensionality) or with multiple dimensions (i.e., within-items multidimensionality); and items between correlated dimensions can contribute to the scoring of individuals on either dimension (Reckase, 1997b). MIRT models can be viewed as nonlinear cases of factor analysis (Bock, Gibbons, & Muraki, 1988). In the context of multiscale assessments, a MIRT extension of the 3PL has been applied successfully to a ninth-grade educational battery with mathematics, spelling, math-computation, and social studies scales (de la Torre & Patz, 2005), achieving high measurement precision for these abilities. In simulation work, the same study also showed that MIRT-based measurement precision improves as trait correlations increase and as test lengths decrease, which highlights how MIRT models can use correlated scales to achieve better score estimates when some scales have poor measurement precision. Wang, Chen, and Cheng (2004) achieved similar results for polytomous items with a multidimensional extension of the rating scale model to the 10-scale Teacher Personality Inventory (Wang, 1997).

Across different domains, IRT models demonstrate utility in modeling both binary and polytomous items, both in unidimensional and multidimensional settings. While large educational assessments are often designed using IRT throughout scale development, clinical and personality assessments typically have not been based on IRT. Most applications of IRT to clinical or personality assessments represent *ad hoc* analyses based on conventional tests, showing fit and measurement accuracies likely lower than what would be expected had IRT been involved in scale development. Even so, the 2PL, 3PL, and GRM models have good measurement properties and could form the basis for multiscale assessments in the clinical or personality domains. MIRT extensions also demonstrate usefulness but have limitations: foremost, the interpretation of scores from any single dimension can become more complicated as correlations among traits increase. Moreover, MIRT model complexity might make generalizability more challenging in some settings, such as in certain cross-cultural settings where culture-specific dimensions could exist. Overall, IRT research shows how precise individual scoring can be achieved in multiscale assessments; but previous research has not investigated how to compare IRT-based scores to one another directly to facilitate score profile interpretations.

Adaptive Measurement of Change

One possibility for comparing IRT-based scores to one another can be found in the computerized adaptive testing (CAT) literature. A CAT is a procedure that administers a series of items according to how an examinee responds to each item. A CAT consists of an IRT model, an item bank, rules for how to begin the procedure, a rule for choosing each item, a method to score an examinee after each item is administered, and a rule for terminating the procedure (Weiss & Kingsbury, 1984). For example, the first administered item's difficulty could be high or low, depending on the objective of the CAT. Moreover, a CAT might terminate when a particular IRT-based SEM has been achieved for an examinee's score estimate. Over time, CATs have been studied extensively and have been used in numerous applications–including in educational, personality, and psychopathology assessments (Weiss & Sahin, 2024). For example, Reise and Henson (2000) used the GRM to develop a CAT version of the NEO Personality Inventory (NEO-PI; Costa & McCrae, 1985), consisting of 30 unidimensional facet scales.

Within the CAT literature, an emerging area of research focuses on the adaptive measurement of change (AMC). AMC uses IRT and CAT to detect psychometrically significant change occurring between two or more occasions for a single individual. In its original conceptualization for two occasions (Weiss & Kingsbury, 1984), one individual receives two CATs, one for each occasion; and IRT-based, observed SEMs are used to construct confidence intervals (CIs) of trait estimates at each occasion. More specifically, the observed SEM is expressed as

$$\mathrm{SE}\left(\hat{\theta}_j\right) = \left(I\left(\hat{\theta}_j\right)\right)^{-1/2} = \left(\frac{-\,\partial^2 \ell_j(\hat{\theta}_j)}{\partial \hat{\theta}_j^2}\right)^{-1/2}, \tag{13}$$

where $\hat{\theta}_j$ is the MLE estimate of the latent trait for the j-th examinee; $I\left(\hat{\theta}_j\right)$ denotes the response pattern information; and ℓ_j is the log-likelihood function for the set of items administered to the j-th examinee. Then, an IRT-based psychometric CI, at a 95% degree of confidence, would be $\hat{\theta}_j \pm 1.96\left(\mathrm{SE}\left(\hat{\theta}_j\right)\right)$, for the j-th examinee. Non-overlapping CIs are used to indicate significant change between two occasions. Critically, assuming a pre-calibrated item bank already exists, $\mathrm{SE}(\hat{\theta}_j)$ contains no group-based information. AMC is, therefore, an intra-individual procedure for measuring change within a single person (Weiss & Kingsbury, 1984).

Although referring to one trait measured at two occasions, the AMC concept of examining overlapping CIs relates to the question posed by Anastasi and Urbina (1997), who asked whether different traits on a multiscale battery could be compared to one another using CTT-based confidence bands. One new method for interpreting score profiles could be to apply the AMC concept to multiscale assessments. Using IRT and CAT-based CI could achieve better measurement precision and could advance the individualized ranking of traits (e.g., how best to prioritize some subjects, like science or reading, over others for studying) or the identification of unique sets of score profiles (e.g., discovering subpopulations). Moreover, using CAT could improve test efficiency by reducing the number of items administered and could tailor tests to an individual's specific trait levels on each scale.

King-Kang and Weiss (2008) conducted the first validation study of the CI method—comparing CTT-based methods, such as the difference score and the residual change score (Manning & DuBois, 1962), to an IRT-based difference score and to the AMC CI method in a Monte Carlo simulation study. Given that the CI method depends on IIFs, and given that the magnitude of some i-th item's information $I_i(.)$ is largely dependent on item discrimination, they simulated three item banks containing items with three different levels of discrimination to observe how discrimination could impact CIs. A 3PL model was used to calibrate items in these banks with difficulty parameters centered approximately on 0 and spanning the range $[-4.5, 4.5]$, and with guessing parameters set at 0.2. Conventional tests were formed from the simulated item banks with a length of 50 items across six parallel forms, and they administered fixed-length CATs (FL-CATs) with 50 items using maximum likelihood scoring. In terms of both bias and root mean squared error (RMSE), the AMC CI method outperformed all other methods. Moreover, as item bank discriminations increased, or as true change magnitudes increased, the recovery of change improved. In a separate study, Weiss and von Minden (2011) extended this work to examine different change patterns across five occasions, discovering again that the CI method outperformed CTT-based methods and IRT-based difference scores in terms of bias and RMSE.

From an NHST perspective, Finkelman et al. (2010) introduced two AMC omnibus hypothesis tests to refine the method further: (1) the AMC Z-test and (2) the AMC likelihood-ratio test (LRT). For a single examinee, the AMC Z-test statistic is defined as

$$Z_{\text{obs}} = \left|\hat{\theta}_{j2} - \hat{\theta}_{j1}\right| \left(\left(\sum_{v=1}^{V} I_v\left(\hat{\theta}_{\text{pool}}\right) \right)^{-1} + \left(\sum_{l=1}^{L} I_l\left(\hat{\theta}_{\text{pool}}\right) \right)^{-1} \right)^{-1/2}, \tag{14}$$

where L items were administered at the first occasion, V items were administered at the second occasion, and a CAT delivered the items at both occasions. The j-th examinee's final trait estimate at the first occasion is $\hat{\theta}_{j1}$ and at the second occasion is $\hat{\theta}_{j2}$. The pooled estimate $\hat{\theta}_{\text{pool}}$ is obtained by scoring the examinee using the responses from both occasions after both CATs have completed. In the case of two occasions, the significance rule is whether $Z_{\text{obs}} \geq z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ defines a cutoff value on a standard normal distribution. More generally, Phadke (2017) extended the Z-test to any number of occasions. For T occasions, $H = T(T-1)/2$ unique Z_{obs} statistics exist for a single examinee. Assuming approximate independence among such statistics, the H statistics can be combined with

$$\chi^2_{Z_{\text{obs}}} = \sum_{h=1}^{H} Z_h^2, \tag{15}$$

where $\chi^2_{Z_{\text{obs}}}$ approximately follows a chi-square distribution with H degrees of freedom. If $\chi^2_{Z_{\text{obs}}} \geq \chi^2_{(H,1-\alpha)}$, at least one significant difference exists among the H possible comparisons.

Together with the Z-test, Finkelman et al. (2010) introduced the AMC LRT statistic for two occasions. Using the direct extension from Phadke, 2017, the LRT statistic is defined as

$$\begin{aligned} \lambda_{\text{obs}} &= -2\log\left(\frac{L(\hat{\theta}_{\text{pool}}|\mathbf{u}_{\text{pool}})}{\prod_{t=1}^{T} L_t(\hat{\theta}_t|\mathbf{u}_t)}\right) \\ &= -2\left(\ell(\hat{\theta}_{\text{pool}}|\mathbf{u}_{\text{pool}}) - \sum_{t=1}^{T} \ell_t\left(\hat{\theta}_t|\mathbf{u}_t\right)\right), \end{aligned} \tag{16}$$

where $L(.)$ defines the likelihood of the observed response pattern when combining all responses, which are contained in the response vector $\mathbf{u}_{\text{pool}}$; $L_t(.)$ is the likelihood of the response vector at the t-th occasion; ℓ is the log-likelihood; $\mathbf{u}_t$ contains observed responses at the t-th occasion; and $\hat{\theta}_t$ is the trait estimate at the t-th occasion, for $t = 1, \ldots, T$ occasions, for a single examinee (the examinee index is dropped for simplicity). If $\lambda_{\text{obs}} \geq \chi^2_{(T-1,1-\alpha)}$, at least one significant difference exists among the occasions. Critically, while related to statistical theory, these AMC hypothesis tests are psychometric tests because they achieve better performance as the number of items (and thus, as overall test information based on those items) increases, not as the number of observations/persons increases.

In a Monte Carlo simulation study, Finkelman et al. (2010) found that both the Z-test and LRT outperformed the CI method in terms of power (i.e., true positive rate). To control Type I error, a value of $\alpha = 0.05$ was used in the simulation, and both hypothesis tests controlled the Type I error rate close to the nominal value of 0.05. Examining three change magnitudes at two occasions—magnitudes of 0.5, 1.0, and 1.5, which were based on an analysis of elementary school achievement tests—they showed that power close to 0.9 could be achieved with their hypothesis tests for magnitudes of 1.0 or greater. However, at a change magnitude of 0.5, hypothesis tests demonstrated power below approximately 0.5. In addition, they employed variable-length CATs (VL-CATs) and found moderate reductions (approximately a 10% reduction) in the number of administered items for a change magnitude of 0.5 and major reductions (approximately 40% to 50%) for magnitudes of 1.0 and 1.5, demonstrating the potential of VL-CATs to improve test efficiency in the AMC framework. Notably, Finkelman et al. (2010) explored different strategies for item selection—based on either Fisher information or Kullback–Leibler information (KLI; see Cover & Thomas, 1991)—finding similar results but a very slight advantage for KLI.

Building on the findings from Finkelman et al. (2010), Lee (2015) proposed an AMC score test. Drawing on the work of Rao (1948, 1965), Lee defined the AMC score statistic as

$$S_{\text{obs}} = \sum_{t=1}^{T} s\left(\hat{\theta}_{\text{pool}}|\mathbf{u}_t\right)^2 \left(I_t\left(\hat{\theta}_{\text{pool}}\right)\right)^{-1}, \tag{17}$$

where I_t is test information for the t-th trait, and $s(.)$ is the score function, the first derivative of the log-likelihood function. If $S_{\text{obs}} \geq \chi^2_{(T-1,1-\alpha)}$, at least one significant difference exists among the occasions. Note that, like Finkleman et al., 2010, Lee (2015) formulated an AMC test for two occasions while Phadke (2017) extended the test directly to any number of occasions. In this test, Lee (2015) reasoned that the score statistic only depends on parameters estimated under the null hypothesis while the LRT requires parameters from both the null and alternative hypotheses to be estimated. Further, while the score statistic tends to converge to the LRT statistic asymptotically (Cox & Hinkley, 1979), the two statistics could yield different outcomes in finite samples (e.g., see Johnston & DiNardo, 1997, p. 150). In a Monte Carlo simulation design, Lee discovered that the score test achieved power and Type I error results very close to the LRT, under various data-generating conditions, offering an additional AMC test for future research.

As previously mentioned, Phadke (2017) extended the Z-test, LRT, and score test to any number of occasions. Phadke (2017) also introduced two types of AMC F-tests. The F1 statistic is defined as

$$\mathrm{F1}_{\mathrm{obs}} = \frac{\left(n_t^{-1}\sum_{t=1}^{T}(\hat{\theta}_t - \hat{\theta}_G)\right)(T-1)^{-1}}{\left(\sum_{t=1}^{T}\left(I_t(\hat{\theta}_{\mathrm{pool}})\right)^{-1}\right)(N-T)^{-1}}, \tag{18}$$

where n_t is the number of items administered at the t-th occasion, $\hat{\theta}_G$ is the average of all T such $\hat{\theta}_t$ estimates, and $N = \sum_{t=1}^{T} n_t$. Moreover, a second F statistic was defined as

$$\mathrm{F2}_{\mathrm{obs}} = (T)(\mathrm{var}(\hat{\boldsymbol{\theta}})\left(\sum_{t=1}^{T}\left(I_t(\hat{\theta}_{\mathrm{pool}})\right)^{-1}\right)^{-1}, \tag{19}$$

where $\mathrm{var}(.)$ defines the sample variance observed among the T trait estimates contained within the $T \times 1$ vector $\hat{\boldsymbol{\theta}}$. For either statistic, significance is indicated when $\mathrm{F}_{\mathrm{obs}} \geq \mathrm{F}_{T-1,N-T}$, where $\mathrm{F}_{T-1,N-T}$ defines a cutoff value on an F-distribution with $T-1$ and $N-T$ degrees of freedom. In simulation research, Phadke (2017) demonstrated that these

AMC hypothesis tests have good power (ranging from 0.7 to 0.8) and control of Type I error (ranging from 0.05 to 0.09) for three or more occasions. Like Finkleman et al., 2010, Phadke used change magnitudes of 0.5, 1.0, and 1.5, and found that power improved as magnitudes increased. Under various conditions, the $F1_{obs}$ statistic and $F2_{obs}$ statistic performed similarly to the Z-test and score tests; but across these studies, the LRT provided the best balance between power and controlling Type I error for multiple occasions.

While these studies examined a single trait measured at two or more occasions, Wang and Weiss (2018) examined AMC hypothesis tests in the context of multidimensional scenarios with two or more traits measured at two occasions, extending the Z-test, LRT, and score test to multivariate cases. In a simulation study, they found that a multivariate score test and multivariate LRT can achieve a good balance of power and Type I error when three latent traits are measured at two occasions. Notably, they introduced a new test—the Kullback–Leibler divergence test—but did not determine that it showed an improvement over existing AMC hypothesis tests. Like previous research, they found that increasing the number of items or the magnitude of true (multivariate) change improved power. In a real data example, using five dimensions from an elementary school mathematics assessment, they also confirmed that both the score test and LRT worked well in detecting significant change, especially compared to other tests. The multidimensional case was further extended by Wang, Weiss, & Suen (2021) who examined change across two and four occasions for three dimensions. Unlike Wang and Weiss, 2018, they examined Likert-type items and used a multidimensional GRM for item bank calibration. Their findings largely supported previous work, showing that

AMC hypothesis tests perform well for change magnitudes near 1.0 and above, and when more change is present overall (across multiple traits). They also found that the LRT provided the best balance of power and Type I error control. In real applications, the multidimensional AMC framework has been used successfully to identify significant changes in patient-reported outcomes in hospital patients (Weiss et al., 2021).

Purpose

Although designed for assessing traits measured at two or more occasions, the AMC CI method and hypothesis tests relate to the question posed by Anastasi and Urbina (1997), who asked whether scores from a multiscale battery—administered on one occasion—could be compared to one another using CTT-based confidence bands. Then, one novel approach for interpreting individual score profiles could be to apply the AMC hypothesis testing framework to multiscale assessments. Specifically, rather than administering one CAT on each occasion to measure a single trait across time, multiple CATs could be administered to measure multiple traits at one occasion. After administering CATs, AMC hypothesis tests could be applied to detect significant differences among traits. Using IRT and CAT-based scores could achieve better measurement precision, which could advance the individualized comparison of traits or the identification of unique sets of score profiles. Moreover, using CAT could improve test efficiency by reducing the number of items administered to any one individual.

Multiscale assessments could benefit from an AMC-based approach. For example, if one individual's score on a measure of Extraversion differs significantly from the same individual's score on a measure of Agreeableness, knowing whether these two personality traits differ significantly could provide useful information about an individual's personality tendencies. Extending the concept to all Big Five traits, understanding how such traits differ within a single person could be used to identify specific subpopulations, particularly given the intra-individual nature of the AMC approach. Such subpopulations could then be linked to further outcomes. In educational assessment, understanding whether a student's scores in subjects like science and mathematics differ from their score in reading could be useful for prioritizing subjects that a single student needs to study more. In the clinical domain, knowing if multiple scores near a threshold, such as a T-score of 65 on the MMPI, significantly differ could be used to inform diagnoses by clinicians after a single administration of a clinical assessment. The new procedure, called *adaptive profile difference analysis* (APDA), could provide an additional tool for interpreting multiscale assessments, potentially improving the interpretation of individual score profiles across multiple domains.

Overview

This study examined the APDA procedure, extending the AMC framework to cases with multiple traits within a multiscale assessment administered at a single occasion. A Monte Carlo simulation study was used to evaluate the effectiveness of the APDA procedure across multiple conditions. First, like previous studies, both magnitude and test length were examined with the expectation that greater magnitudes and longer tests would provide better power and better control of Type I error. Second, the number of traits in the multiscale assessment were examined. Previous studies focusing on change have evaluated up to five dimensions; therefore, the current study examined 5, 11, and 18 dimensions. Of particular interest, then, were the power and Type I error rate for 11 and 18 dimensions. The expectation was that the presence of more dimensions would make it easier for AMC omnibus hypothesis tests to detect at least one difference, but whether the Type I error rate remains controlled has not been studied. Third, the number of true differences among scales was evaluated. For example, for $T = 5$ traits, there would be $H = 10$ pairwise comparisons. A low number of true differences could be set at four (i.e., 40% of possible comparisons differ) while a high number could be set at six (i.e., 60% differ). Fourth, in line with previous research, the performance of FL-CATs against VL-CATs was evaluated to determine the efficiency gained by using VL-CATs. Fifth, the Z-test, score test, and LRT were compared to one another in terms of performance: the expectation was that the LRT would outperform the other tests, which would coincide with previous research. Because AMC tests depend on information, and because information differs across the latent trait continuum, intra-individual differences were examined at areas of low, medium, and high information. For instance, for most conventional tests, areas of high information tend to occur near zero on the latent trait continuum with information decreasing near the extremes, corresponding to poorer measurement precision in the extremes of a trait. The expectation was that differences occurring at high information areas—when both trait estimates are closer to zero, for example—would be easier to detect. Two types of item banks were also examined: realistic and ideal. The realistic item bank contained item discriminations based on real data while the ideal bank contained discriminations likely to occur when a multiscale assessment is developed using IRT. Finally, two additional types of item banks were examined: realistic or flat. The realistic item bank contained item location parameters based on real data while the flat bank contained item location parameters that were spread evenly across the latent trait continuum.

Importantly, APDA examines intra-individual differences and is not group-based. However, pre-calibrated item banks must exist for multiple CATs to be delivered to a single individual (i.e., one CAT for each scale in the assessment). To make the simulation study more realistic, item parameters for simulated item banks were informed by fitting a series of unidimensional IRT models to scales from three multiscale personality assessments. Assessments included (1) a 50-item assessment of the Big Five with five scales based on the International Personality Item Pool (IPIP) and an internet-based sample (Goldberg, 1992); (2) a 300-item version of the Multidimensional Personality Questionnaire with a total of 11 primary scales, based on a sample from the Minnesota Center for Twin and Family Research (Tellegen & Waller, 2008; Iacono & McGue, 2002); and (3) a 180-item assessment of the metatrait of Stability with a total of 18 scales, also based on the IPIP and an internet-based sample (Johnson, 2014). The 2PL model was used to calibrate MPQ items, which are binary; and the GRM was used to calibrate items for the remaining two assessments.

Chapter 2. Method

Informative Real Data Analyses

Goldberg's Big Five Personality Measure. The Big Five personality traits represent the most studied set of broad personality dimensions in psychology (Widiger, 2017). Goldberg (1992) developed Big Five factor markers to construct an openly available, CTT-based analog to the NEO Personality Inventory (NEO-PI; Costa & McCrae, 1985), which measures the Big Five. Goldberg created a version of this multiscale assessment by giving a large group of items to a community sample of adults, and selecting sets of 10 items that achieved classical reliabilities between about 0.8 and 0.9 for each scale. Goldberg showed that the assessment—referred to here as Goldberg's Big Five Measure (GBFM)—corresponds well to the NEO-PI (Goldberg, 1992). Overall, Goldberg's assessment totals 50 Likert-type items.

The Open Source Psychometrics Project (OSPP) hosts an internet-based administration of the GBFM. In addition to the GBFM, OSPP has administered many multiscale assessments across various domains and makes their data freely available. OSPP data sets have been used in numerous publications in statistics, psychometrics, and quantitative psychology (recent examples include Liu et al., 2021; Chen, Culpepper, & Liang, 2020; and Schneider et al., 2020). To complete an assessment, participants navigate to the project's website, answer items for an assessment of their choosing, and receive a report about their results. Individuals taking the GBFM read a series of 50 personality statements (e.g., "I am the life of the party.") and rate how true each statement is for them on a Likert scale with five ordered response options. The response options include: "Disagree", "Slightly disagree", "Neutral", "Slightly agree", and "Agree". About half the items are negatively keyed. The population of individuals seeking such tests is likely biased; for example, people who are higher on Openness to Experience might be more likely to complete such assessments. However, the purpose of fitting IRT models in this study was to acquire realistic item parameters, not to make inferences about the general population. Thus, in this study, some bias in the sample was deemed acceptable.

To inform simulated item banks, the 50-item GBFM from the OSPP was analyzed using unidimensional IRT. Five graded response models (GRMs) were fit, one to each of the Big Five scales. Participants completed the GBFM between 2016 and 2018 on the OSPP website. The available sample size was about one million but was reduced through data cleaning. Specifically, internet-based samples typically include invalid or missing data. To clean these data, the following procedures were implemented. First, if a participant had any missing values on any items, the participant was excluded. Second, any participants with a non-mixed response vector were excluded to remove careless or extreme response patterns. Third, an overall assessment completion time was recorded by OSPP. Participants who took less than 1.5 minutes to complete all 50 items, or who took more than 75 minutes, were excluded; these exclusions removed individuals who moved through the assessment too quickly or who might not have completed the assessment in one sitting. Fourth, the OSPP records the response time for each GBFM item (i.e., the time between when a new item appears on a screen and when the participant selects a response). Participants averaging less than two seconds per item (i.e., less than 100 seconds on the assessment overall) or who averaged more than 90 seconds per item (i.e., more than 4,500 seconds overall) were excluded. Some participants moved rapidly through the assessment but waited at one or more questions; moreover, some participants with means above 90 seconds might have some items with extremely long response times, which could indicate that attention was not directed towards the assessment. Using this exclusion criterion eliminated such participants.

In addition to assessment timing, the OSPP records the number of GBFM assessments that have been administered to any given IP address. A participant with a value of one on this indicator would be the only person to have completed the GBFM from their IP address. To reduce participants who might have taken the assessment more than once, participants were only included if the IP address associated with a participant appeared only once in the data set. Furthermore, OSPP approximates a participant's country of origin using their IP address. Only individuals from the United States (US) were included, which reduced the presence of cross-cultural effects that might have made the analysis more complex. After these exclusions, a random sample of 50,000 was chosen from the remaining subset. Then, each participant's mean response for each scale was calculated, where categories were represented by integer weights such that "Disagree" is 1; "Slightly disagree" is 2; etc. Next, each participant's set of five mean responses, among the five scales, was summed. The lowest possible value would be just above 5 while the highest would be just below 25. Individuals with summed means below 6 (e.g., 5.4) or above 24 (e.g., 24.7) were eliminated. This removal screened extreme respondents. Finally, from the remaining participants, a random sample of 10,000 was chosen. On the final sample of 10,000, an IRT analysis was conducted to obtain realistic parameters for item banks, with a separate GRM model fit for each of the five scales.

Multidimensional Personality Questionnaire. Tellegen (1982) introduced the Multidimensional Personality Questionnaire (MPQ) to measure a set of broad personality traits. The MPQ includes 11 primary scales consisting of binary items indicating whether a respondent agrees or disagrees with a personality statement. Compared to the NEO-PI, the MPQ focuses on a lower level of personality (Tellegen & Waller, 2008). The original version of the MPQ consists of primary scales with between 20 to 34 items, and the assessment totals 300 items. A sample of individuals from the Minnesota Center for Twin and Family Research (MCTFR) completed the MPQ in the 1980s using an in-person, paper-and-pencil version of the assessment. A series of 2PL IRT models have been fit to the 11 primary scales of the MPQ for this sample previously (Reise & Waller, 1990; see "Introduction"). 6,371 respondents were available from this sample. To inform simulated item banks, the MPQ from the MCTFR sample was analyzed. Individuals with missing data were removed; but because the data from this sample were mostly clean, no other steps were implemented to clean data further. A total of 5,432 respondents remained after cleaning. As in Reise & Waller, 1990, a 2PL model was fit to these data for each of the 11 primary scales to obtain estimated item parameters, totaling eleven 2PL models.

Goldberg's IPIP-300. Goldberg (1992, 1999) developed the International Personality Item Pool (IPIP) to construct openly available, CTT-based analogs to popular personality inventories. The IPIP-300 is one such inventory—measuring the five broad, normal-range personality factors of the Big Five. Goldberg's assessment consists of 300 items with individual scales each containing 60 Likert-type items. Thus, the IPIP-300 can be considered as a longer version of the GBFM. In addition to the Big Five primary scales, the IPIP-300 contains six facet scales for each of the Big Five. Goldberg (1999) showed that facet scales have classical reliabilities ranging from 0.7 to 0.9. Using a large, internet-based sample, Johnson (2014) found the same range of reliabilities for the 30 facet scales of the IPIP-300.

To inform simulated item banks, 180 items from the IPIP-300 were analyzed using openly available data from Johnson (2014). The available sample contained data for 307,313 participants who completed the IPIP-300. The 180 items correspond to the 18 facet scales associated with Agreeableness, Conscientiousness, and Neuroticism. These three traits constitute Big Five traits associated with the metatrait of Stability (DeYoung, 2015, p. 47). Then, this assessment constituted a multiscale assessment of the metatrait of Stability. For this assessment, cleaning steps differed from those used for the GBFM, as available secondary variables (e.g., response times) differed. Specifically, as in the GBFM analysis, only individuals from the US were included; and individuals with any missing values for any items were removed. Moreover, individuals with non-mixed response vectors were removed. However, information on response times was not available for the Johnson (2014) sample and was not used to screen participants. Further, unlike the GBFM analysis, individuals with extreme summed mean scores were not eliminated in order to keep more of the data, given that the sample size was about 30% of the GBFM sample size. A random sample of 10,000 was drawn from the remaining data, as done in the GBFM analysis. Using the final sample, an IRT analysis was conducted to obtain realistic item parameters with 18 GRM models fit, one to each scale.

Software for Real Analyses. For any multiscale assessment, IRT models were fit using XCalibre 4.2 (Guyer & Thompson, 2011), with one model fit for each scale. Negatively keyed items were identified and marked for the software to handle by reversing item categories’ indices. For both the GRM and the 2PL, the model was fit with a constant of $D = 1.7$ with boundary or difficulty parameters set to an allowable range of $b_{jm} \in [-4, 4]$. An E-M algorithm implemented marginal maximum likelihood estimation of item parameters, using 40 quadrature points and a maximum of 200 iterations with a convergence criterion of 0.001. After convergence, item response functions were evaluated within each scale to ensure they operated as expected. Estimated item parameters served as a basis for forming CAT item banks (see "Item Banks" section below).

Person Scores. Weighted maximum likelihood estimation (WLE) was used to score individuals on all scales for the GBFM model calibration sample. For each person, all pairwise difference scores were calculated using WLE scores; and the median absolute difference was calculated (i.e., the median among the absolute values of pairwise difference scores). Across all persons, percentiles of these medians were calculated to determine how large intra-individual differences typically were in the real data.

Item Banks

Estimated item parameters from these models served as a basis for simulating item banks. For the simulation study, the number of traits included 5, 11, or 18 scales, resembling the three real analyses. Moreover, the simulated item banks each had a bank information function (BIF), which is the sum of all item information functions (IIFs) in the bank. Each bank contained 100 items, which aligns with previous research where about 100 items were used per latent dimension (Wang, Weiss, & Suen, 2021). For any one scale, for any type of item parameter (e.g., discrimination a_i), descriptive statistics were calculated across that scale's estimated parameters including: the mean, the variance, the minimum, the maximum, the 25th percentile, and the 75th percentile. Using these descriptive statistics, three types of distributions were formed to create the simulated item banks: (1) a normal distribution with mean and variance equal to the calculated mean and variance for that parameter across items in one scale; (2) a uniform distribution with endpoints equal to the calculated minimum and maximum; and (3) a uniform distribution with endpoints equal to the 25th and 75th percentiles. For any scale, such distributions were formed for all five GRM item parameters (e.g., for five response options, parameters were a_i, b_{i1}, b_{i2}, b_{i3}, and b_{i4}) for the 5-scale or 18-scale conditions, or for both of the 2PL's item parameters (i.e., a_i and b_i) for the 11-scale condition. Using the three types of distributions, three types of item banks were simulated for each scale.

For the three types of banks, the simulated BIF whose shape best resembled its corresponding real test information function (TIF) was selected as the best item bank (i.e., the bank whose information most closely matched the measurement properties of its associated real scale). To determine the best item bank and distribution type, therefore, a comparison was performed between the simulated BIFs and the real TIFs for each scale. Because banks each had 100 items, and because real analyses were performed on scales with much less than 100 items, simulated banks had more information than their real counterparts. To account for this difference in comparisons, relative information functions (RIFs) were calculated. Specifically, the real TIF for each of the scales was evaluated at values of $\theta = -4$ to $\theta = 4$ in increments of 0.05. The maximum information value across this range was calculated for a given scale, and all information values for that scale were divided by the maximum value to produce a real RIF with values between 0 and 1, inclusive. Note, then, that a RIF provides the amount of information available along the θ continuum with respect to the point along the θ continuum with the highest amount of information. The same process was applied to all three of a scale's simulated BIFs. Among the three options, the simulated RIF that best resembled its corresponding real RIF was selected.

GRM boundary parameters should always increase in value across the θ continuum. Because boundary parameters were simulated randomly from distributions, some simulated boundary parameters did not increase for some items. This outcome occurred the least using the uniform distribution with 25^{th} and 75^{th} percentile endpoints because such endpoints removed more extreme boundary parameters and reduced overlap between boundaries. In these cases, boundary parameters that overlapped were modified to be less than (or greater than) the next boundary parameter to ensure that boundary parameters always increased. While this step introduced some bias, the bias was negligible as only a small number of items exhibited this phenomenon.

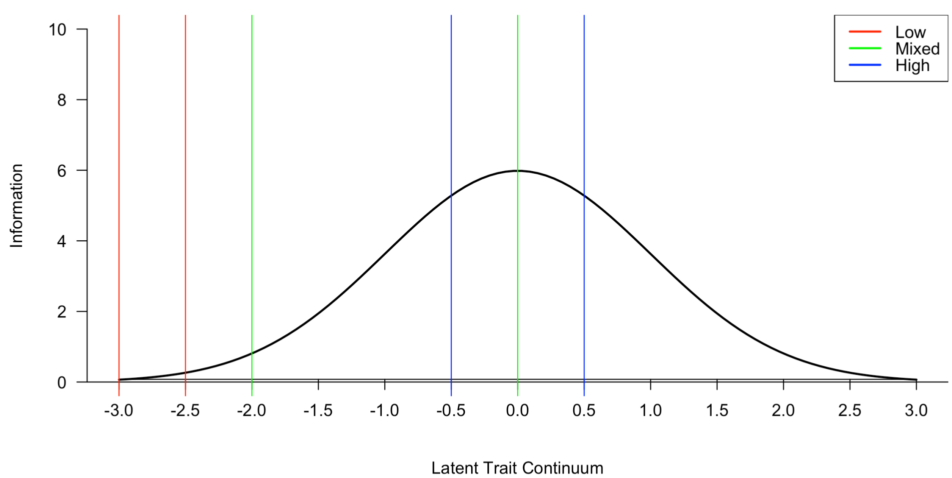
Simulation Design

To evaluate adaptive profile difference analysis (APDA), a Monte Carlo simulation study was conducted and consisted of nine design factors. First, because previous AMC studies examined magnitude of change, the magnitude of intra-individual differences was examined. In line with previous research on change (e.g., Phadke, 2017), difference magnitudes of $d = 0, 0.5, 1, 1.5$ were used. Second, test length was examined. For fixed-length computerized adaptive tests (FL-CATs), test lengths of $I = 10, 30, 50$ items were used. The values of 30 and 50 aligned with previous research exactly while the value of 10 was close to previous studies, which used 15 for the lowest test length (Lee, 2015). Because most scales in the informative real data analyses contained 10 items, using 10 (instead of 15) better reflected these scales. Third, FL-CATs were compared to variable-length CATs (VL-CATs). For VL-CATs, IRT-based standard error of measurement, $\mathrm{SE}(\hat{\theta})$, was used instead of test length as a criterion for terminating the CAT. Specifically, for any single simulee, that simulee's score estimate must achieve a particular degree of measurement precision for a VL-CAT to terminate. Forcing the CAT to achieve a specific precision for all simulees would correspond to real-world applications where all examinees should, if possible, have the same measurement precision for purposes of fairness or comparability of examinees to one another. Values of $\mathrm{SE}(\hat{\theta}) = 0.4, 0.3, 0.2$ were used for VL-CAT termination. Two constraints were also used for VL-CATs: the minimum test length was set at five items while the maximum was set at 99 items (one less than the number of items in the bank).

The fourth factor was scale location. For any two MLE score estimates $\hat{\theta}_{j1}$ and $\hat{\theta}_{j2}$, each exists at a point along the θ continuum with a particular amount of information at that point. As displayed in Figure 4, for any two estimated scores, within a single simulee, the scores could both exist in an area of low information (vertical red lines in the figure); high information (blue lines); or one from each area, called mixed information (green lines). Scores for each of the three types of pairs were examined. To determine a threshold for low versus high areas of information, the 34 RIFs from the real data analyses were averaged to create an average real RIF. The point along the averaged RIF's θ continuum with the highest information was set as the center point. The center point operated as the most typical location along θ where maximum information was achieved in the real analyses, across all scales. The distance from this center point to the maximum value for each real RIF was calculated, for all 34 RIFs. The average distance was used to move outward from the center point to mark a lower θ value and an upper θ value that could serve as general cutoffs for areas of low versus high information. This criterion guided selection of a θ cutoff, but cutoffs were selected on each end of the continuum that were equal in absolute value for simplicity. This threshold identification method worked because the real scales being examined were developed using conventional, CTT-based analyses; so CTT-based precision amassed in the center of the θ continuum for the real data and decreased moving toward the extremes symmetrically on both ends. However, there was some variation in real TIFs and RIFs. This method served as an approximation but was limited for some TIFs as they were not entirely middle-peaked and symmetrical around the middle.

Figure 4

Example Score Locations on a Test Information Function: Both in Low, in High, or in Mixed Information Areas



Fifth, the number of scales in each multiscale assessment was examined. Specifically, multiscale assessments with $T = 5, 11, 18$ scales were used, which matched the number of scales available in the real multiscale assessments. The sixth design factor followed from the fifth: for any single simulee, the number of intra-individual differences among scales was examined. For example, for $T = 5$, a total of $H = T(T-1)/2 = 10$ pairwise comparisons exist. For $T = 11$, $H = 55$; and for $T = 18$, $H = 153$. Two levels were used: approximately 40% of pairwise comparisons differed, or approximately 60% differed. Constraints from other design factors limited the exactness of achieving 40% or 60% differences, but the closest value was sought. For example, two trait scores might need to differ by 1.5 (the magnitude factor) but might also both exist in areas of high information (i.e., the scale location factor). Moving outward from $\theta = 0$, two possibilities are $\theta_1 = -0.75$ and $\theta_2 = 0.75$, assuming, for example, the high-low information threshold is set at $|\theta| < 1$. Then, for $T = 5$ trait scores, with 40% of pairwise comparisons differing, the remaining three trait scores must also be set at $\theta_3 = \theta_4 = \theta_5 = -0.75$. In this case, 40% is achieved with θ_2 being different from the other four scores. However, for $T = 11$ and $T = 18$, the situation becomes more complex.

The seventh design factor was the AMC omnibus hypothesis test. The Z-test, the score test, and the likelihood-ratio test (LRT) were used. Phadke (2017) noted that Z-test observed statistics are not strictly independent; then, when they are squared and summed to form a chi-square distributed variable, the variable is only approximately chi-square distributed. This violation might become more evident as the number of Z observed statistics increases. Phadke (2017), for example, examined up to three statistics associated with $T = 3$ occasions with $H = 3$ pairwise comparisons; but for cases with $H = 55$ or

$H = 153$ comparisons (numbers never examined previously with the Z-test), the Z-test might have inflated power.

The eighth design factor was the use of different distributions of item discriminations for simulated item banks. Two conditions were used: realistic or ideal. The realistic condition used the item discriminations simulated using the real data analyses (i.e., using the type of distribution, such as normal or uniform, that aligned best with real RIFs). The ideal condition replaced these discriminations with higher discrimination values. Specifically, if either of the uniform distribution types provided the best item banks, ideal item discriminations would be simulated from a uniform distribution with endpoints of 1.1 and 2.8, which matched the distribution used to simulate discriminations in Wang, Weiss, and Suen (2021). Otherwise, if the normal distribution provided the best item banks for the simulation, ideal item discriminations would be simulated from a normal distribution with a mean of 1.5 and a variance of 0.15, as done in Phadke (2017). Ideal discriminations might be expected when using IRT to develop a multiscale assessment from the beginning of the development process.

The ninth factor was the use of realistic or equally spread location parameters for simulated item banks. The realistic condition used the item locations (i.e., difficulties/locations for the 2PL or boundary parameters for the GRM) simulated using the informative real data analyses. The condition with equally spread locations replaced realistic location parameters with parameters that tended to produce equal measurement precision across the latent trait continuum for an item bank. In particular, for the GRM that models items with five response categories, the four boundary parameters were simulated from uniform distributions (symbolized with "U") with the following endpoints: U[−3, −1.5001], U[−1.5, −0.001], U[0, 1.5], and U[1.5001, 3]. These distributions correspond closely to those used by Wang, Weiss, and Suen (2021)—except four parameters would be simulated for the current study instead of three parameters simulated in Wang, Weiss, and Suen (2021).

In summary, the design factors included the following: (1) magnitude of differences with four levels: 0, 0.5, 1, or 1.5; (2) FL-CAT versus VL-CAT; (3) test length with three levels of 10, 30, or 50 items, or $\mathrm{SE}(\hat{\theta})$ with three levels of 0.4, 0.3, or 0.2; (4) scale location with three levels: low, mixed, or high information locations; (5) the number of scales in a multiscale assessment with three levels: 5, 11, or 18; (6) the number of differences with two levels: approximately 40% or 60%; (7) AMC hypothesis tests with three levels: the Z-test, score test, and LRT; (8) item discrimination type: realistic or ideal; and (9) boundary/location type: realistic or equal. Note that the number of differences factor cannot be crossed with a difference magnitude equal to zero because no differences exist. Moreover, the scale location condition "mixed" cannot be crossed with a difference magnitude equal to zero because, for two scores to be in areas of differing information, they would have to differ. In total, then, the nine design factors yielded 4,320 conditions. For each condition, responses for 1,000 simulees were generated.

Two main outcome measures were used to evaluate the performance of APDA across the nine design factors. The main outcomes included: (1) the true positive rate (i.e., TPR; the proportion of true differences over the total number of detected differences; also called "power"); and (2) the false positive rate (i.e., FPR; the proportion of detected differences that are not significant over the total number of detected differences; also called "Type I error"). For the VL-CAT condition, two indices were also recorded. For each simulee, the number of items needed to terminate the CAT for each scale was recorded, and the mean across scales was calculated for each simulee. The mean among all simulee means served as the average number of items required to terminate within that condition. The standard deviation among all simulee means was also calculated to determine how much variability existed among different simulees in the average number of items needed for termination.

Trait Scores for Simulation

For each of the 4,320 simulation cells, random model-fitting response patterns were simulated using a specific set of trait scores. For the condition with $T = 5$ scales, the true θ scores used to simulate response patterns are displayed in Table 1, shown across multiple combinations of conditions. Trait scores for $T = 11$ scales and for $T = 18$ scales are provided in Appendix A. For any row of Table 1, 1,000 random response patterns were simulated using the appropriate IRT model, either the GRM or the 2PL. Specifically, the probability of selecting each option was calculated using item bank parameters for the given model and using the true θ value. Then, a deviate was generated randomly from a uniform distribution with endpoints of zero and one. The probability of selecting each option was compared to the random uniform deviate. For the GRM, the cumulative probabilities were calculated and used for this comparison, such that a response option was selected if the uniform deviate was within its cumulative range. As an example, for a Likert-type item with five options, the probabilities of selecting each of those options might be 0.1, 0.2, 0.4, 0.2, 0.1, for the first through fifth response options, respectively, where moving toward the fifth option represents the keyed direction. If a generated uniform deviate is 0.3, the first comparison is $0.3 < 0.1$. This inequality is false. The next comparison is $0.1 \leq 0.3 < (0.1 + 0.2) = 0.3$. This statement is also false. Next, the comparison is $0.3 \leq 0.3 < (0.1 + 0.2 + 0.4) = 0.7$. This statement is true, so the simulee would be recorded as selecting the third option for this Likert-type item, receiving an integer weight of 3. For the 2PL, the comparison is easier. If the probability of the keyed response is greater than the uniform deviate, the simulee receives the keyed response (i.e., an integer weight of 1) and, otherwise, receives the non-keyed response (i.e., a weight of 0). For each simulee, this process was repeated for all items in all item banks across all scales (e.g., $T = 11$ scales). Trait scores were shuffled randomly among the traits to prevent any single trait from impacting results unequally. Moreover, half of the simulees were those with trait scores displayed in the table while half had their scores multiplied by -1 to ensure both extremes of the latent trait continuum were included equally in true score patterns. Note that the full set of analyses suggested the scale location threshold should be a value of 1 to define locations with high, low, or mixed amounts of information (see more details in "Results").

Table 1. Simulation Trait Scores for Assessment with Five Scales.

Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Magnitude	Location	Number of Differences
0	0	0	0	0	0	High	0
2	2	2	2	2	0	Low	0
0	0	0	0	0.5	0.5	High	4
0.75	0.75	0.75	0.75	1.25	0.5	Mixed	4
1.5	1.5	1.5	1.5	2	0.5	Low	4
–0.75	–0.75	–0.75	–0.75	0.25	1	High	4
0.75	0.75	0.75	0.75	1.75	1	Mixed	4
1.5	1.5	1.5	1.5	2	1	Low	4
–0.75	–0.75	–0.75	–0.75	0.75	1.5	High	4
0.75	0.75	0.75	0.75	2.25	1.5	Mixed	4
1.5	1.5	1.5	1.5	3	1.5	Low	4
0	0	0	0.5	0.5	0.5	High	6
0.75	0.75	0.75	1.25	1.25	0.5	Mixed	6
1.5	1.5	1.5	2	2	0.5	Low	6
–0.75	–0.75	–0.75	0.25	0.25	1	High	6
0.75	0.75	0.75	1.75	1.75	1	Mixed	6
1.5	1.5	1.5	2.5	2.5	1	Low	6
–0.75	–0.75	–0.75	0.75	0.75	1.5	High	6

Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Magnitude	Location	Number of Differences
0.75	0.75	0.75	2.25	2.25	1.5	Mixed	6
1.5	1.5	1.5	3	3	1.5	Low	6

Computerized Adaptive Tests

CATs were administered to each simulee using that simulee's simulated item bank responses. For a given simulee and scale, the CAT began with the selection of the bank item that maximized unweighted Fisher information. An initial starting value of $\theta = 0$ was used. MLE was used to score the simulee after each item was given, and unweighted Fisher information was again used to select subsequent items. For the first five items, θ was restricted to values between -1 and 1; after the first five items, the range of possible values for θ was expanded from -6 to 6. For FL-CATs, the CAT terminated at a fixed test length. For VL-CATs, the CAT terminated either at a fixed $\mathrm{SE}(\hat{\theta})$ or at the maximum number of items allowed (i.e., 99 items out of the 100 items in the bank).

Simulation Data Analysis

For the TPR, an eight-way analysis of variance (ANOVA) was used to model TPR variance. A separate ANOVA was used for each of the three AMC hypothesis tests. Main effects included the simulation design factors (every factor except the type of hypothesis test). In addition, for the FPR, a six-way ANOVA was used to model FPR variance for each hypothesis test where magnitude and the number of differences were both excluded, given that FPR corresponds to a magnitude of zero and to no differences. Finally, for the mean number of items, a seven-way ANOVA was used to model variance where the type of CAT factor was excluded, given that the number of items only varied for VL-CATs. For all ANOVA models, all two-way and three-way interactions were included in the model. Note that test length and the standard error termination value were nested within the type of CAT—either FL-CAT or VL-CAT, respectively.

Because the number of simulees could be increased arbitrarily to produce statistically significant effects, only effect sizes were examined for each main effect and interaction in each ANOVA model. Specifically, the classical effect size measure of η^2 (i.e., eta-squared) was used where η^2 is the ratio of the sum of squares between groups (i.e., for a specific effect or interaction) over the total sum of squares for the model (i.e., across all effects and interactions). This effect size measured the relative importance of each main effect or interaction and was used to interpret results.

Software Implementation

To clean real data, conduct the simulation study, and analyze results, the R programming language was used (R Core Team, 2024). Custom functions were written to calculate item information and to conduct AMC hypothesis tests. Custom scripts were created to implement the simulation. In addition, the R package **catIrt** was used to simulate response patterns for simulees and to administer CATs (Nydick, 2022). ANOVA models were fit in R using the R function aov(); and η^2 values were calculated with Type II Sums of Squares, which was implemented through the R package **car** (Fox & Weisberg, 2019). Simulation R scripts and functions written in R are provided in Appendix B so that others can reproduce the simulation's results.

Post-Simulation Real Data Analysis

After the simulation study, APDA was applied to a holdout sample from the GBFM data set. Specifically, from the remaining individuals in the GBFM data set, a separate sample of 10,000 was randomly drawn (i.e., this sample did not overlap with the calibration sample from the informative real analysis). Weighted maximum likelihood estimation was used to estimate trait scores for these individuals, and pairwise difference scores were calculated for each individual (10 in total per person). Individuals were grouped based on the median absolute difference among their difference scores. In particular, individuals with median absolute differences below 0.4 were included in a group corresponding to a magnitude of zero; the value of 0.4 was used to ensure enough individuals were available. Compared to the magnitude of the zero group, more individuals became available as the median absolute difference increased, which allowed for a smaller inclusion interval. For the small magnitude group, individuals with median absolute differences between 0.4 and 0.6 were included; for medium magnitudes, those between 0.9 and 1.1 were included; and for large magnitudes, those between 1.4 and 1.6 were included. AMC hypothesis tests was applied to each person within these groupings to determine the number of individuals demonstrating a psychometrically significant difference among their Big Five trait score estimates. These detection rates were compared to a specific set of simulation conditions: (1) a test length of 10 items; (2) a FL-CAT; (3) realistic discriminations; (4) realistic location parameters; (5) 5 scales; and (6) a high number of differences (i.e., about 60%). This sixth condition was used because the trait score pattern used for a "high" number of differences (i.e., near about 60%) for 5 scales produced a median absolute difference of exactly 0.5, 1, or 1.5. (see Table 1). Moreover, enough individuals were not available to disaggregate by scale locations with high, mixed, and low information in the holdout sample; thus, when comparing real detection rates to the simulation's results, simulation results based on scale location were averaged together. In the analysis, two comparisons were of interest: first, comparing the detection rate for individuals with near-zero median absolute differences to the FPR for the simulation condition with a magnitude of zero; and second, comparing detection rates for small, medium, and large magnitudes to TPRs calculated in the simulation study for their corresponding conditions.

Chapter 3. Results

Informative Real Data Analyses

Item Analyses. Appendix C contains the observed, relative test information functions (RIFs) for all three multiscale assessments across different values of θ. Each RIF has a θ value at which the maximum amount of information is present. Averaging across all 34 RIFs at each θ value produced an average RIF. The average, maximum relative information occurred at $\theta = 0.05$. The standard deviation among relative information maxima across all 34 scales was SD = 0.8831. One standard deviation below 0.05 occurred at $\theta = -0.8331$ while one standard deviation above occurred at $\theta = 0.9331$. Respectively, these values round to -1 and 1. Then, a reasonable threshold for low-high information—for the simulation factor of scale location—was set at $\theta = 1$ where $|\theta| < 1$ indicated high information and where $|\theta| \geq 1$ indicated low information.

Table D1 in Appendix D provides estimated item parameters for each GBFM scale. In addition, Table 2 provides means and standard deviations for item parameter estimates for each GBFM scale. Average discriminations ranged from 0.824 (Openness to Experience) to 1.187 (Extraversion). Average boundaries showed that Openness to Experience items tended to be centered somewhat above $\theta = 0$. As displayed in Table C1 of Appendix C, maximum information for observed RIFs occurred at locations from $\theta = -0.10$ (Neuroticism) to $\theta = 1.90$ (Openness to Experience).

Table 2. Descriptive Statistics for Estimated Item Parameters for Scales from Goldberg's Big Five Measure Assessment.

Scale	Statistic	Discrimination	Boundary 1	Boundary 2	Boundary 3	Boundary 4
Openness	Mean	0.824	–0.525	0.979	2.178	3.456
	SD	0.221	0.498	0.595	0.715	1.012
Conscientiousnes s	Mean	0.85	–1.238	0.17	1.271	2.587
	SD	0.207	0.38	0.661	1.012	1.363
Extraversion	Mean	1.187	–1.501	–0.449	0.348	1.312
	SD	0.212	0.466	0.471	0.479	0.551
Agreeableness	Mean	1.032	–0.637	0.677	1.589	2.724
	SD	0.378	0.339	0.151	0.29	0.638
Neuroticism	Mean	1.053	–1.498	–0.328	0.474	1.591
	SD	0.28	0.544	0.518	0.522	0.572

Table D2 in Appendix D provides estimated item parameters for each MPQ scale. Moreover, Table 3 provides descriptive statistics for estimated item parameters for each scale. Average discriminations ranged from 0.676 (Traditionalism) to 0.982 (Wellbeing). Item difficulties/locations showed Alienation and Aggression items as being located toward the lower range of the θ continuum while Harm Avoidance, Traditionalism, and Wellbeing items were located toward the upper range. As shown in Tables C2–C4 (Appendix C), maximum information for observed RIFs was observed from $\theta = -1.75$ (Alienation) to $\theta = 1.25$ (Wellbeing).

Table 3. Descriptive Statistics for Estimated Item Parameters for Scales from the Multidimensional Personality Questionnaire Assessment.

Scale	Statistic	Discrimination	Difficulty/Location
Absorption	Mean	0.692	–0.192
	SD	0.125	0.803
Achievement	Mean	0.760	0.362
	SD	0.219	0.693
Aggression	Mean	0.817	–1.378
	SD	0.268	0.675
Alienation	Mean	0.844	–1.771
	SD	0.199	0.472
Control	Mean	0.765	0.765
	SD	0.256	0.695
Harm Avoidance	Mean	0.775	1.039
	SD	0.195	0.525
Social Closeness	Mean	0.868	0.813
	SD	0.264	0.859
Social Potency	Mean	0.952	–0.545
	SD	0.280	0.618
Stress Reaction	Mean	0.924	–0.298
	SD	0.205	0.472
Traditionalism	Mean	0.676	1.014
	SD	0.230	0.745
Wellbeing	Mean	0.982	1.098
	SD	0.293	0.532

For the 18 Big Five facets associated with Stability, Table D3 in Appendix D displays estimated item parameters. Table 4 provides descriptive statistics. Average discriminations ranged from 0.758 (Sympathy) to 1.428 (Anger). Average boundaries indicated that Altruism, Self-Efficacy, Dutifulness, and Achievement-Striving items were shifted lower while Vulnerability items were shifted higher on the θ continuum. Finally, as shown in Tables C5–C8 in Appendix C, Stability RIFs had maxima at locations from $\theta = -1.65$ (Dutifulness) to $\theta = 0.70$ (Vulnerability).

Table 4. Descriptive Statistics for Estimated Item Parameters for Scales from the Stability Assessment.

Scale	Statistic	Discrimination	Boundary 1	Boundary 2	Boundary 3	Boundary 4
Trust	Mean	1.158	–2.172	–0.817	–0.019	1.498
	SD	0.382	0.537	0.539	0.495	0.568
Morality	Mean	0.943	–2.974	–1.442	–0.551	0.757
	SD	0.437	0.763	0.711	0.725	0.924
Altruism	Mean	0.957	–3.315	–1.895	–0.982	0.652
	SD	0.243	0.336	0.282	0.362	0.476
Cooperation	Mean	0.833	–2.555	–0.984	–0.156	1.262
	SD	0.314	1.168	0.766	0.667	0.873
Modesty	Mean	0.918	–3.425	–0.874	0.571	2.708
	SD	1.11	2.004	1.349	1.308	1.743
Sympathy	Mean	0.758	–2.854	–1.23	–0.215	1.485
	SD	0.269	0.651	0.68	0.766	0.909
Self-Efficacy	Mean	0.928	–3.438	–1.855	–0.896	1.024
	SD	0.235	0.375	0.347	0.398	0.446
Orderliness	Mean	1.052	–2.016	–0.735	–0.064	1.271
	SD	0.248	0.865	0.629	0.363	0.311
Dutifulness	Mean	0.854	–3.44	–1.947	–1.036	0.61
	SD	0.203	0.486	0.426	0.367	0.344
Achievement-Striving	Mean	0.971	–2.99	–1.559	–0.639	0.956
	SD	0.321	0.671	0.447	0.346	0.525
Self-Discipline	Mean	1.186	–1.909	–0.465	0.248	1.693
	SD	0.315	0.616	0.429	0.329	0.4
Cautiousness	Mean	1.074	–2.395	–0.686	0.259	1.906
	SD	0.623	1.537	0.92	0.585	0.825
Anxiety	Mean	1.069	–1.744	–0.34	0.259	1.673
	SD	0.252	0.413	0.587	0.68	0.818
Anger	Mean	1.428	–1.436	–0.237	0.348	1.598
	SD	0.413	0.458	0.451	0.533	0.618
Depression	Mean	1.357	–1.122	–0.098	0.506	1.62
	SD	0.375	0.453	0.355	0.34	0.303
Self-Consciousness	Mean	0.94	–1.592	–0.225	0.393	1.722
	SD	0.199	0.552	0.572	0.558	0.635
Immoderation	Mean	0.826	–2.37	–0.957	–0.161	1.457
	SD	0.297	0.877	0.597	0.454	0.439
Vulnerability	Mean	1.031	–1.329	0.193	0.893	2.173
	SD	0.251	0.497	0.583	0.614	0.674

Simulated Banks. Three methods were explored for generating item banks for the simulation study, depending on three possible sets of descriptive statistics of estimated item parameters. Statistics included: (1) mean and variance, (2) minimum and maximum, or (3) 25th and 75th percentile. Among the three methods, the 25th and 75th percentile method worked best as (1) it provided a good approximation to real RIFs and (2) reduced the number of instances where simulated boundary parameters were not ordered correctly (e.g., Boundary 1 was −2 and Boundary 2 was −3, even though boundary parameters must be in ascending order). Appendix E provides visualizations of observed TIFs for real scales (colored in red) and each corresponding simulated BIF (colored in purple) for the simulation condition with realistic discriminations and equally spaced location parameters. Note that, because the 25th and 75th percentile method worked best, ideal item discriminations were simulated from a uniform distribution with endpoints of 1.1 and 2.8, which matched the distribution used to simulate discriminations in Wang, Weiss, and Suen (2021).

As shown in Figure E1 of Appendix E, real RIFs for the GBFM showed high measurement precision near the center of θ continua; and precision remained somewhat high moving toward the extremes. Simulated bank RIFs closely matched real RIFs but generally provided slightly more information, especially in the extremes of the latent trait continuum. Comparatively, the real RIFs for the MPQ were more peaked in the middle (see Figures E2 and E3) and, similar to the GBFM assessment, simulated RIFs had slightly more precision across the θ continuum for each scale. Finally, Figures E4–E6 show real and simulated RIFs for the Stability facets of the Big Five. Most scales were middle-peaked; however, the left panels of Figure E5 and the upper-right panel of Figure

E6 display scales that were shifted to the left on their respective latent trait continua. Moreover, the scale in the bottom-center of Figure E6 showed some inconsistency in precision where it is peaked. As with the GBFM and MPQ, simulated RIFs approximated real RIFs but provided slightly more precision across the θ continuum for each scale.

Individual Scores. Weighted maximum likelihood estimation was used to estimate examinee trait scores for all 10,000 individuals from the GBFM assessment, and intra-individual pairwise differences were calculated for each individual (i.e., 10 differences per person). For example, if a person's Big Five IRT-based scores were 0, 0, 0, 1, and 1, that individual's pairwise differences were 0, 0, 0, 0, 1, 1, 1, 1, 1, and 1; and the individual's median absolute difference was 1. The median of the absolute values of pairwise differences was calculated for each individual. Across all 10,000 individuals, the 10^{th}, 30^{th}, 50^{th}, 70^{th}, and 90^{th} percentiles of these individual medians were also calculated. For the 10^{th} percentile, the median of medians was 0.48; for the 30^{th}, 0.75; the 50^{th}, 0.98; the 70^{th}, 1.27; and for the 90^{th}, 1.78. Then, at least half of the 10,000 individuals had median absolute differences of at least 0.98. The 10^{th}, 50^{th}, and 90^{th} percentiles were similar to the small, medium, and large magnitude conditions of the simulation study.

True Positive Rate

Overview. Table F1 in Appendix F displays η^2 effect sizes from ANOVA results for the true positive rate (TPR) while Table F4 provides sums of squares and degrees of freedom. Each η^2 value corresponds to the proportion of observed variance in TPR that was accounted for by a main effect or interaction. The largest source of variation was the magnitude of true differences, which accounted for between 39.0% and 47.4% of the observed variance, depending on the AMC hypothesis test. Precision (i.e., test length for

FL-CATs or standard error of measurement termination value for VL-CATs) accounted for 7.5% and 8.0% of observed TPR variance for the *Z*-test and score test, respectively, while accounting for 11.9% of the variance for the LRT. The number of scales accounted for 12.1%, 8.1%, and 6.9% of observed variance for the *Z*-test, LRT, and score test, respectively. Both discrimination and the number of differences accounted for a small amount of variance. Notably, the type of CAT (i.e., FL versus VL) and type of item bank (i.e., middle-peaked versus equally spaced) accounted for a negligible amount of observed variance. Scale location accounted for a small amount of variance (1.2%) for the LRT but a negligible amount for the remaining hypothesis tests. Collectively, the main effects accounted for between 67.5% and 76.6% of observed variance across hypothesis tests—driven by magnitude, precision, the number of scales, discrimination, and the number of differences. In addition, two-way interactions accounted for between 14.1% and 21.9% of variance across tests while three-way interactions accounted for between 5.90% and 6.90% of variance across tests. Note that, while it refers to locations, the ninth design factor is denoted as “Boundaries” in Table F1 and other ANOVA results in order to prevent confusion with the “Scale Location” design factor.

Table 5 provides TPR means and standard deviations for each hypothesis test collapsed across all conditions. For TPR, the *Z*-test provided the best performance with the largest mean (0.846) and smallest standard deviation (0.220) among hypothesis tests (note, as shown later, the *Z*-test also performed the worst in terms of the false positive rate). Means were smaller for the LRT and score test, and both tests had larger standard deviations than the *Z*-test. Compared to the *Z*-test, the LRT and score test performed more similarly to one another.

Table 5. Mean and Standard Deviation of True Positive Rate by AMC Hypothesis Test.

Statistic	LR	Z	Score
Mean	0.727	0.846	0.775
Standard Deviation	0.312	0.220	0.284

Effect of Magnitude. Across hypothesis tests, magnitude was the greatest source of variation for TPR. Table 6 displays TPR means conditional on magnitude across hypothesis tests. Across the three tests, means were low for small magnitudes (i.e., d = 0.5); but the differences in means between the *Z*-test and other tests were 0.221 (LRT) and 0.150 (score test), showcasing better performance for the *Z*-test when detecting small differences. This gap diminished for medium magnitudes (i.e., d = 1.0) where all three hypothesis tests had means between 0.8 and 0.9. For large magnitudes (i.e., d = 1.5), means were greater than 0.9 with the *Z*-test providing the best performance. Notably, across all hypothesis tests, moving from a magnitude of 0.5 to 1.0 provided a larger increase in TPR than moving from a magnitude of 1.0 to 1.5.

Table 6. Mean of True Positive Rate by AMC Hypothesis Test Conditional on Magnitude.

Magnitude	LR	Z	Score
0.5	0.436	0.657	0.507
1.0	0.804	0.899	0.847
1.5	0.942	0.981	0.969

As shown in Table F1, the interaction between magnitude and precision accounted for between 2.1% and 3.1% of observed variance, depending on the hypothesis test. Figure 5a displays how TPR means differed by magnitudes conditional on test length (i.e., for FL-CATs). Specifically, the gaps among test length means decreased as magnitude increased. In Figure 5b, this trend continued for magnitudes conditional on standard error of measurement termination value (i.e., for VL-CATs) with gaps again decreasing as magnitude increased. Notably, for FL-CATs, across magnitudes, performance differences were smaller between test lengths of 30 and 50, relative to differences between test lengths of 10 and 30. For VL-CATs, a similar difference was not present across magnitudes. For FL-CATs, across hypothesis tests, TPR means at or above 0.8 occurred for a magnitude of at least 1.0 and for a test length of at least 30; for VL-CATs, means at or above 0.8 also occurred for magnitudes of at least 1.0 and for standard error terminations of 0.3 or less.

Figure 5a

Mean of True Positive Rate for Magnitude by Test Length for Fixed-length Computerized Adaptive Tests

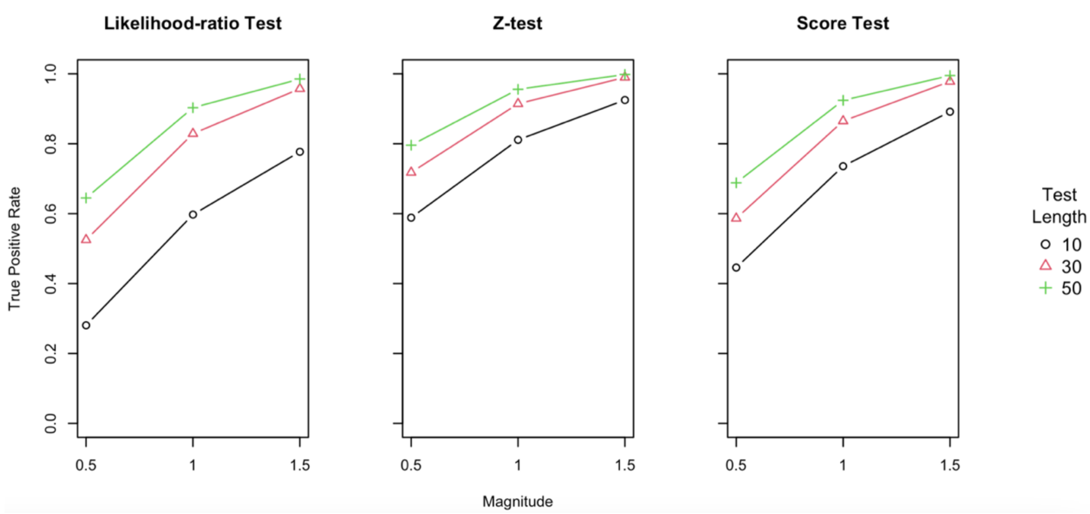
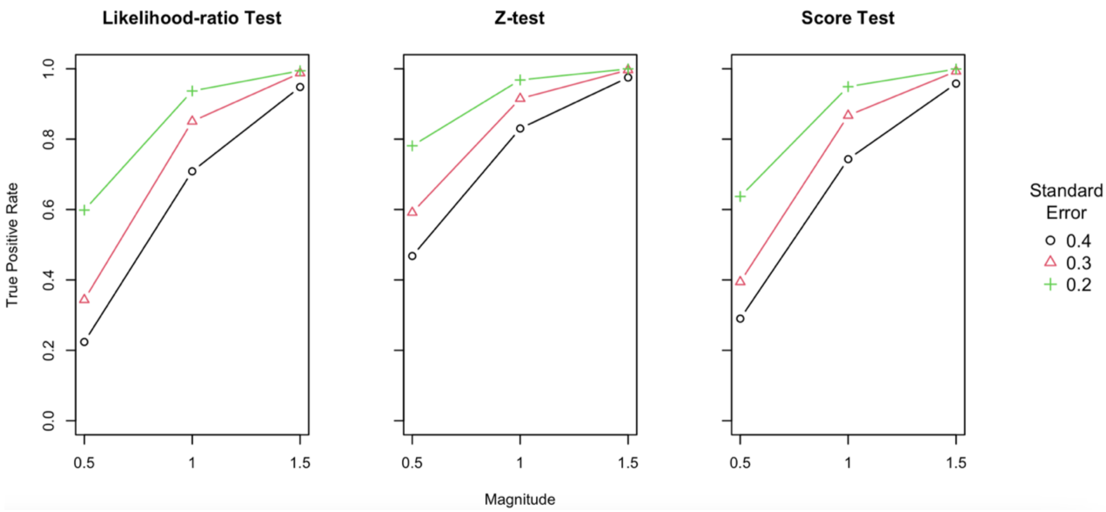


Figure 5b

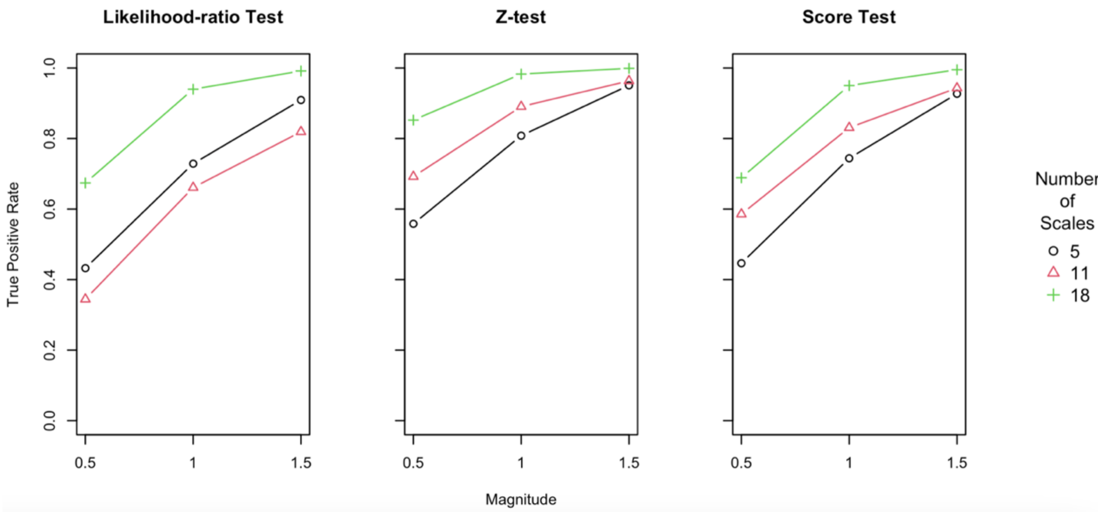
Mean of True Positive Rate for Magnitude by Standard Error Termination for Variable-length Computerized Adaptive Tests



As displayed in Figure 6, magnitude also interacted importantly with the number of scales when accounting for TPR variance, accounting for between 1.6% and 5.1% of observed variance (see Table F1). TPR performance increased and converged as magnitude increased across all numbers of scales and across all hypothesis tests. Across tests, a ceiling effect was nearly observed for medium and large magnitudes for 18 scales. Notably, for the LRT, performance was similar for 5 and 11 scales for small, medium, and large magnitudes. In contrast, for the *Z*-test, performance for 11 and 18 scales was somewhat similar—compared to 5 scales—but the two diverged at small magnitudes. Finally, performance for the score test was somewhat similar for 11 and 18 scales across all magnitudes, compared to 5 scales. For 11 scales or more, and for magnitudes of 1.0 or greater, TPR means occurred at 0.8 or more for the *Z*-test and score test.

Figure 6

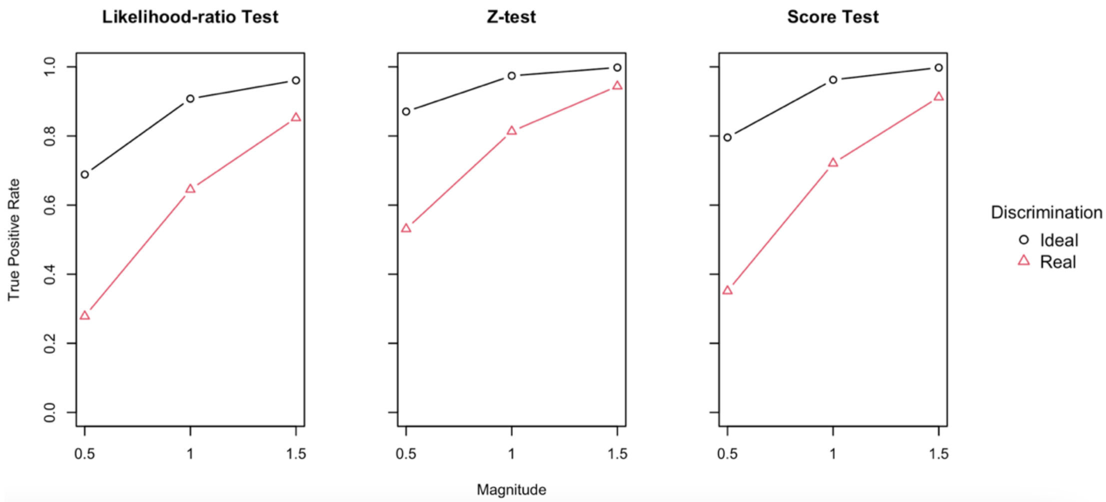
Mean of True Positive Rate for Magnitude by Number of Scales



In addition to precision and the number of scales, magnitude interacted with discrimination to account for between 1.4% and 2.9% of observed TPR variance (see Table F1). Figure 7 displays this interaction across hypothesis tests. While ideal discriminations improved performance across magnitude levels, the improvement was greatest for small magnitudes but reduced as magnitudes reached 1.5. Noticeably, for a small magnitude, a mean TPR of at least 0.8 was achieved for the *Z*-test for ideal discriminations with the score test following closely.

Figure 7

Mean of True Positive Rate for Magnitude by Discrimination



Effect of Precision. Across AMC hypothesis tests, precision accounted for between 8.0% and 11.9% of observed variance in TPR (see Table F1). For FL-CATs, Table 7a displays mean TPRs across hypothesis tests for each test length. For each hypothesis test, TPR increased as test length increased. At each level of test length, the *Z*-test provided the best performance, followed by the score test and then the LRT. Notably, a test length of 30 items achieved TPRs near or above 0.8. For VL-CATs, Table 7b displays mean TPRs across hypothesis tests for each standard error of measurement termination value. Similar to FL-CATs, VL-CAT performance improved with greater precision (i.e., for VL-CATs, as the standard error of measurement termination value decreased). Also similar to FL-CATs, the *Z*-test provided the best performance at each standard error level. Under the VL-CAT condition, TPR performance was more similar for the LRT and score test, compared to the FL-CAT condition. Across hypothesis tests, TPRs of at least 0.8

were achieved for standard error values of 0.2; but the *Z*-test also achieved a TPR of at least 0.8 at a standard error of 0.3.

Table 7a. Mean of True Positive Rate by AMC Hypothesis Test Conditional on Test Length for Fixed-length Computerized Adaptive Tests.

Test Length	LR	Z	Score
10	0.552	0.775	0.691
30	0.770	0.874	0.810
50	0.844	0.917	0.869

Table 7b. Mean of True Positive Rate by AMC Hypothesis Test Conditional on Standard Error for Variable-length Computerized Adaptive Tests.

Standard Error	LR	Z	Score
0.4	0.627	0.758	0.664
0.3	0.727	0.835	0.752
0.2	0.843	0.916	0.862

Effect of Number of Scales. Across hypothesis tests, the number of scales accounted for between 6.9% and 12.1% of observed variance in TPR (see Table F1). Table 8 provides TPR means for each level of the number of scales, across AMC hypothesis tests. Generally, for each hypothesis test, TPR improved as the number of scales increased. The one exception was for the LRT where performance for 11 scales was slightly worse than performance for 5 scales. Across hypothesis tests, and across levels of the number of scales, performance was best for the *Z*-test. Except for the condition with 11 scales, performance was similar between the LRT and the score test. For 11 scales, the *Z*-test achieved a mean TPR of at least 0.8; and for 18 scales, all three hypothesis tests achieved mean TPRs of at least 0.8.

Table 8. Mean of True Positive Rate by AMC Hypothesis Test Conditional on the Number of Scales.

Number of Scales	LR	Z	Score
5	0.669	0.750	0.678
11	0.660	0.851	0.786
18	0.853	0.937	0.859

Effect of Discrimination. Discrimination accounted for between 5.6% and 7.7% of observed TPR variance across hypothesis tests (see Table F1); and Table 9 provides mean TPRs conditional on discrimination conditions. For realistic item discriminations (i.e., derived from the informative real data analyses), hypothesis tests performed worse compared to the ideal item discriminations condition. Notably, under the realistic condition, the *Z*-test performed similarly to the LRT under the ideal condition. For both discrimination conditions, the LRT and score test achieved similar performance, relative to the *Z*-test, although the score test outperformed the LRT for both conditions. The *Z*-test provided the best performance overall. For the ideal condition, all three hypothesis tests achieved mean TPRs of at least 0.8.

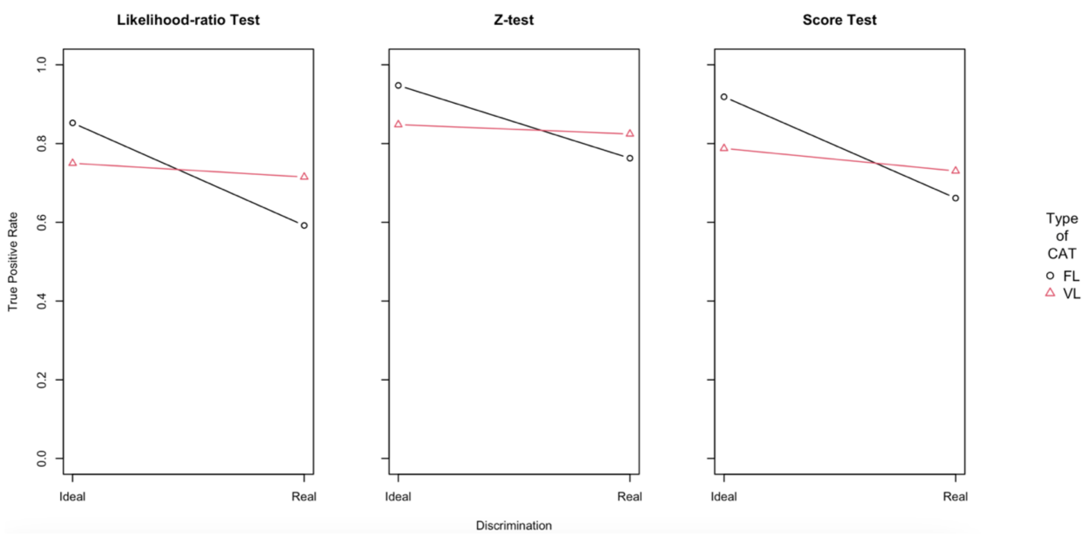
As shown in Table F1, the interaction between discrimination and type of CAT accounted for between 3.1% and 3.4% of observed variance across hypothesis tests. Figure 8 displays this interaction for each hypothesis test. For VL-CATs, mean TPRs improved slightly when moving from realistic to ideal discriminations for all three tests. However, for FL-CATs, the improvement was larger. Moreover, for FL-CATs, the improvement from realistic to ideal for the LRT was greater than the improvement for both the *Z*-test and the score test, relative to VL-CATs.

Table 9. Mean of True Positive Rate by AMC Hypothesis Test Conditional on Discrimination.

Discrimination	LR	Z	Score
Realistic	0.653	0.794	0.696
Ideal	0.801	0.898	0.853

Figure 8

Mean of True Positive Rate for Discrimination by Type of CAT



Effect of Number of Differences. The number of differences accounted for between 2.5% and 3.1% of observed TPR variance across hypothesis tests (see Table F1). Table 10 displays mean TPRs conditional on the number of differences (i.e., the “Less” condition corresponded to about 40% of possible differences differing while the “More” condition corresponded to about 60%). Across hypothesis tests, performance improved when more differences were present. The *Z*-test performed best for each condition, followed by the score test and the LRT. The *Z*-test and the score test achieved mean TPRs of at least 0.8 with more differences; but the LRT did not. Noticeably, the *Z*-test did achieve a mean TPR of at least 0.8 for less differences although, as shown later, this performance was likely an artifact tied to poor performance with the false positive rate.

Table 10. Mean of True Positive Rate by AMC Hypothesis Test Conditional on the Number of Differences.

Number of Differences	LR	Z	Score
Less	0.672	0.811	0.728
More	0.783	0.881	0.821

False Positive Rate

Overview. Table F2 in Appendix F displays ANOVA results for the false positive rate (FPR) across hypothesis tests while Table F5 provides sums of squares and degrees of freedom. The number of scales and scale location accounted for 31.3% and 12.2% of observed FPR variance for the *Z*-test, respectively, while they accounted for 26.5% and 13.5% of the variance for the score test and around 3% of variance for the LRT. Among the main effects, precision (i.e., test length for FL-CATs or standard error termination value for VL-CAT) accounted for the largest amount of variation for the LRT, describing 8.70% of variance. For the LRT, the type of CAT (i.e., FL-CAT versus VL-CAT) was the second largest source of FPR variation, describing 5.90% of observed variance.

Main effects and interactions accounted for FPR variance differently, depending on the hypothesis test. For the *Z*-test and score test, main effects accounted for 55.2% and 53.6% of observed variance, respectively; but for the LRT, main effects only accounted for 21.4% of observed variance. In addition, two-way interactions accounted for 31.7% and 34.1% of observed variance for the *Z*-test and score test, respectively. For the LRT, two-way interactions accounted for 8.4% of variance. In contrast, for the *Z*-test and score test, three-way interactions accounted for 9.2% and 8% of variance, respectively, but accounted for 30.9% of variance for the LRT.

Table 11 displays results for the FPR across the three hypothesis tests, collapsed across all simulation factors and conditions. The mean FPR for the LRT closely matched the nominal Type I error rate of 0.05, but both the *Z*-test and score test had mean FPRs much larger than the nominal Type I error rate. Moreover, the standard deviation was small for the LRT (0.011) but much larger for the *Z*-test and score test, suggesting mean FPRs varied widely across simulation conditions for these last two hypothesis tests. Among the three tests, the LRT performed the best while the *Z*-test performed the worst, which was the opposite outcome compared to TPR performance.

Table 11. Mean and Standard Deviation of False Positive Rate by AMC Hypothesis Test.

Statistic	LR	Z	Score
Mean	0.054	0.287	0.169
Standard Deviation	0.011	0.221	0.252

Effect of Number of Scales. For the *Z*-test and score test, the number of scales was the greatest source of variation for the FPR, accounting for 31.3% and 26.3% of observed variance, respectively. However, for the LRT, the number of scales accounted for 3.7% of observed variance (see Table F2). Table 12 displays FPR means conditional on the number of scales for the hypothesis tests. The LRT controlled the Type I error nearly at the nominal rate of 0.05, even as the number of scales increased. In contrast, mean FPRs increased from 5 to 18 scales for the *Z*-test and score test. For 11 scales, mean FPRs were higher than other conditions for both the *Z*-test and score test. For 5 or 18 scales, notably, the LRT and score test had similar performance with the score test having a more inflated Type I error rate. For all numbers of scales, the *Z*-test provided the worst performance.

Table 12. Mean of False Positive Rate by AMC Hypothesis Test Conditional on the Number of Scales.

Number of Scales	LR	Z	Score
5	0.051	0.133	0.068
11	0.057	0.447	0.356
18	0.054	0.281	0.084

Among the two-way interactions, the interaction between scale location and the number of scales accounted for 15.3% and 14.5% of FPR variance for the *Z*-test and score test, respectively (see Table F2). Figure 9 displays this interaction for all three hypothesis tests. For both the *Z*-test and score test, mean FPRs were higher for scale locations with low information, across numbers of scales. However, for both tests, mean FPRs were much higher for 11 scales at areas of low information. This outcome was also found in the interaction between discrimination and the number of scales—which accounted for 7.9% and 9.9% of FPR variance for the *Z*-test and score test, respectively. Figure 10 shows that ideal discriminations provided higher mean FPRs across numbers of scales; but for 11 scales, ideal discriminations had higher FPRs for the *Z*-test and score test.

Figure 9

Mean of False Positive Rate for the Number of Scales by Scale Location

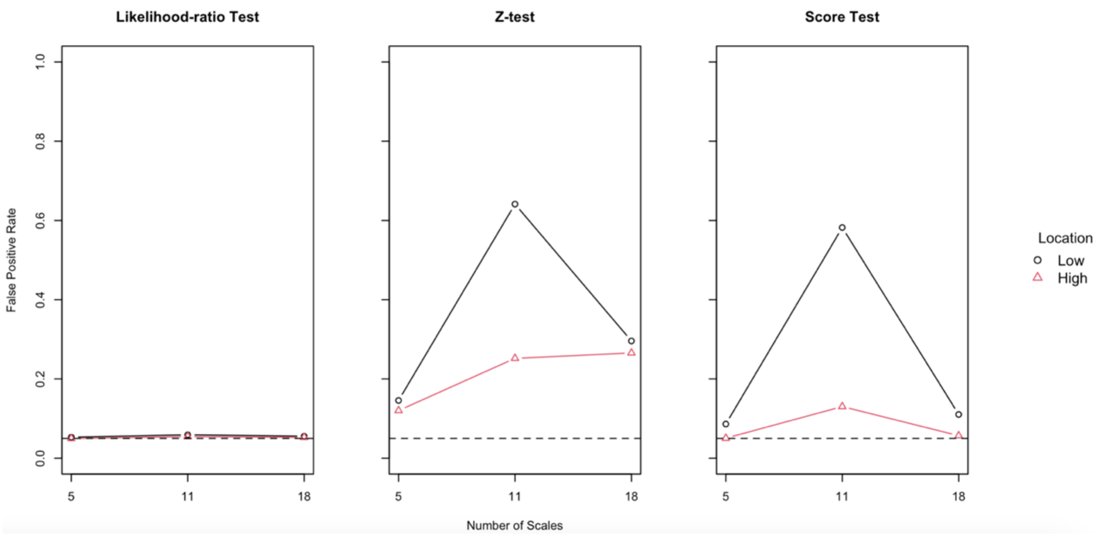
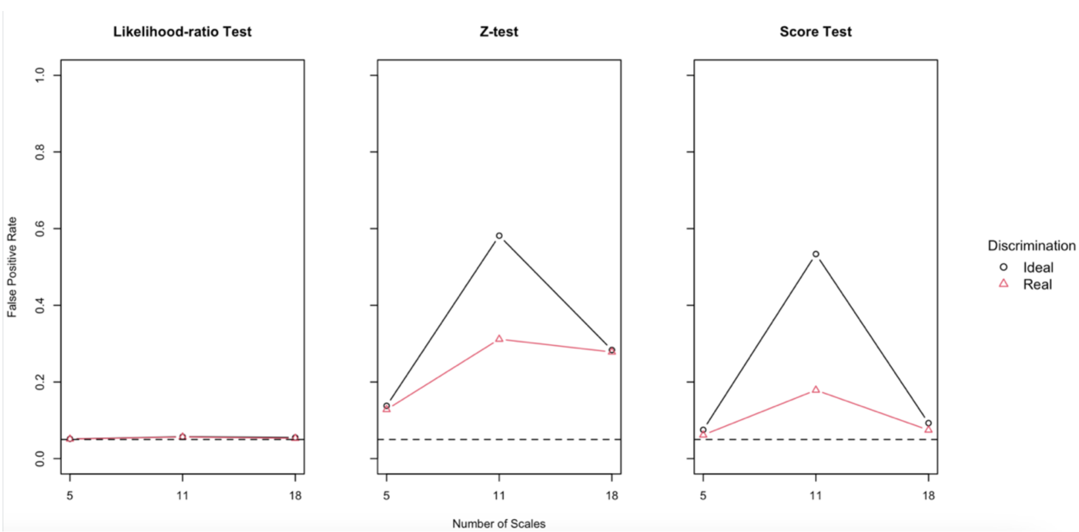


Figure 10

Mean of False Positive Rate for the Number of Scales by Discrimination



The three-way interaction among scale location, discrimination, and the number of scales accounted for 7.7% and 7.1% of the variance for the *Z*-test and score test, respectively. Figure 11a and Figure 11b show this interaction graphically in two parts. Figure 11a shows mean FPRs for the three hypothesis tests conditional on the number of scales and scale location for ideal item discriminations; and Figure 11b displays the same effects for realistic item discriminations. For scale locations with low information, for 11 scales, and for ideal discriminations, mean FPRs were highly elevated for the *Z*-test and score test, which likely contributed to the presence of important two-way interactions among the same effects. This set of conditions also likely reduced the performance of the *Z*-test and score test conditional on the number of scales. Table 13 displays mean FPRs across hypothesis tests conditional on the number of scales for (a) scale locations with high information and (b) realistic item discriminations. While mean FPRs remained inflated for the *Z*-test and score test for 11 scales, they were reduced greatly compared to the corresponding FPRs in Table 12. In addition, for 5 and 18 scales, Table 13 shows that the LRT and score test performed similarly. Mean FPRs for the *Z*-test remained inflated for all numbers of scales.

Figure 11a

Mean of False Positive Rate for the Number of Scales by Location for Ideal Discriminations

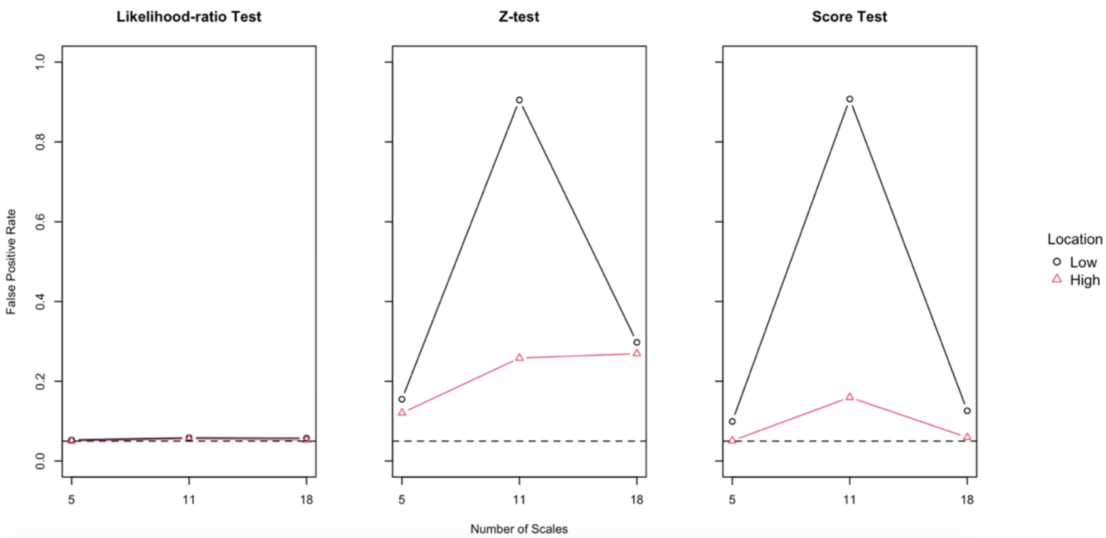


Figure 11b

Mean of False Positive Rate for the Number of Scales by Location for Realistic Discriminations

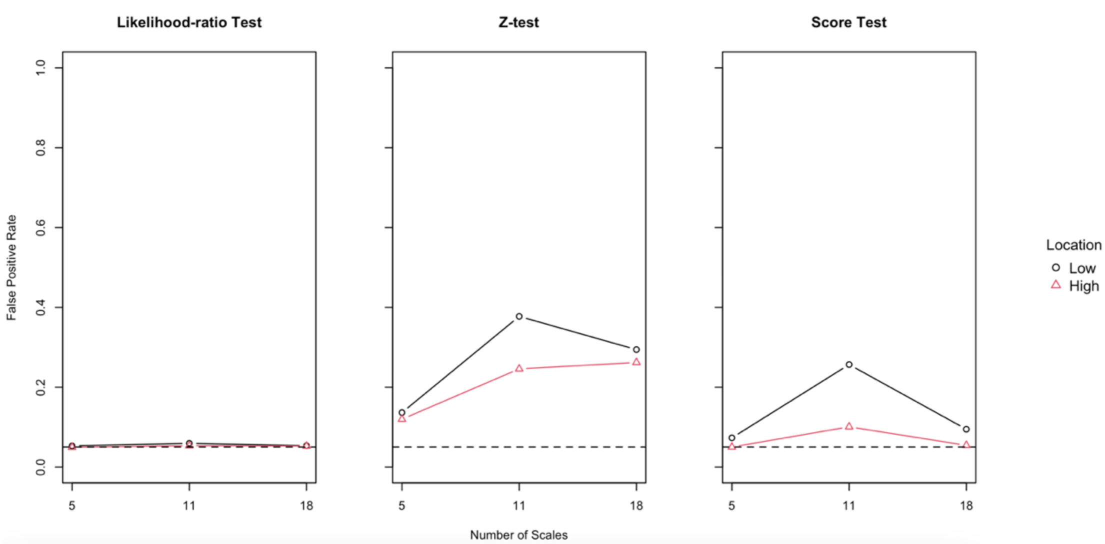


Table 13. Mean of False Positive Rate by AMC Hypothesis Test Conditional on the Number of Scales for Scale Locations with High Information and for Realistic Discriminations.

Number of Scales	LR	Z	Score
5	0.049	0.120	0.050
11	0.053	0.246	0.101
18	0.052	0.262	0.054

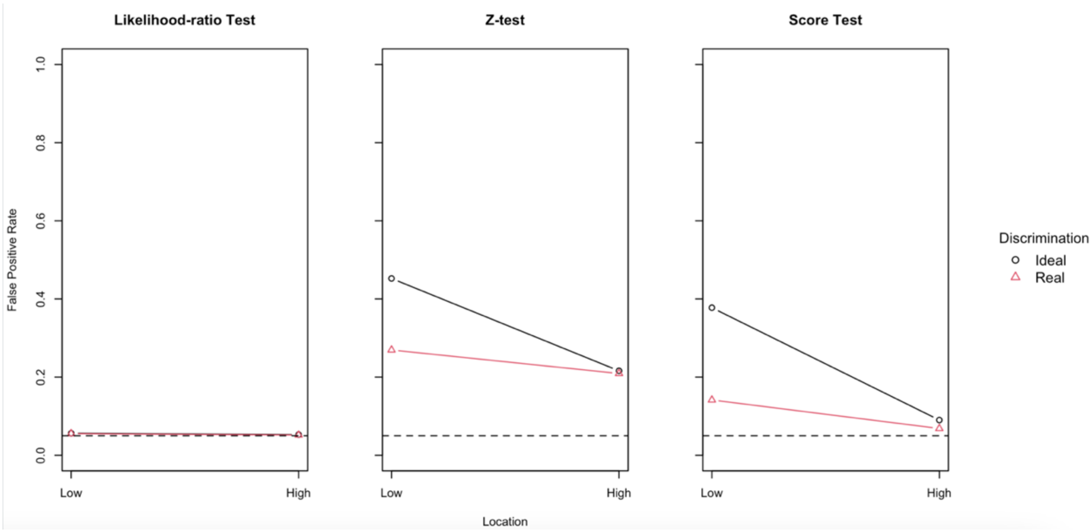
Effect of Scale Location. Scale location accounted for 12.2% and 13.5% of observed FPR variance for the *Z*-test and score test, respectively, and for 2.7% of variance for the LRT (see Table F2). As displayed in Table 14, all three hypothesis tests performed better in scale locations with high information although the LRT performed the best. Moreover, for high information, performance for the score test approached performance for the LRT. In addition to the main effect, the two-way interaction between scale location and discrimination accounted for 4.5% and 5.0% of observed FPR variance for the *Z*-test and score test, respectively. Figure 12 displays this interaction. While moving from scale locations with low information to locations with high information reduced mean FPRs for realistic discriminations, the improvement was greater for ideal discriminations. As previously discussed, the three-way interaction among scale location, discrimination, and the number of scales likely contributed to the importance of this two-way interaction.

Table 14. Mean of False Positive Rate by AMC Hypothesis Test Conditional on Scale Location.

Scale Location	LR	Z	Score
Low	0.056	0.361	0.260
High	0.052	0.213	0.079

Figure 12

Mean of False Positive Rate for Scale Location by Discrimination



Effect of Precision. Precision accounted for 8.7%, 2.7%, and 3.4% of observed FPR variance for the LRT, *Z*-test, and score test, respectively (see Table F2). Table 15a shows mean FPRs for test length (for FL-CATs) while Table 15b displays FPRs for standard error of measurement termination value (for VL-CATs), collapsed across all other simulation factors. As displayed in Table 15a, mean FPRs improved as test length increased for all three hypothesis tests. Table 15b displays results for VL-CATs: mean FPRs remained stable for the LRT and score test and increased slightly for the *Z*-test as standard error decreased (i.e., as precision increased), but then decreased for the smallest standard error.

Table 15a. Mean of False Positive Rate by AMC Hypothesis Test Conditional on Test Length for Fixed-length Computerized Adaptive Tests.

Test Length	LR	Z	Score
10	0.063	0.393	0.294
30	0.054	0.295	0.166
50	0.054	0.285	0.151

Table 15b. Mean of False Positive Rate by AMC Hypothesis Test Conditional on the Standard Error for Variable-length Computerized Adaptive Tests.

Standard Error	LR	Z	Score
0.4	0.051	0.242	0.137
0.3	0.053	0.247	0.135
0.2	0.050	0.258	0.134

Effect of Discrimination. The main effect of discrimination accounted for 0.2%, 5.3%, and 7.3% of observed FPR variance for the LRT, *Z*-test, and score test, respectively (see Table F2), indicating that it was important for the *Z*-test and score test. As displayed in Table 16, FPRs remained stable across discrimination conditions for the LRT but were worse for ideal discriminations for both the *Z*-test and score test. The three-way interaction likely explained some of the observed FPR inflation for the *Z*-test and score test—especially given the condition with 11 scales, with scale locations that have low information, and with ideal discriminations. Across discrimination conditions, the LRT remained near the nominal Type I error rate of 0.05.

Table 16. Mean of False Positive Rate by AMC Hypothesis Test Conditional on Discrimination.

Discrimination	LR	Z	Score
Realistic	0.054	0.239	0.105
Ideal	0.055	0.334	0.234

Effect of Type of CAT. The type of CAT accounted for 5.9%, 3.0%, and 1.9% of observed FPR variance for the LRT, *Z*-test, and score test, respectively (see Table F2). Table 17 displays mean FPRs conditional on the type of CAT, collapsed across other simulation factors. Performance was better for VL-CATs across all hypothesis tests. However, the improvement from FL-CATs to VL-CATs was greater for the *Z*-test and score test, compared to the LRT, which likely represented a floor effect for the LRT (i.e., it already controlled the Type I error rate well).

Table 17. Mean of False Positive Rate by AMC Hypothesis Test Conditional on Type of Computerized Adaptive Test.

Type of CAT	LR	Z	Score
Fixed	0.057	0.325	0.203
Variable	0.051	0.249	0.135

Mean Number of Items

Overview. Table F3 displays ANOVA results for the mean number of items needed to terminate variable-length CATs (VL-CATs) while Table F6 provides sums of squares and degrees of freedom. Discrimination, precision, and the number of scales together accounted for 84.8% of the observed variance—accounting for 35.4%, 25.7%, and 23.7% of variance, respectively. Across all simulation conditions for VL-CATs, the mean among all mean number of items was 39.1. The number of items differed greatly by simulation condition as the standard deviation among means was 29.6.

Effect of Discrimination. Across simulation factors, discrimination was the most important source of variation for the mean number of items. Collapsed across other simulation factors, the mean among the mean number of items for realistic discriminations was 56.7; for ideal discriminations, 21.6. The two-way interaction between discrimination and precision (i.e., the standard error termination value) accounted for 6.5% of variance in the mean number of items. Table 18 displays the means for the mean number of items for discrimination conditions across standard error values. The number of items required to terminate VL-CATs was lower for ideal discriminations, across all levels of standard error. In fact, the number of items needed at a standard error of 0.2 for ideal discriminations was close to the value needed at a standard error of 0.4 for realistic discriminations. Moving from higher standard errors to lower standard errors, the change was larger for realistic discriminations, compared to the change for ideal discriminations.

Table 18. Mean Number of Items by Standard Error Termination Value Conditional on Discrimination for Variable-length Computerized Adaptive Tests.

Discrimination	Standard Error		
	SE = 0.4	SE = 0.3	SE = 0.2
Realistic	31.371	52.599	86.066
Ideal	13.941	18.854	31.729

Other Effects. After discrimination, the effects of precision, the number of scales, and scale location collectively accounted for the majority of the remaining observed variance in the mean number of items. For precision, the means for the mean number of items for each standard error termination value included: a mean of 22.7 for a standard error of 0.4; a mean of 35.7 for 0.3; and a mean of 58.9 for 0.2. As the standard error decreased (i.e., as precision increased), the number of items required to terminate the VL-CAT increased. In addition, the means for the number of scales included: a mean of 29.2 for 5 scales; 59.4 for 11 scales; and 28.7 for 18 scales. The mean number of items was similar between 5 and 18 scales but was much higher for 11 scales, likely because the condition with 11 scales used the 2PL model while the conditions with 5 or with 18 scales used the GRM. Finally, the means for scale location included: a mean of 45.2 for scale locations with low information; 37.9 for mixed information; and 34.1 for high information. When scores came from locations with more information, the number of items needed to terminate a VL-CAT decreased.

Real Data Analysis

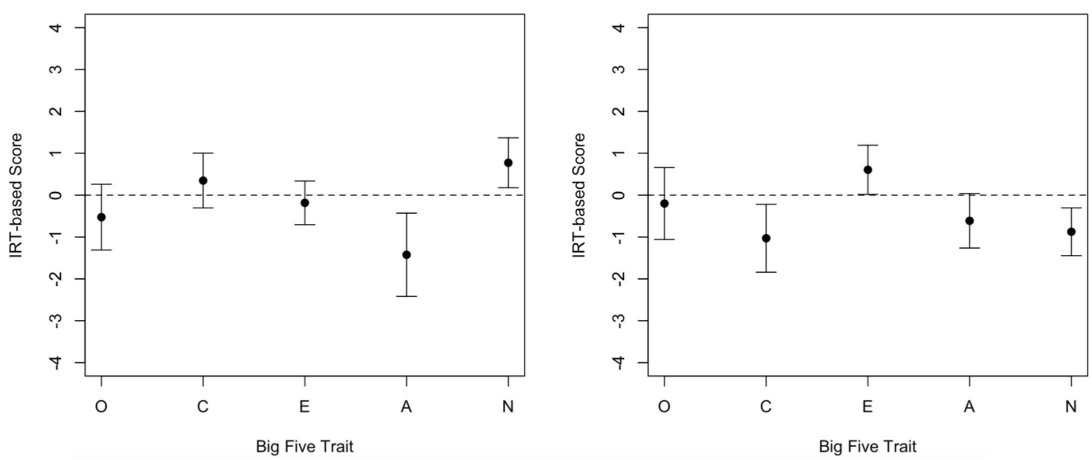
Table 19 displays results from the post-simulation real data analysis. Of 10,000 participants in a random holdout sample who completed Goldberg's Big Five Measure, only 4,096 participants met criteria for inclusion in the comparison because their scores were similar to the trait scores used in the simulation in terms of magnitude. Not enough participants were available to disaggregate by scale location (i.e., areas of low, mixed, or high information). Sample sizes used to calculate each real detection rate are included in the first column on the left of Table 19. For the LRT, the real detection rate was close to the corresponding FPR for the LRT in the simulation. For a magnitude of 0.5, the real detection rate for the LRT was nearly identical to the TPR observed in the simulation. At magnitudes of 1.0 and 1.5, the real detection rates for the LRT were lower than the corresponding TPRs. In contrast, across magnitude levels, real detection rates using the *Z*-test were consistently lower than the FPR or TPR for each corresponding simulation condition. Moreover, compared to the LRT, the *Z*-test had a higher real detection rate across all magnitude levels. Among individuals in the holdout sample, two individuals with significant LRTs were randomly selected; and their IRT-based score profiles are displayed in Figure 13. For Figure 13, IRT-based confidence bands (which were not used in detecting significance of differences) are within 95% confidence; and letters on the horizontal axis are associated with the same traits as those in Figure 2.

Table 19. Detection Rate for Two AMC Hypothesis Tests for Real Data for Goldberg's Big Five Measure Compared to Relevant False Positive or True Positive Rate for Simulation Conditions.

Sample Size	Magnitude	LR		Z	
		Real	Simulated	Real	Simulated
684	0	0.088	0.063	0.127	0.194
1,221	0.5	0.111	0.119	0.161	0.257
1,498	1.0	0.253	0.353	0.339	0.533
693	1.5	0.407	0.646	0.556	0.807

Figure 13

IRT-based Score Profiles with Significant LRTs from a Holdout Sample for Goldberg's Big Five Measure for Two Random Respondents



Chapter 4. Discussion and Conclusions

Major Findings

True Positive Rate. The ability to detect intra-individual differences successfully using adaptive profile difference analysis (APDA) depended largely on the magnitude of intra-individual differences among an individual's trait scores. Specifically, while APDA provided an adequate detection rate with the *Z*-test for magnitudes of 0.5 (but also had the worst FPR among the hypothesis tests), APDA performed better across hypothesis tests with magnitudes of intra-individual differences of at least 1.0. For at least one multiscale assessment, the real data revealed that such difference magnitudes are common: half of all individuals in the calibration sample from Goldberg's Big Five Measure had intra-individual, median absolute differences of 0.98 or larger.

In addition to magnitude, simulation results indicated that item discriminations were important for understanding TPRs. In particular, APDA performed better when scales were developed using IRT (i.e., had ideal discriminations). Even at small magnitudes of 0.5, APDA hypothesis tests performed adequately for ideal discriminations, achieving average TPRs near 0.65 for the likelihood-ratio test (LRT) and near 0.80 for the *Z*-test and score test. This result is reasonable: as item discriminations increased, information (i.e., measurement precision) increased, improving the ability to detect differences. For the *Z*-test and score test, higher information allowed for a more accurate evaluation of test information at the pooled MLE, increasing the value of the *Z*-test statistic or score statistic when intra-individual differences existed. For the likelihood-ratio test (LRT), higher discriminations improved the estimation of likelihoods.

Together with magnitude and discrimination, precision (i.e., test length for fixed-length CATs and standard error of measurement termination value for variable-length CATs) accounted for a large amount of TPR variance. As test length increased, hypothesis tests improved in their detection capabilities because the presence of more items improved observed test information. Specifically, observed test information is the sum of individual item information functions with

$$I(\hat{\theta}_j) = \sum_{v=1}^{V} I_v(\hat{\theta}_j), \tag{20}$$

for $v = 1, \ldots, V$ items. With more items, test information increases in magnitude, improving accuracy when evaluating test information at the pooled MLE for the *Z*-test or score test. For the LRT, multiplying more items together when constructing likelihoods improved the estimation of those likelihoods, assuming local independence among items. Moreover, for VL-CATs, as standard error decreased, information increased because the two are inversely related, as seen in Equation 13. The result was a higher TPR. Across various conditions, results suggest that the *Z*-test provided the best TPRs at smaller test lengths and at higher standard errors; but it also had the highest FPR, making its use less desirable.

False Positive Rate. The number of scales and scale location largely accounted for the observed variance in the false positive rate (FPR) for the *Z*-test and score test; however, these effects did not account for variance in LRT performance. The importance of these effects was partially explained by the set of conditions with 11 scales, low information, and ideal discriminations; but the condition with 11 scales alone also inflated FPRs greatly for both the *Z*-test and score test. One possible explanation lies in Figures E2 and E3 in Appendix E, which show the relative bank information functions

(BIFs) for the condition with 11 scales, realistic discriminations, and realistic (i.e., middle-peaked) location parameters. Like the real test information functions (TIFs) that they emulate, the BIFs were highly peaked, providing little to no information at the extreme ends of latent traits for some scales. However, in the simulation, trait scores used for scale locations with low information were sometimes extreme, as shown in the second row of Tables A1, A2, or A3 in Appendix A. The set of scores in this condition would assume that an examinee had a trait level of 2 for all 11 scales. Then, one possibility is that estimating these trait levels using MLE was inaccurate due to the lack of information in the extremes for simulated BIFs based on real, middle-peaked TIFs—producing unstable scores in the simulation and, as a result, artificially inflating hypothesis test statistics due to the presence of extreme estimated scores. Even producing extreme estimates for 2 out of 11 scales could cause an omnibus hypothesis test to detect a difference when none exists.

Examining the set of conditions with 11 scales, scale locations with low information, and ideal discriminations, each simulee had 11 MLE trait scores. When looking at a random subset of 250 simulees' scores from the VL-CAT condition, where each simulee had 11 actual trait scores of 2, the average largest, intra-individual MLE trait scores were 4.72 and 5.80, which were much larger than the actual scores of 2. Examining the same set of conditions but for realistic discriminations, the corresponding average largest, intra-individual MLE scores were 2.45 and 2.83. Moreover, examining the same conditions but for 18 scales with realistic discriminations, the average largest MLE scores were 2.26 and 2.39. These results suggest that the simulation produced unstable MLE scores for the condition with 11 scales because insufficient information existed at the extremes for both the real and simulated scales. Then, FPR performance was made worse by the presence of scale locations with low information and by extreme trait score patterns.

In contrast to the *Z*-test and score test, the LRT performed well in terms of FPR for the condition with 11 scales and across all simulation conditions, controlling the FPR near its Type I nominal value of 0.05. While a test length of 10 items inflated the FPR for the LRT slightly for FL-CATs (see Table 15a), the LRT maintained mean FPRs near the nominal value of 0.05 for 30 items and above. Moreover, while the LRT controlled the FPR well for FL-CATs overall, the LRT performed slightly better for VL-CATs. While FPR variance for the *Z*-test and score test could be largely accounted for by the simulation design, FPR variance for the LRT could not, likely because the LRT had a consistently low FPR with little variance across simulation conditions. Comparatively, after removing the condition with 11 scales, the score test performed as well as the LRT. This result likely stems from the mathematical similarity between the LRT and score test: both tests make use of the likelihood of the null hypothesis when calculating their respective statistics. In contrast, the *Z*-test does not use the likelihood function when calculating its statistic.

Mean Number of Items. Simulation results reveal that discrimination, standard error, and the number of scales largely accounted for observed variance in the mean number of items needed to terminate VL-CATs. With ideal discriminations, the number of items needed was lower, compared to realistic discriminations. Across levels of standard error, ideal discriminations offered reductions between about 55% and 65% in the number of items needed per scale, compared to conditions with realistic discriminations; across many scales in a multiscale assessment, this amount of item savings could be impactful. Developing scales using IRT from the beginning would be important for achieving this benefit. Moreover, for the number of scales factor, the 5-scale and 18-scale conditions differed from the 11-scales condition: the 11-scales condition required about twice as many items per scale as the other conditions. This result is expected because the 11-scales condition used binary items, which had less information than polytomous items from the 5-scales or 18-scales conditions. More items are needed to achieve the same level of precision (i.e., standard error termination value) for scales with binary items, relative to scales with polytomous items. Finally, for the standard error of measurement, the number of items needed for termination increased as the standard error decreased across all other simulation conditions. As a simulee's estimated trait level became more precise (i.e., lower standard errors), more items were needed to achieve a given level of precision.

Other Effects

The type of location parameters (i.e., flat versus peaked BIFs) did not have a major impact on FPR or TPR. Figure E7 in Appendix E provides an example of simulated BIFs (in purple) using flatter locations for the GBFM assessment, compared to real TIFs (in red). While simulated BIFs were flatter for this condition—relative to realistic locations for the same set of scales (see Figure E1)—they were not flat at the extremes of the latent trait continuum for each scale. Rather, information decreased rapidly near the extremes. Thus, including this factor might have added redundancy to the simulation design as simulated flat BIFs were not as flat as needed to observe variation based on this factor (e.g., such as when examining scale location).

While IRT item locations had little impact on results, the number of differences did have a small impact on TPRs: as the actual number of differences increased in the simulation, TPR performance improved for all three hypothesis tests. In the informative real analysis for the GBFM, the median of intra-individual median absolute differences was 0.98, suggesting that most individuals had five or six differences close to 1.0 or above (i.e., medium difference magnitudes or above). Given how common differences were for this analysis, it could be plausible to expect a larger (close to 60%) number of true differences in real data for some multiscale assessments. However, before performing APDA, an analysis might be conducted to determine the typical size of intra-individual differences for a given multiscale assessment in a real setting.

Comparison to Previous AMC Research

Similar to previous AMC studies, magnitude was a major factor when accounting for observed variance in TPR (e.g., see Tai et al., 2023, p. 45). In addition, as in previous AMC studies, small magnitudes of 0.5 produced small, possibly inadequate TPRs that might be unacceptable for some real-world applications. In contrast, TPRs improved greatly for magnitudes of 1.0 or above across a variety of simulation factors (see, for comparison, Lee, 2015, p. 44–49). Conditional on discrimination, difference magnitudes of 0.5 were difficult to detect for realistic discrimination conditions at scale locations with low information, which matched previous research suggesting higher item discriminations provide large TPR improvements for low magnitudes (e.g., see Phadke, 2017, p. 79).

Unlike past research, both the *Z*-test and score test performed much worse than the LRT in terms of FPR. For example, under various simulation conditions, Phadke (2017) typically found FPRs at 0.1 or less for all three AMC hypothesis tests. As explained under the section "Major Findings", in the current study, the inflated FPRs for the *Z*-test and the score test were largely specific to the set of conditions that included 11 scales with middle-peaked information functions, scale locations with low information, and ideal discriminations that further peaked information functions, drawing information away from extreme ends of latent trait continua. This set of conditions showcases how both the *Z*-test and score test can be sensitive to the presence of extreme scores and low information.

However, unlike Phadke (2017) who used three scores (across occasions), the number of scales (across traits) in this simulation study was larger. In fact, 11 scores or 18 scores was a larger number of scores than the number used in previous AMC research studies. For example, Weiss & von Minden (2011) examined change scores over five different occasions. Moreover, in the context of multivariate change, Wang, Weiss, & Suen (2021) investigated AMC hypothesis tests in the context of two or four scores per individual. Wang, Weiss, & Suen (2021) discovered that increasing the number of scores improved TPRs across multivariate versions of the three AMC hypothesis tests; but increasing also inflated FPRs for the multivariate *Z*-test (see Wang, Weiss, & Suen, 2021, p. 8). The current study showed similar FPR inflation for the univariate *Z*-test even after accounting for the condition with 11 scales (see Table 13). Like Wang, Weiss, and & Suen (2021), the current study found that both the LRT and score test controlled FPRs as the number of scales increased. However, three considerations limit this comparison: in the current study (1) hypothesis tests were univariate, (2) the numbers of scales were larger, and (3) different traits at a single occasion were examined.

Recommendations for Assessment Design

The findings of the current study suggest certain recommendations for designing multiscale assessments with APDA. First, item discrimination impacts TPRs when using APDA; and better TPRs can be achieved by developing scales with higher (i.e., ideal) item discriminations. If developing new scales with new items, a test developer might pilot more items than necessary for each scale on a pilot sample, calibrate banks using IRT models, and identify items with larger item discriminations for inclusion in test banks—such as discriminations between about 1.5 and 2 or above (on the $D = 1.7$ scale), depending on the IRT model.

Second, because magnitude is important for TPRs, a test developer should consider the expected, intra-individual difference magnitudes for a specific multiscale assessment application. For an educational assessment, for instance, future work might be done to determine how much IRT-based scores tend to differ intra-individually on reading, science, and mathematics scales in a comprehensive state educational assessment. If the differences are expected to be lower (e.g., closer to 0.5), the Z-test would offer better TPRs although large FPRs would be expected. However, for three scales, FPRs might not be as elevated. If difference magnitudes are more typically medium and close to 1.0 (e.g., as seen in the GBFM assessment), the LRT would offer the best balance between FPRs and TPRs.

Third, BIFs or TIFs for multiscale assessments should be examined to determine if enough information is available at scale locations where many examinees are expected. Not all trait distributions will be normally distributed, so the population distribution of trait scores should be considered to determine if enough measurement precision exists at locations of expected high score frequency on that distribution. The 11-scales condition in the simulation study offers a cautionary example where lacking information produces inflated FPRs for the *Z*-test and score test. Nevertheless, even in the presence of this condition, the LRT remains a useful option, providing the best balance between the FPR and the TPR.

Fourth, a test length of 30 for FL-CATs and a standard error termination value of 0.3 for VL-CATs provide adequate TPRs across various conditions. In addition, using a *Z*-test, a test length of 10 items could be sufficient for some FL-CAT designs although FPRs might be elevated in some applications if using the *Z*-test. If not employing CAT, results from the post-simulation real analysis (see Table 19) suggest that 10 items per scale might not be enough to achieve adequate TPRs: even for the *Z*-test for large magnitudes, the mean TPR was still below 0.6. This result, however, comes from the GBFM, which is based on classical test theory techniques. An IRT-based, fixed-length multiscale assessment could have a set of 10 items with higher discriminations and might, therefore, be able to achieve high TPRs with 10 items. TIFs for assessment scales could be evaluated to determine how much precision is available. However, if optimal items cannot be acquired for an assessment, or if IRT models have been fit on scales developed using classical test theory, the simulation results suggest that the LRT is a good alternative.

Limitations

The current study based realistic conditions (i.e., realistic item parameters) on a particular set of multiscale assessments, which limits the generalizability of results. First, assessments were developed using classical test theory where real TIFs tended to be middle-peaked. As such, the performance for realistic conditions is likely worse than the performance would have been if IRT-based multiscale assessments had been used to inform the simulation. While the ideal discriminations condition mitigated this issue somewhat, more work could be done to investigate how informing such a simulation with real IRT-based assessments impacts both the TPR and the FPR. Second, only personality multiscale assessments were used to inform simulated banks. Future work might inform simulation designs using real multiscale assessments from educational or clinical domains, for example—which could yield different BIF shapes or typical, intra-individual magnitudes.

In addition to informative real analyses, the item location factor was limited. While the flat condition for item locations did provide flatter BIFs in the middle of latent trait continua, such functions were not as flat at the extremes, which likely reduced potential variance for this design factor. For example, if extremes had more information under the condition with flat information functions, FPRs might have improved for the condition with 11 scales for the *Z*-test and score test; but this possibility remains speculative and would need to be tested.

Finally, more types of trait score patterns should be examined to further validate APDA. The trait scores in Table 1 and Appendix A correspond well to their respective simulation factors; nevertheless, if trait score patterns were drawn randomly from a distribution specific to each set of factors, results might be further generalizable. Moreover, such random draws might better protect against extreme results, such as poor FPRs for the *Z*-test or score test in the condition with 11 scales and with extreme trait score patterns (i.e., scale locations with low information). Still, this process might not eliminate such a problem entirely.

Future Directions

TPR performance was not adequate for small magnitudes of 0.5 in the simulation study, particularly when using realistic item discriminations based on scales constructed using classical test theory. One option would be to improve the items themselves using IRT-based scales with better item discriminations; this option should improve the TPRs. Another possibility could be to develop a novel AMC omnibus hypothesis test to better detect smaller differences. One such novel test is offered in Appendix G: a new Bayesian AMC omnibus hypothesis index. This approach penalizes the LRT statistic using prior distributions, observed TIFs, and the number of scales. By incorporating prior knowledge, it might be possible to detect small differences more easily. While not strictly a hypothesis test, this approach could be used as an index to detect psychometrically significant change under certain assumptions. This new index was applied in the context of the exact same simulation study as the other three tests, and results can be viewed in Appendix G. Findings indicate that this Bayesian approach might be useful for detecting small differences, but more research would be needed to validate it further.

In addition, while this research study examined omnibus hypothesis tests, future work should explore the utility of post hoc tests in the context of APDA. After an omnibus hypothesis test suggests that at least one psychometrically significant difference exists, post hoc tests could be used to determine the specific pairs of traits that differ significantly. Wang and Weiss (2018, p. 227) define post hoc tests in the context of the AMC framework. For a single individual, an omnibus hypothesis test's significance does not need to be corrected for FPR inflation because only a single test is conducted. However, when conducting many post hoc comparisons within a single individual, corrections might be necessary to control FPR inflation. Future work should examine if such post hoc comparisons result in FPR inflation for each of the primary AMC hypothesis tests—including the LRT, the *Z*-test, and the score test—in an APDA context.

In addition, more research could be done to explore applications of AMC hypothesis tests in the context of multiscale assessments without CAT. IRT-based assessments would likely have reduced detection rates, especially if a multiscale assessment is based on scales constructed using classical test theory (e.g., based on maximizing coefficient alpha). For example, a study might be conducted to evaluate the detection rates of the LRT using IRT-based scores for an educational assessment like the Minnesota Comprehensive Assessment with its three knowledge scales or for a clinical assessment like the MMPI-3 with its Restructured Clinical Scales. Applications to different conceptual domains could reveal new use cases or problems with APDA.

Finally, research could be conducted to determine if IRT-based score profiles—derived from CATs or from fixed-form IRT-based tests—could be used to search for subpopulations within a larger population. For example, the right panel of Figure 13 shows the IRT-based score profile for one random respondent from the GBFM holdout sample. This individual's Extraversion score (using the confidence interval overlap method) is psychometrically different than this individual's Conscientiousness score or Neuroticism score, but the remaining eight comparisons do not show significant differences. This information might be used to define a particular type of person with a certain set of tendencies. However, an enormous amount of variation would exist not only in the possible differences among the ten comparisons but also among the locations of the scores themselves. Then, one place to begin might be to examine whether scores are psychometrically different from the mean point of zero (e.g., the dashed line in Figure 13): they could be psychometrically below, above, or at zero, which would yield $3^5 = 243$ possible outcomes in the GBFM example. For this assessment or others, a practitioner with substantive knowledge would need to determine if such an approach is useful, depending on theory.

Conclusions

The APDA procedure works well under certain conditions. In particular, when magnitudes of 1.0 or more are common in real multiscale assessment data, APDA can detect differences well when they exist. At least in one multiscale assessment (i.e., GBFM), a magnitude of about 1.0 or above is a common, intra-individual difference magnitude, suggesting APDA could be used for such an assessment. Moreover, among the AMC hypothesis tests used, the LRT appears to offer the best balance of FPRs and TPRs across a wide range of realistic simulation conditions. In fact, in terms of FPR, the LRT is surprisingly robust to situations where extreme scores are present and where little information is available. If the cost of inflated FPRs is low, however, using the *Z*-test could also be beneficial for some real-world applications, given that the *Z*-test demonstrates high TPRs.

For AMC research more broadly, this study is the first to examine AMC omnibus hypothesis tests with 11 or 18 scores and shows that the LRT, in particular, provides a good balance of FPRs and TPRs when many scores are present. Moreover, this study is the first to explore comparisons among traits using IRT-based score profiles that incorporate individualized measurement error. Based on the study's findings, future IRT-based multiscale assessments could be used with AMC omnibus hypothesis tests—or with the APDA procedure using CATs—to enhance the objective interpretation of individual score profiles.

References

Anastasi, A., & Urbina, S. (1997). Technical and methodological principles. *Psychological Testing* (7th ed.). Prentice Hall/Pearson Education.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. https://doi.org/10.1007/BF02293814

Angoff, W. H. (1971). *The College Board admissions testing program: A technical report on research and development activities relating to the scholastic aptitude test and achievement tests*. Educational Testing Service.

Ben-Porath, Y. S. (2012). Interpreting the MMPI-2-RF substantive scales. *Interpreting the MMPI-2-RF*. University of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2020). *MMPI-3: Minnesota multiphasic personality inventory-3 manual for administration, scoring, and interpretation*. University of Minnesota Press.

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. https://doi.org/10.1007/BF02291411

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280. https://doi.org/10.1177/014662168801200305

Carroll, J. D., & Arabie, P. (1998). Multidimensional scaling. In M. H. Birnbaum (Ed.), *Measurement, judgment and decision making* (pp. 179–250). Academic Press. https://doi.org/10.1016/B978-012099975-0.50005-1

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283-319. https://doi.org/10.1007/BF02310791

Cattell, R. B. (1952). The three basic factor-analytic research designs—their interrelations and derivatives. *Psychological Bulletin, 49*(5), 499–520. https://doi.org/10.1037/h0054245

Cattell, H. E., & Mead, A. D. (2008). The sixteen personality factor questionnaire (16PF). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment* (pp. 135–159). SAGE Publications.

Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, *85*(1), 121-153. https://doi.org/10.1007/s11336-019-09693-2

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523-562. https://doi.org/10.1207/S15327906MBR3604_03

Conrad, L. (1977). *Graduate record examinations: Technical manual* (Report No. ED163085). Educational Testing Service.

Costa, P. T., & McCrae, R. R. (2009). The five-factor model and the NEO inventories. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 299–322). Oxford University Press.

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual*. Psychological Assessment Resources.

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. In *Elements of information theory*. John Wiley & Sons.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Crocker, L., & Algina, J. (2008). Reliability and the Classical True Score Model. In *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we?. *Psychological Bulletin*, *74*(1), 68-80. https://doi.org/10.1037/h0029382

Davison, M. L., Kim, S. K., & Ding, S. (2001). *Profile analysis via multidimensional scaling (PAMS): exploring the predominant profile patterns in data* (Report No. ED453270). U.S. Department of Education.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295-311. https://doi.org/10.3102/10769986030003295

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, *91*(6), 1138-1151. https://doi.org/10.1037/0022-3514.91.6.1138

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality*, *56*, 33-58. https://doi.org/10.1016/j.jrp.2014.07.004

Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*(6), 1246-1256.

https://doi.org/10.1037/0022-3514.73.6.1246

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Embretson, S. E., & Reise, S. P. (2000a). Binary IRT Models. In *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Embretson, S. E., & Reise, S. P. (2000b). Polytomous IRT Models. In *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*(4), 238-254. https://doi.org/10.1177/0146621609344844

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. Sage Publications.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*(2), 350-365. https://doi.org/10.1037/0022-3514.78.2.350

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26-42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7-28.

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, *40*(2), 171-178. https://doi.org/10.1016/0021-9681(87)90069-5

Guyer, R., & Thompson, N. A. (2011). *User's manual for XCalibre item response theory calibration software, version 4.1.3*. Assessment Systems Corporation.

Hicks, B. M., Markon, K. E., Patrick, C. J., Krueger, R. F., & Newman, J. P. (2004). Identifying psychopathy subtypes on the basis of personality structure. *Psychological Assessment*, *16*(3), 276-288.

Iacono, W. G., & McGue, M. (2002). Minnesota twin family study. *Twin Research and Human Genetics*, *5*(5), 482-487. https://doi.org/10.1375/twin.5.5.482

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*(4), 336-352. https://doi.org/10.1016/S0005-7894(84)80002-7

Jacobson, N. S., & Truax, P. (1992). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 631–648). American Psychological Association. https://doi.org/10.1037/10109-042

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*(4), 407-415. https://doi.org/10.1016/0197-2456(89)90005-6

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78-89. https://doi.org/10.1016/j.jrp.2014.05.003

Johnston, J., & DiNardo, J. (1997). *Econometric methods*. The McGraw-Hill Companies, Inc.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391-406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kang, T., & Chen, T. T. (2011). Performance of the generalized SX^2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*, 89-96. https://doi.org/10.1007/s12564-010-9082-4

Kim, S. K., Frisby, C. L., & Davison, M. L. (2004). Estimating cognitive profiles using profile analysis via multidimensional scaling (PAMS). *Multivariate Behavioral Research*, *39*(4), 595-624. https://doi.org/10.1207/s15327906mbr3904_2

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie/Journal of Psychology*, *216*(1), 49-58. https://doi.org/10.1027/0044-3409.216.1.49

Kinnunen, M. L., Metsäpelto, R. L., Feldt, T., Kokko, K., Tolvanen, A., Kinnunen, U., Leppänen, E., & Pulkkinen, L. (2012). Personality profiles and health: Longitudinal evidence among Finnish adults. *Scandinavian Journal of Psychology*, *53*(6), 512-522. https://doi.org/10.1111/j.1467-9450.2012.00969.x

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1-27. https://doi.org/10.1007/BF02289565

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115-129. https://doi.org/10.1007/BF02289694

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2014). Assessing individual change using short tests and questionnaires. *Applied Psychological Measurement*, *38*(3), 201-216. https://doi.org/10.1177/0146621613510061

Lassere, M. N., van der Heijde, D., & Johnson, K. R. (2001). Foundations of the minimal clinically important difference for imaging. *The Journal of Rheumatology*, *28*(4), 890-891.

Lattin, J. M., Carroll, J. D., & Green, P. E. (2003a). Cluster Analysis. In *Analyzing Multivariate Data*. Brooks/Cole.

Lattin, J. M., Carroll, J. D., & Green, P. E. (2003b). Multidimensional Scaling. In *Analyzing Multivariate Data*. Brooks/Cole.

Lee, J. E. (2015). *Hypothesis testing for adaptive measurement of individual change* (Unpublished doctoral dissertation). University of Minnesota.

Liu, D., Li, S., Yu, Y., & Moustaki, I. (2021). Assessing partial association between ordinal variables: quantification, visualization, and hypothesis testing. *Journal of the American Statistical Association*, *116*(534), 955-968. https://doi.org/10.1080/01621459.2020.1796394

Lombard, T. J. (2006). *An alignment study of the Minnesota comprehensive assessment-II with state standards in reading and literature for grades 3-8 and 10*. Minnesota Department of Education.

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

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21-39. https://doi.org/10.1037/1082-989X.10.1.21

Maassen, G. H. (2004). The standard error in the Jacobson and Truax Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society*, *10*(6), 888-893. https://doi.org/10.1017/S1355617704106097

Manning, W. H., & DuBois, P. H. (1962). Correlational methods in research on human learning. *Perceptual and Motor Skills*, *15*(2), 287-321. https://doi.org/10.2466/pms.1962.15.2.287

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

McCrae, R. R. (2003). *Personality in adulthood: A five-factor theory perspective*. Guilford Press.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and its Application*, *6*(1), 355-378. https://doi.org/10.1146/annurev-statistics-031017-100325

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.

Merz, E. L., & Roesch, S. C. (2011). A latent profile analysis of the five factor model of personality: Modeling trait interactions. *Personality and Individual Differences*, *51*(8), 915-919. https://doi.org/10.1016/j.paid.2011.07.022

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*, 159-179. https://doi.org/10.1007/BF02294245

Mouelhi, Y., Jouve, E., Castelli, C., & Gentile, S. (2020). How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health and Quality of Life Outcomes*, *18*, 1-17. https://doi.org/10.1186/s12955-020-01344-w

Nydick, S. (2022). *catIrt: Simulate IRT-based computerized adaptive tests* (R package version 0.5.1). The Comprehensive R Archive Network (CRAN). https://doi.org/10.32614/CRAN.package.catIrt

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: an item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289-298. https://doi.org/10.1177/0146621603027004004

Phadke, C. (2017). *Measuring intra-individual change at two or more occasions with hypothesis testing methods* (Unpublished doctoral dissertation). University of Minnesota.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 44(1), 50–57. Cambridge University Press. https://doi.org/10.1017/S0305004100023987

Rao, C. R. (1965). *Linear statistical inference and its applications*. Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49-57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

Reckase, M. D. (1984). Scaling techniques. In R. J. McCallum (Ed.), *Handbook of psychological assessment* (pp. 38–53). Wiley.

Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer.

Reckase, M. D. (1997b). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36. https://doi.org/10.1177/0146621697211002

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, *7*(4), 347-364. https://doi.org/10.1177/107319110000700404

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*(1), 45-58. https://doi.org/10.1177/014662169001400105

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?. *Psychological Methods*, *8*(2), 164-184. https://doi.org/10.1037/1082-989X.8.2.164

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: a comparison of item response theory and covariance structure approaches. *Psychological Methods*, *4*(1), 3-21. https://doi.org/10.1037/1082-989X.4.1.3

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, *61*(2), 102-109. https://doi.org/10.1016/j.jclinepi.2007.03.012

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education*, *7*(2), 159-170. https://doi.org/10.1207/s15324818ame0702_4

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, *34*(S1) 1-97. https://doi.org/10.1007/BF03372160

Schneider, L., Chalmers, R. P., Debelak, R., & Merkle, E. C. (2020). Model selection of nested and non-nested item response models using Vuong tests. *Multivariate*

Behavioral Research, *55*(5), 664-684. https://doi.org/10.1080/00273171.2019.1664280

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125-140. https://doi.org/10.1007/BF02289630

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219-246. https://doi.org/10.1007/BF02289621

Spaans, M., Barendregt, M., Muller, E., de Beurs, E., Nijman, H., & Rinne, T. (2009). MMPI profiles of males accused of severe crimes: A cluster analysis. *Psychology, Crime & Law*, *15*(5), 441-450. https://doi.org/10.1080/10683160802356704

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, *63*(6), 1044-1048. https://doi.org/10.1037//0022-006x.63.6.1044

Tai, M. H., Cooperman, A. W., DeWeese, J. N., & Weiss, D. J. (2023). How do trait change patterns affect the performance of adaptive measurement of change?. *Journal of Computerized Adaptive Testing*, *10*(3). https://doi.org/10.7333/2307-1003032

Tellegen, A. (1982). *Manual for the Multidimensional Personality Questionnaire*. University of Minnesota Press.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle,

G. Matthews, & D. H. Saklofske (Ed.), *The SAGE handbook of personality theory and assessment* (pp. 261-292). Sage Publications.

Tisu, L., Lupșa, D., Vîrgă, D., & Rusu, A. (2020). Personality characteristics, job performance and mental health: The mediating role of work engagement. *Personality and Individual Differences*, *153*, 109644. https://doi.org/10.1016/j.paid.2019.109644

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures anova: In defense of difference scores for data analysis. *Educational and Psychological Measurement*, *72*(1), 37-43. https://doi.org/10.1177/0013164411409929

van der Linden, W. J. (2016). Unidimensional logistic response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 13–30). Chapman and Hall/CRC.

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education*, *4*, 1-23. https://doi.org/10.1186/s40536-016-0025-3

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In *Applied Latent Class Analysis.* Cambridge University Press.

Wang, W. C. (1997). The Teacher Personality Inventory (National Science Council Research Report No. NSC86-2413-H-194-022). The National Science Council.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116. https://doi.org/10.1037/1082-989X.9.1.116

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement*, *42*(3), 221-239. https://doi.org/10.1177/0146621617726787

Wang, C., Weiss, D. J., & Suen, K. Y. (2021). Hypothesis testing methods for multivariate multi-occasion intra-individual change. *Multivariate Behavioral Research*, *56*(3), 459-475. https://doi.org/10.1080/00273171.2020.1730739

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weiss, D. J., & Sahin, A. (2024). *Computerized Adaptive Testing: From Concept to Implementation*. Guilford Publications.

Weiss, D. J., & von Minden, S. (2011). Measuring individual growth with conventional and adaptive tests. *Journal of Methods and Measurement in the Social Sciences*, *2*(2), 80-101. https://doi.org/10.2458/azu_jmmss_v2i2_weiss

Weiss, D. J., Wang, C., Cheville, A. L., Basford, J. R., & DeWeese, J. (2022). Adaptive measurement of change: A novel method to reduce respondent burden and detect significant individual-level change in patient-reported outcome measures. *Archives of Physical Medicine and Rehabilitation*, *103*(5), S43-S52. https://doi.org/10.1016/j.apmr.2021.07.814

Woodcock, R. W. (1977). Woodcock-Johnson Psycho-Educational Battery. Technical Report (Report No. ED171748).

Widiger, T. A. (2017). *The Oxford handbook of the five factor model*. Oxford University Press.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245-262. https://doi.org/10.1177/014662168100500212

Yin, K., Lee, P., Sheldon, O. J., Li, C., & Zhao, J. (2021). Personality profiles based on the FFM: A systematic review with a person-centered approach. *Personality and Individual Differences*, *180*, 110996. https://doi.org/10.1016/j.paid.2021.110996

Zwick, R. (2012). Admissions testing in higher education. In *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 401-423). Routledge.

Appendices

Appendix A. Trait Scores for Simulation Study for 11 and 18 Scales

11 Scales – Multidimensional Personality Questionnaire

Table A1. Simulation Trait Scores for Assessment with 11 Scales – Traits 1 through 5.

Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Magnitude	Location	Number of Differences
0	0	0	0	0	0	High	0
2	2	2	2	2	0	Low	0
0	0	0	0	0	0.5	High	18
0.75	0.75	0.75	0.75	0.75	0.5	Mixed	18
1.5	1.5	1.5	1.5	1.5	0.5	Low	18
-0.75	-0.75	-0.75	-0.75	-0.75	1	High	18
0.75	0.75	0.75	0.75	0.75	1	Mixed	18
1.5	1.5	1.5	1.5	1.5	1	Low	18
-0.75	-0.75	-0.75	-0.75	-0.75	1.5	High	18
0.75	0.75	0.75	0.75	0.75	1.5	Mixed	18
1.5	1.5	1.5	1.5	1.5	1.5	Low	18
0	0	0	0	0	0.5	High	36
0.75	0.75	0.75	0.75	0.75	0.5	Mixed	36
1.5	1.5	1.5	1.5	1.5	0.5	Low	36
-0.75	-0.75	-0.75	-0.75	-0.75	1	High	36
0.75	0.75	0.75	0.75	0.75	1	Mixed	36
1.5	1.5	1.5	1.5	1.5	1	Low	36
-0.75	-0.75	-0.75	-0.75	-0.75	1.5	High	36
0.75	0.75	0.75	0.75	0.75	1.5	Mixed	36
1.5	1.5	1.5	1.5	1.5	1.5	Low	36

Table A2. Simulation Trait Scores for Assessment with 11 Scales – Traits 6 through 10.

Trait 6	Trait 7	Trait 8	Trait 9	Trait 10	Magnitude	Location	Number of Differences
0	0	0	0	0	0	High	0
2	2	2	2	2	0	Low	0
0	0	0	0	0.5	0.5	High	18
0.75	0.75	0.75	0.75	1.25	0.5	Mixed	18
1.5	1.5	1.5	1.5	2	0.5	Low	18
-0.75	-0.75	-0.75	-0.75	0.25	1	High	18
0.75	0.75	0.75	0.75	1.75	1	Mixed	18
1.5	1.5	1.5	1.5	2.5	1	Low	18
-0.75	-0.75	-0.75	-0.75	0.75	1.5	High	18
0.75	0.75	0.75	0.75	2.25	1.5	Mixed	18
1.5	1.5	1.5	1.5	3	1.5	Low	18
0	0.5	0.5	0.5	0.5	0.5	High	36
0.75	1.25	1.25	1.25	1.25	0.5	Mixed	36
1.5	2	2	2	2	0.5	Low	36
-0.75	0.25	0.25	0.25	0.25	1	High	36
0.75	1.75	1.75	1.75	1.75	1	Mixed	36
1.5	2.5	2.5	2.5	2.5	1	Low	36
-0.75	0.75	0.75	0.75	0.75	1.5	High	36
0.75	2.25	2.25	2.25	2.25	1.5	Mixed	36
1.5	3	3	3	3	1.5	Low	36

Table A3. Simulation Trait Scores for Assessment with 11 Scales – Trait 11.

Trait 11	Magnitude	Location	Number of Differences
0	0	High	0
2	0	Low	0
0.5	0.5	High	18
1.25	0.5	Mixed	18
2	0.5	Low	18
0.25	1	High	18
1.75	1	Mixed	18
2	1	Low	18
0.75	1.5	High	18
2.25	1.5	Mixed	18
3	1.5	Low	18
0.5	0.5	High	36
1.25	0.5	Mixed	36
2	0.5	Low	36
0.25	1	High	36
1.75	1	Mixed	36
2.5	1	Low	36
0.75	1.5	High	36
2.25	1.5	Mixed	36
3	1.5	Low	36

18 Scales – Stability Facets

Table A4. Simulation Trait Scores for Assessment with 18 Scales – Traits 1 through 5.

Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Magnitude	Location	Number of Differences
0	0	0	0	0	0	High	0
2	2	2	2	2	0	Low	0
0	0	0	0	0	0.5	High	56
0.75	0.75	0.75	0.75	0.75	0.5	Mixed	56
1.5	1.5	1.5	1.5	1.5	0.5	Low	56
-0.75	-0.75	-0.75	-0.75	-0.75	1	High	56
0.75	0.75	0.75	0.75	0.75	1	Mixed	56
1.5	1.5	1.5	1.5	1.5	1	Low	56
-0.75	-0.75	-0.75	-0.75	-0.75	1.5	High	56
0.75	0.75	0.75	0.75	0.75	1.5	Mixed	56
1.5	1.5	1.5	1.5	1.5	1.5	Low	56
0	0	0	0	0	0.5	High	81
0.75	0.75	0.75	0.75	0.75	0.5	Mixed	81
1.5	1.5	1.5	1.5	1.5	0.5	Low	81
-0.75	-0.75	-0.75	-0.75	-0.75	1	High	81
0.75	0.75	0.75	0.75	0.75	1	Mixed	81
1.5	1.5	1.5	1.5	1.5	1	Low	81
-0.75	-0.75	-0.75	-0.75	-0.75	1.5	High	81
0.75	0.75	0.75	0.75	0.75	1.5	Mixed	81
1.5	1.5	1.5	1.5	1.5	1.5	Low	81

Table A5. Simulation Trait Scores for Assessment with 18 Scales – Traits 6 through 10.

Trait 6	Trait 7	Trait 8	Trait 9	Trait 10	Magnitude	Location	Number of Differences
0	0	0	0	0	0	High	0
2	2	2	2	2	0	Low	0
0	0	0	0	0	0.5	High	56
0.75	0.75	0.75	0.75	0.75	0.5	Mixed	56
1.5	1.5	1.5	1.5	1.5	0.5	Low	56
-0.75	-0.75	-0.75	-0.75	-0.75	1	High	56
0.75	0.75	0.75	0.75	0.75	1	Mixed	56
1.5	1.5	1.5	1.5	1.5	1	Low	56
-0.75	-0.75	-0.75	-0.75	-0.75	1.5	High	56
0.75	0.75	0.75	0.75	0.75	1.5	Mixed	56
1.5	1.5	1.5	1.5	1.5	1.5	Low	56
0	0	0	0	0.5	0.5	High	81
0.75	0.75	0.75	0.75	1.25	0.5	Mixed	81
1.5	1.5	1.5	1.5	2	0.5	Low	81
-0.75	-0.75	-0.75	-0.75	0.25	1	High	81
0.75	0.75	0.75	0.75	1.75	1	Mixed	81
1.5	1.5	1.5	1.5	2.5	1	Low	81
-0.75	-0.75	-0.75	-0.75	0.75	1.5	High	81
0.75	0.75	0.75	0.75	2.25	1.5	Mixed	81
1.5	1.5	1.5	1.5	3	1.5	Low	81

Table A6. Simulation Trait Scores for Assessment with 18 Scales – Traits 11 through 15.

Trait 11	Trait 12	Trait 13	Trait 14	Trait 15	Magnitude	Location	Number of Differences
0	0	0	0	0	0	High	0
2	2	2	2	2	0	Low	0
0	0	0	0	0.5	0.5	High	56
0.75	0.75	0.75	0.75	1.25	0.5	Mixed	56
1.5	1.5	1.5	1.5	2	0.5	Low	56
-0.75	-0.75	-0.75	-0.75	0.25	1	High	56
0.75	0.75	0.75	0.75	1.75	1	Mixed	56
1.5	1.5	1.5	1.5	2.5	1	Low	56
-0.75	-0.75	-0.75	-0.75	0.75	1.5	High	56
0.75	0.75	0.75	0.75	2.25	1.5	Mixed	56
1.5	1.5	1.5	1.5	3	1.5	Low	56
0.5	0.5	0.5	0.5	0.5	0.5	High	81
1.25	1.25	1.25	1.25	1.25	0.5	Mixed	81
2	2	2	2	2	0.5	Low	81
0.25	0.25	0.25	0.25	0.25	1	High	81
1.75	1.75	1.75	1.75	1.75	1	Mixed	81
2.5	2.5	2.5	2.5	2.5	1	Low	81
0.75	0.75	0.75	0.75	0.75	1.5	High	81
2.25	2.25	2.25	2.25	2.25	1.5	Mixed	81
3	3	3	3	3	1.5	Low	81

Table A7. Simulation Trait Scores for Assessment with 18 Scales – Traits 16 through 18.

Trait 16	Trait 17	Trait 18	Magnitude	Location	Number of Differences
0	0	0	0	High	0
2	2	2	0	Low	0
0.5	0.5	0.5	0.5	High	56
1.25	1.25	1.25	0.5	Mixed	56
2	2	2	0.5	Low	56
0.25	0.25	0.25	1	High	56
1.75	1.75	1.75	1	Mixed	56
2.5	2.5	2.5	1	Low	56
0.75	0.75	0.75	1.5	High	56
2.25	2.25	2.25	1.5	Mixed	56
3	3	3	1.5	Low	56
0.5	0.5	0.5	0.5	High	81
1.25	1.25	1.25	0.5	Mixed	81
2	2	2	0.5	Low	81
0.25	0.25	0.25	1	High	81
1.75	1.75	1.75	1	Mixed	81
2.5	2.5	2.5	1	Low	81
0.75	0.75	0.75	1.5	High	81
2.25	2.25	2.25	1.5	Mixed	81
3	3	3	1.5	Low	81

Appendix B. Simulation R Code

Functions

```
#######################################################################
# Function: amc_LRT
# Purpose: conduct adaptive measurement of change likelihood-ratio test
#          or conduct experimental Bayes Factor index test (BFI)
# Arguments:
#     1.) resp.list: a list of responses for an examinee where each
#                    list element corresponds to one trait; each list
#                    element should be numeric
#     2.) param.list: a list of item response model parameters
#                     associated with the items that the examinee
#                     completed where each list element corresponds
#                     to one trait; each list element should be a
#                     matrix
#     3.) mle.theta.vec: a vector of maximum likelihood estimates
#                        (MLEs) for each trait for one examinee; if not
#                        provided, the function estimates scores using
#                        provided responses and item response model
#                        parameters
#     4.) alpha: the Type I error rate for the test
#     5.) model: the type of item response model to use; can be either
#                "brm" for a binary response model like the
#                two-parameter logistic model or "grm" for the graded
#                response model
#     6.) mod: logical indicator to determine whether to use the
#              experimental Bayes Factor test; defaults to FALSE, using
#              the likelihood-ratio test
#     7.) prior: for BFI; use "normal" to use a multivariate normal
#                distribution (MVN) as a prior distribution for trait
#                scores; other types of priors are not currently
#                implemented
#     8.) mean.pool: for BFI, specifies the mean of the normal
#                    distribution that serves as a prior distribution
#                    for the pooled MLE
#     9.) var.pool: for BFI, specifies the variance of the normal
#                   distribution that serves as a prior distribution
#                   for the pooled MLE
#     10.) mean.trait: for BFI, specifies vector of means for MVN
#                      distribution that serves as a prior distribution
#                      for trait scores
#     11.) corr.trait: for BFI, specifies the correlation matrix for
#                      the MVN distribution that serves as a prior
#                      distribution for trait scores
#     12.) threshold: for BFI, threshold value to indicate importance
#######################################################################
amc_LRT <- function(resp.list, param.list, mle.theta.vec = NULL, alpha
= .05,
                    model = "grm", mod = FALSE, prior = c("normal"),
                    mean.pool = NULL, var.pool = NULL, mean.trait =
NULL,
                    corr.trait = NULL, threshold = NULL) {
```

```
  if (length(resp.list) != length(param.list)) {
    stop("Length of response list and item parameter list must match.")
  } else {
    n_tests <- length(resp.list)
  }
  if ((model != "grm") && (model != "brm")) {
    stop("Input 'model' must be either 'grm' or 'brm'.")
  }
  ntraits <- length(param.list)
  if (is.null(mle.theta.vec)) {
    mles <- mapply(lrtcustom(resp.list, param.list, model = model),
                   resp.list, param.list)
  } else {
    mles <- mle.theta.vec
  }
  resps.pooled <- unlist(resp.list)
  params.pooled <- do.call(rbind, param.list)
  mle.pooled <- catIrt::mleEst(resp = resps.pooled, params =
params.pooled,
                               mod = model, range = c(-4, 4))$theta
  if (model == "grm") {
    ll0 <- catIrt:::logLik.grm(u = resps.pooled, theta = mle.pooled,
                               params = params.pooled)
    ll1 <- sum(mapply(catIrt:::logLik.grm, u = resp.list, theta = mles,
                      params = param.list))
  } else {
    ll0 <- catIrt:::logLik.brm(u = resps.pooled, theta = mle.pooled,
                               params = params.pooled)
    ll1 <- sum(mapply(catIrt:::logLik.brm, u = resp.list, theta = mles,
                      params = param.list))
  }
  lrt_stat <- -2 * (ll0 - ll1)
  if (mod) {
    if (is.null(threshold)) {threshold <- -6}
    priors <- c("normal")
    numtype <- sum(tolower(prior) %in% priors)
    if (numtype == 0) {
      stop("Input 'prior' does not contain a valid value. Must specify
\n
           'normal' for 'prior'.")
    } else if (numtype > 1L) {
      stop("Input 'prior' contains two or more values. Must specify
'normal' \n
           for 'prior'.")
    } else {
      prior <- tolower(prior)
    }
    if (is.null(mean.pool)) {mean.pool <- 0}
    if (is.null(var.pool)) {var.pool <- 1}
    poolprob <- 2 * log(dnorm(as.numeric(mle.pooled), mean = mean.pool,
                              sd = sqrt(var.pool)))
    if (is.null(mean.trait)) {mean.trait <- rep(0, ntraits)}
    if (is.null(corr.trait)) {
      corr.trait <- matrix(0, nrow = ntraits, ncol = ntraits)
      diag(corr.trait) <- 1
```

```
    }
    traitprob <- 2 * log(dmvnorm(as.numeric(mles), mean = mean.trait,
                                 sigma = corr.trait))
    testvals <- NULL
    for (ii in 1:ntraits) {
       if (model == "grm") {
         infoval <- test_information(theta = mles[ii],
                                     a = as.matrix(param.list[[ii]][,
1]),
                                     b = param.list[[ii]][, 2:5])
         testvals <- c(testvals, infoval)
       } else {
         infoval <- test_information_2PL(theta = mles[ii],
                                         a =
as.matrix(param.list[[ii]][, 1]),
                                         b =
as.matrix(param.list[[ii]][, 2]))
         testvals <- c(testvals, infoval)
       }
    }
    if (model == "grm") {
      testpool <- test_information(theta = mle.pooled,
                                   a = as.matrix(params.pooled[, 1]),
                                   b = params.pooled[, 2:5])
    } else {
      testpool <- test_information_2PL(theta = mle.pooled,
                                   a = as.matrix(params.pooled[, 1]),
                                   b = as.matrix(params.pooled[, 2]))
    }
    testinfo <- log(testpool)
    traitinfos <- sum(log(testvals))
    constant <- (1 - ntraits) * log(2 * pi)
    C_stat <- constant - traitprob + traitinfos - lrt_stat + poolprob -
testinfo
    brt_decision <- as.numeric(C_stat <= threshold)
    return(list(constant = constant, traitprob = traitprob,
                traitinfos = traitinfos, poolprob = poolprob,
                testinfo = testinfo, lrt_stat = lrt_stat, C_stat =
C_stat,
                lrt_decision = brt_decision, mle.pooled = mle.pooled))
  } else {
    lrt_p <- pchisq(lrt_stat, n_tests - 1, lower.tail = FALSE)
    lrt_decision <- as.numeric(lrt_p < alpha)
    return(list(ll0 = ll0, ll1 = ll1, lrt_stat = lrt_stat, lrt_p =
lrt_p,
                lrt_decision = lrt_decision, mle.pooled = mle.pooled))
  }
}

#######################################################################
# Function: amc_score
# Purpose: conduct adaptive measurement of change score test
# Arguments:
#     1.) resp.list: a list of responses for an examinee where each
```

```
#                     list element corresponds to one trait; each list
#                     element should be numeric
#     2.) param.list: a list of item response model parameters
#                      associated with the items that the examinee
#                      completed where each list element corresponds to
#                      one trait; each list element should be a matrix
#     3.) mle.theta.vec: a vector of maximum likelihood estimates
#                         (MLEs) for each trait for one examinee; if not
#                         provided, the function estimates scores using
#                         provided responses and item response model
#                         parameters
#     4.) alpha: the Type I error rate for the test
#     5.) model: the type of item response model to use; can be either
#                "brm" for a binary response model like the
#                two-parameter logistic model or "grm" for the graded
#                response model
#######################################################################
amc_score <- function(resp.list, param.list, mle.theta.vec = NULL,
                      model = "grm", alpha = .05) {
  if (length(resp.list) != length(param.list)) {
    stop("Length of response list and item parameter list must match.")
  } else {
    n_tests <- length(resp.list)
  }
  if ((model != "grm") && (model != "brm")) {
    stop("Input 'model' must be either 'grm' or 'brm'.")
  }
  if (is.null(mle.theta.vec)) {
    mles <- mapply(lrtcustom(resp.list, param.list, model = model),
resp.list,
                   param.list)
  } else {
    mles <- mle.theta.vec
  }
  resps.pooled <- unlist(resp.list)
  params.pooled <- do.call(rbind, param.list)
  mle.pooled <- catIrt::mleEst(resp = resps.pooled, params =
params.pooled,
                               mod = model, range = c(-4, 4))$theta
  testvals <- NULL
  for (ii in 1:n_tests) {
     if (model == "grm") {
       infoval <- test_information(theta = mles[ii],
                                   a = as.matrix(param.list[[ii]][,
1]),
                                   b = param.list[[ii]][, 2:5])
     } else {
       infoval <- test_information_2PL(theta = mles[ii],
                                       a = as.matrix(param.list[[ii]][,
1]),
                                       b = as.matrix(param.list[[ii]][,
2]))
     }
     testvals <- c(testvals, infoval)
  }
```

```
  if (model == "grm") {
    testpool <- test_information(theta = mle.pooled,
                                 a = as.matrix(params.pooled[, 1]),
                                 b = params.pooled[, 2:5])
  } else {
    testpool <- test_information_2PL(theta = mle.pooled,
                                     a = as.matrix(params.pooled[, 1]),
                                     b = as.matrix(params.pooled[, 2]))
  }
  spieces <- NULL
  for (jj in 1:n_tests) {
     resps <- resp.list[[jj]]
     params <- param.list[[jj]]
     if (model == "grm") {
       scval <- score(resps = resps, theta = mle.pooled, params =
params)
     } else {
       scval <- score_2PL(resps = resps, theta = mle.pooled, params =
params)
     }
     spval <- (scval^2) / (testvals[jj])
     spieces <- c(spieces, spval)
  }
  s_stat <- sum(spieces)
  s_p <- pchisq(s_stat, n_tests - 1, lower.tail = FALSE)
  s_decision <- as.numeric(s_p < alpha)
  return(list(s_stat = s_stat, s_p = s_p, s_decision = s_decision,
              mle.pooled = mle.pooled))
}

####################################################################
# Function: amc_Z
# Purpose: conduct adaptive measurement of change Z test
# Arguments:
#     1.) resp.list: a list of responses for an examinee where each
#                    list element corresponds to one trait; each list
#                    element should be numeric
#     2.) param.list: a list of item response model parameters
#                     associated with the items that the examinee
#                     completed where each list element corresponds to
#                     one trait; each list element should be a matrix
#     3.) mle.theta.vec: a vector of maximum likelihood estimates
#                        (MLEs) for each trait for one examinee; if not
#                        provided, the function estimates scores using
#                        provided responses and item response model
#                        parameters
#     4.) alpha: the Type I error rate for the test
#     5.) model: the type of item response model to use; can be either
#                "brm" for a binary response model like the
#                two-parameter logistic model or "grm" for the graded
#                response model
#     6.) mod: logical indicating whether to modify the degrees of
#              freedom of the test to be H + ceiling(H/3); experimental
####################################################################
```

```
amc_Z <- function(resp.list, param.list, mle.theta.vec = NULL, model =
"grm",
                  alpha = .05, mod = F) {
  if (length(resp.list) != length(param.list)) {
    stop("Length of response list and item parameter list must match.")
  } else {
    n_tests <- length(resp.list)
  }
  if ((model != "grm") && (model != "brm")) {
    stop("Input 'model' must be either 'grm' or 'brm'.")
  }
  if (is.null(mle.theta.vec)) {
    mles <- mapply(lrtcustom(resp.list, param.list, model = model),
resp.list,
                   param.list)
  } else {
    mles <- mle.theta.vec
  }
  resps.pooled <- unlist(resp.list)
  params.pooled <- do.call(rbind, param.list)
  mle.pooled <- catIrt::mleEst(resp = resps.pooled, params =
params.pooled,
                               mod = model, range = c(-4, 4))$theta
  testvals <- NULL
  for (ii in 1:n_tests) {
     if (model == "grm") {
       infoval <- test_information(theta = mle.pooled,
                                   a = as.matrix(param.list[[ii]][,
1]),
                                   b = param.list[[ii]][, 2:5])
     } else {
       infoval <- test_information_2PL(theta = mle.pooled,
                                       a = as.matrix(param.list[[ii]][,
1]),
                                       b = as.matrix(param.list[[ii]][,
2]))
     }
     testvals <- c(testvals, infoval)
  }
  data <- cbind(mles, testvals)
  npairs <- t(combn(1:n_tests, 2))
  H <- nrow(npairs)
  zvals <- NULL
  for (i in 1:nrow(npairs)) {
     val1 <- npairs[i, ][1]
     val2 <- npairs[i, ][2]
     mle1 <- data[val1, 1]
     mle2 <- data[val2, 1]
     itinf1 <- 1 / data[val1, 2]
     itinf2 <- 1 / data[val2, 2]
     modf <- 1 / sqrt(itinf1 + itinf2)
     dist <- abs(mle2 - mle1)
     Zobs <- dist * modf
     zvals <- c(zvals, Zobs)
  }
```

```
  z_stat <- sum(zvals^2)
  if (mod) {
    dfval <- H + ceiling(H/3)
  } else {
    dfval <- H
  }
  z_p <- pchisq(z_stat, dfval, lower.tail = FALSE)
  z_decision <- as.numeric(z_p < alpha)
  return(list(z_stat = z_stat, z_p = z_p, z_decision = z_decision,
              mle.pooled = mle.pooled))
}

######################################################################
# Function: item_information_2PL
# Purpose: calculate item information for two-parameter logistic (2PL)
#          model
# Arguments:
#     1.) theta: trait score for an examinee
#     2.) a: item discrimination parameter for an item
#     3.) b: item difficulty/location parameter for an item
######################################################################
item_information_2PL <- function(theta, a, b) {
  infos <- (a^2) * Pstar(theta, a, b) * (1 - Pstar(theta, a, b))
  return(infos)
}

######################################################################
# Function: item_information
# Purpose: calculate item information for a graded response model (GRM)
# Arguments:
#     1.) theta: trait score for an examinee
#     2.) a: item discrimination parameter for an item
#     3.) b: numeric containing boundary parameters for an item
######################################################################
item_information <- function(theta, a, b) {
  infos <- 0
  for (i in 0:length(b)) {
     if (i == 0) {
       numerator <- (0 - Pstarp(theta, a, b[i + 1]))^2
       denominator <- 1 - Pstar(theta, a, b[i + 1])
       infos <- infos + (numerator / denominator)
     } else if (i == length(b)) {
       numerator <- (Pstarp(theta, a, b[i]) - 0)^2
       denominator <- Pstar(theta, a, b[i]) - 0
       infos <- infos + (numerator / denominator)
     } else {
       numerator <- (Pstarp(theta, a, b[i]) - Pstarp(theta, a, b[i +
1]))^2
       denominator <- Pstar(theta, a, b[i]) - Pstar(theta, a, b[i + 1])
       infos <- infos + (numerator / denominator)
     }
  }
  return(infos)
}
```

```
########################################################################
# Function: lrtcustom
# Purpose: utility function to estimate trait scores using maximum
#          likelihood estimation
# Arguments:
#     1.) resp.list: a list of responses for an examinee where each
#                    list element corresponds to one trait; each list
#                    element should be numeric
#     2.) param.list: a list of item response model parameters
#                     associated with the items that the examinee
#                     completed where each list element corresponds to
#                     one trait; each list element should be a matrix
#     3.) model: the type of item response model to use; can be either
#                "brm" for a binary response model like the
#                two-parameter logistic model or "grm" for the graded
#                response model
########################################################################
lrtcustom <- function(resp.list, param.list, model = "grm") {
  return(catIrt::mleEst(resp = resp.list, params = param.list, mod =
model,
                        range = c(-4, 4))$theta)
}

########################################################################
# Function: Pstar
# Purpose: calculate probability of a keyed response
# Arguments:
#     1.) theta: trait score for an examinee
#     2.) a: item discrimination parameter for an item
#     3.) b: item difficulty/location or boundary parameter for an item
########################################################################
Pstar <- function(theta, a, bk) {return(1 / (1 + exp(-1 * a * (theta -
bk))))}

########################################################################
# Function: Pstarp
# Purpose: calculate first derivative of the probability of a keyed
#          response
# Arguments:
#     1.) theta: trait score for an examinee
#     2.) a: item discrimination parameter for an item
#     3.) b: item difficulty/location or boundary parameter for an item
########################################################################
Pstarp <- function(theta, a, bk) {
  return((a * exp(a * (theta - bk))) / (1 + exp(a * (theta - bk)))^2)
}

########################################################################
# Function: score_2PL
# Purpose: calculate score statistic for two-parameter logistic model
#          (2PL)
# Arguments:
#     1.) resps: numeric containing responses for an examinee
#     2.) theta: trait score for an examinee
#     3.) params: matrix containing item parameters for the 2PL where
```

```
#                  discriminations are in the first column and where
#                  difficulties/locations are in the second column
########################################################################
score_2PL <- function(resps, theta, params) {
  iscores <- NULL
  for (gg in 1:length(resps)) {
     x <- resps[gg]
     a <- params[gg, 1]
     bg <- params[gg, 2]
     if (x == 0) {
       numerator <- -1 * a * exp(a * (theta - bg))
       denominator <- 1 + exp(a * (theta - bg))
       out <- (numerator / denominator)
     } else {
       numerator <- a
       denominator <- 1 + exp(a * (theta - bg))
       out <- (numerator / denominator)
     }
     iscores <- c(iscores, out)
  }
  scoreval <- sum(iscores)
  return(scoreval)
}

########################################################################
# Function: score
# Purpose: calculate score statistic for the graded response model
#          (GRM)
# Arguments:
#     1.) resps: numeric containing responses for an examinee
#     2.) theta: trait score for an examinee
#     3.) params: matrix containing item parameters for the GRM where
#                  discriminations are in the first column and where
#                  boundary parameters are in the remaining columns;
#                  only works currently for items with five responses
########################################################################
score <- function(resps, theta, params) {
  iscores <- NULL
  for (gg in 1:length(resps)) {
     x <- resps[gg]
     a <- params[gg, 1]
     b <- params[gg, 2:5]
     if (x == 1) {
       bg <- b[1]
       numerator <- -1 * a * exp(a * (theta - bg))
       denominator <- 1 + exp(a * (theta - bg))
       out <- (numerator / denominator)
     } else if (x == 5) {
       bg <- b[4]
       numerator <- a
       denominator <- 1 + exp(a * (theta - bg))
       out <- (numerator / denominator)
     } else {
       if (x == 2) {
         bg1 <- b[1]
```

```
        bg2 <- b[2]
      } else if (x == 3) {
        bg1 <- b[2]
        bg2 <- b[3]
      } else {
        bg1 <- b[3]
        bg2 <- b[4]
      }
      num <- -1 * a * (exp((a * (theta - bg1)) + (a * (theta - bg2)))
- 1)
      denom <- (1 + exp(a * (theta - bg1))) * (1 + exp(a * (theta -
bg2)))
      out <- (num / denom)
    }
    iscores <- c(iscores, out)
  }
  scoreval <- sum(iscores)
  return(scoreval)
}

########################################################################
# Function: simAPDA
# Purpose: primary function that runs core simulation
# Arguments:
#     1.) nsimulees: number of simulees for the simulation
#     2.) ncond: number of conditions for different true score patterns
#                that specify levels of magnitude and scale location
#     3.) nitemx: vector containing number of items for test length
#     4.) alphaval: nominal Type I error rate for AMC omnibus
#                   hypothesis tests
#     5.) nitems: number of items to generate for simulated item banks
#     6.) nSEM: standard error termination values for variable-length
#               computerized adaptive tests (VL-CATs)
#     7.) thresval: threshold value for experimental Bayes Factor Index
#                   (BFI)
#     8.) corrmat: for BFI, specifies correlation matrix for the
#                  multivariate normal distribution (MVN) that serves
#                  as a prior distribution for trait scores
#     9.) nullvar: for BFI, specifies the variance of the normal
#                  distribution that serves as a prior distribution for
#                  the pooled trait estimate
#     10.) priorname: for BFI; use "normal" to use a MVN distribution
#                     as a prior distribution for trait scores; other
#                     types of priors are not currently implemented
#     11.) penalties: defunct parameter used for experimental penalty
#                     work
#     12.) temp.start: integer specifying true score pattern to begin
#                      with in the simulation, among the 1:ncond such
#                      patterns; can be used to skip some patterns to
#                      examine others more quickly
#     13.) temp.end: integer specifying true score pattern to end with
#                    in the simulation, among the 1:ncond such
#                    patterns; can be used to skip some patterns to
#                    examine others more quickly
#     14.) vlcat: logical specifying whether to use VL-CATs
```

```
#     15.) betterbank: logical specifying whether to use ideal
#                      discriminations
#     16.) equalbank: logical specifying whether to make simulated item
#                     banks flatter in their bank information functions
#     17.) numtraits: character indicating number of scales to use:
#                     'bigfive' refers to 5 scales; 'mpq', to 11
#                     scales; and 'facets', to 18 scales
#     18.) finalname: character indicating name of output files created
#                     after each true score pattern condition is
#                     complete
#     19.) realtestinfos0: character indicating location of observed
#                          relative test information functions from
#                          real data analyses
########################################################################
simAPDA <- function(nsimulees = 100, ncond = 20, nitemx = c(10, 30,
50),
                    alphaval = 0.05, nitems = 100, nSEM = c(0.4, 0.3,
0.2),
                    thresval = 0, corrmat = NULL, nullvar = 0.85,
                    priorname = "normal", penalties = c(0), temp.start
= 1,
                    temp.end = 20, vlcat = T, betterbank = F,
                    equalbank = F, numtraits = c("bigfive", "mpq",
"facets"),
                    finalname = "sim.csv", realtestinfos0 = getwd()) {
  types <- c("bigfive", "mpq", "facets")
  numtype <- sum(tolower(numtraits) %in% types)
  if (numtype == 0) {
    stop("Input 'numtraits' does not contain a valid value.")
  } else if (numtype > 2L) {
    stop("Input 'numtraits' does not contain a valid value. \n
         Provide a single value.")
  } else if (numtype == 3L) {
    numtraits <- "bigfive"
  } else {
    numtraits <- tolower(numtraits)
  }
  n1 <- c("none", "ihigh", "none")
  n2 <- c("none", "ilow", "none")
  n3 <- c("mag05", "ihigh", "four")
  n4 <- c("mag05", "imix", "four")
  n5 <- c("mag05", "ilow", "four")
  n6 <- c("mag10", "ihigh", "four")
  n7 <- c("mag10", "imix", "four")
  n8 <- c("mag10", "ilow", "four")
  n9 <- c("mag15", "ihigh", "four")
  n10 <- c("mag15", "imix", "four")
  n11 <- c("mag15", "ilow", "four")
  n12 <- c("mag05", "ihigh", "six")
  n13 <- c("mag05", "imix", "six")
  n14 <- c("mag05", "ilow", "six")
  n15 <- c("mag10", "ihigh", "six")
  n16 <- c("mag10", "imix", "six")
  n17 <- c("mag10", "ilow", "six")
  n18 <- c("mag15", "ihigh", "six")
```

```
  n19 <- c("mag15", "imix", "six")
  n20 <- c("mag15", "ilow", "six")
  if (numtraits == "bigfive") {
    ntraits <- 5
    realtestinfos <- paste0(realtestinfos0, "ipip_relative_real.csv")
    dmeans <- c(0.824, 0.85, 1.187, 1.032, 1.053)
    b1means <- c(-0.525, -1.238, -1.501, -0.637, -1.498)
    b2means <- c(0.979, 0.17, -0.449, 0.677, -0.328)
    b3means <- c(2.178, 1.271, 0.348, 1.589, 0.474)
    b4means <- c(3.456, 2.587, 1.312, 2.724, 1.591)
    dsds <- c(0.221, 0.207, 0.212, 0.378, 0.28)
    b1sds <- c(0.498, 0.38, 0.466, 0.339, 0.544)
    b2sds <- c(0.595, 0.661, 0.471, 0.151, 0.518)
    b3sds <- c(0.715, 1.012, 0.479, 0.29, 0.522)
    b4sds <- c(1.012, 1.363, 0.551, 0.638, 0.572)
    NBanks <- vector(mode = "list", length = ntraits)
    set.seed(123)
    for (i in 1:ntraits) {
      NBanks[[i]] <- cbind(a = rnorm(nitems, mean = dmeans[i], sd =
dsds[i]),
                          b1 = rnorm(nitems, mean = b1means[i], sd =
b1sds[i]),
                          b2 = rnorm(nitems, mean = b2means[i], sd =
b2sds[i]),
                          b3 = rnorm(nitems, mean = b3means[i], sd =
b3sds[i]),
                          b4 = rnorm(nitems, mean = b4means[i], sd =
b4sds[i]))
    }
    dmin <- c(0.481, 0.499, 0.894, 0.468, 0.641)
    dmax <- c(1.305, 1.145, 1.485, 1.813, 1.436)
    b1min <- c(-1.643, -1.953, -2.318, -1.356, -2.368)
    b1max <- c(-0.031, -0.642, -0.894, -0.085, -0.488)
    b2min <- c(-0.066, -0.901, -1.133, 0.37, -0.894)
    b2max <- c(2.245, 1.598, 0.418, 0.94, 0.819)
    b3min <- c(1.051, -0.067, -0.272, 1.263, -0.096)
    b3max <- c(3.695, 3.138, 1.241, 2.143, 1.565)
    b4min <- c(2.319, 1.017, 0.613, 2.101, 0.811)
    b4max <- c(5.663, 5.274, 2.391, 3.897, 2.635)
    UBanks <- vector(mode = "list", length = ntraits)
    for (i in 1:ntraits) {
      UBanks[[i]] <- cbind(a = runif(nitems, min = dmin[i], max =
dmax[i]),
                           b1 = runif(nitems, min = b1min[i], max =
b1max[i]),
                           b2 = runif(nitems, min = b2min[i], max =
b2max[i]),
                           b3 = runif(nitems, min = b3min[i], max =
b3max[i]),
                           b4 = runif(nitems, min = b4min[i], max =
b4max[i]))
    }
    if (betterbank) {
      dmin <- rep(1.1, ntraits)
      dmax <- rep(2.8, ntraits)
```

```
    } else {
      dmin <- c(0.713, 0.733, 1.033, 0.909, 0.834)
      dmax <- c(0.914, 0.990, 1.319, 1.177, 1.253)
    }
    if (equalbank) {
      b1min <- rep(-3, ntraits)
      b1max <- rep(-1.5001, ntraits)
      b2min <- rep(-1.5, ntraits)
      b2max <- rep(-0.001, ntraits)
      b3min <- rep(0, ntraits)
      b3max <- rep(1.5, ntraits)
      b4min <- rep(1.5001, ntraits)
      b4max <- rep(3, ntraits)
    } else {
      b1min <- c(-0.794, -1.474, -1.806, -0.696, -1.631)
      b1max <- c(-0.230, -1.005, -1.136, -0.548, -1.393)
      b2min <- c(0.732, -0.054, -0.802, 0.62, -0.696)
      b2max <- c(1.155, 0.380, -0.143, 0.766, -0.036)
      b3min <- c(1.885, 0.572, 0.07, 1.409, 0.105)
      b3max <- c(2.290, 1.881, 0.663, 1.753, 0.659)
      b4min <- c(2.934, 1.882, 0.882, 2.226, 1.224)
      b4max <- c(3.541, 3.157, 1.55, 2.91, 1.685)
    }
    U2Banks <- vector(mode = "list", length = ntraits)
    for (i in 1:ntraits) {
       U2Banks[[i]] <- cbind(a = runif(nitems, min = dmin[i], max =
dmax[i]),
                             b1 = runif(nitems, min = b1min[i], max =
b1max[i]),
                             b2 = runif(nitems, min = b2min[i], max =
b2max[i]),
                             b3 = runif(nitems, min = b3min[i], max =
b3max[i]),
                             b4 = runif(nitems, min = b4min[i], max =
b4max[i]))
    }
    AllBanks <- list(NBanks, UBanks, U2Banks)
    theta_values <- seq(-4, 4, by = 0.05)
    dists <- vector(mode = "list", length = 3)
    for (hh in 1:3) {
       relatives <- matrix(0, nrow = length(theta_values), ncol =
ntraits)
       for (gg in 1:ntraits) {
          storage <- matrix(0, nrow = length(theta_values), ncol =
nitems)
          for (j in 1:nitems) {
             storage[, j] <- sapply(theta_values, item_information,
                                    a = AllBanks[[hh]][[gg]][j, 1],
                                    b = AllBanks[[hh]][[gg]][j, 2:5])
          }
          bankinfo <- apply(storage, 1, sum)
          relatives[, gg] <- bankinfo / max(bankinfo)
       }
       dists[[hh]] <- relatives
    }
```

```
    relreal <- read.csv(realtestinfos, header = FALSE)
    par(mfrow = c(2, 3))
    for (gg in 1:ntraits) {
       plot(theta_values, dists[[1]][, gg], type = "l", col = "blue",
lwd = 2,
            ylab = "Relative Information", xlab = "Theta",
            main = "Simulated and Real Information")
            lines(theta_values, dists[[2]][, gg], col = "green", lwd =
2)
            lines(theta_values, dists[[3]][, gg], col = "purple", lwd =
2)
            lines(theta_values, relreal[, gg], col = "red", lwd = 2)
            legend("topleft", legend = c("Normal", "MinMax Uniform",
                                         "Robust Uniform", "Real Test
Info."),
                   col = c("blue", "green", "purple", "red"), lwd = 2,
                   bty = "n", cex = 0.5)
    }
    traitvals <- matrix(0, nrow = 20, ncol = ntraits)
    traitvals[1, ] <- rep(0, 5)
    traitvals[2, ] <- rep(2, 5)
    traitvals[3, ] <- c(0, 0, 0, 0, 0.5)
    traitvals[4, ] <- c(0.75, 0.75, 0.75, 0.75, 1.25)
    traitvals[5, ] <- c(1.5, 1.5, 1.5, 1.5, 2)
    traitvals[6, ] <- c(-0.75, -0.75, -0.75, -0.75, 0.25)
    traitvals[7, ] <- c(0.75, 0.75, 0.75, 0.75, 1.75)
    traitvals[8, ] <- c(1.5, 1.5, 1.5, 1.5, 2)
    traitvals[9, ] <- c(-0.75, -0.75, -0.75, -0.75, 0.75)
    traitvals[10, ] <- c(0.75, 0.75, 0.75, 0.75, 2.25)
    traitvals[11, ] <- c(1.5, 1.5, 1.5, 1.5, 3)
    traitvals[12, ] <- c(0, 0, 0, 0.5, 0.5)
    traitvals[13, ] <- c(0.75, 0.75, 0.75, 1.25, 1.25)
    traitvals[14, ] <- c(1.5, 1.5, 1.5, 2, 2)
    traitvals[15, ] <- c(-0.75, -0.75, -0.75, 0.25, 0.25)
    traitvals[16, ] <- c(0.75, 0.75, 0.75, 1.75, 1.75)
    traitvals[17, ] <- c(1.5, 1.5, 1.5, 2.5, 2.5)
    traitvals[18, ] <- c(-0.75, -0.75, -0.75, 0.75, 0.75)
    traitvals[19, ] <- c(0.75, 0.75, 0.75, 2.25, 2.25)
    traitvals[20, ] <- c(1.5, 1.5, 1.5, 3, 3)
    designs <- rbind(n1, n2, n3, n4, n5, n6, n7, n8, n9, n10, n11, n12,
n13,
                     n14, n15, n16, n17, n18, n19, n20)
    setup <- data.frame(trait1 = traitvals[, 1], trait2 = traitvals[,
2],
                        trait3 = traitvals[, 3], trait4 = traitvals[,
4],
                        trait5 = traitvals[, 5], mag = designs[, 1],
                        info = designs[, 2], ndiff = designs[, 3])
  } else if (numtraits == "mpq") {
    ntraits <- 11
    realtestinfos <- paste0(realtestinfos0, "mpq_relative_real.csv")
    dmeans <- c(0.692, 0.760, 0.817, 0.843, 0.765, 0.775, 0.868, 0.952,
0.924,
                0.676, 0.982)
```

```
    bmeans <- c(-0.192, 0.362, -1.378, -1.771, 0.765, 1.039, 0.813, -
0.545,
                -0.298, 1.014, 1.098)
    dsds <- c(0.125, 0.219, 0.268, 0.198, 0.256, 0.195, 0.264, 0.280,
0.205,
              0.230, 0.293)
    bsds <- c(0.803, 0.693, 0.675, 0.472, 0.695, 0.525, 0.859, 0.618,
0.472,
              0.745, 0.532)
    NBanks <- vector(mode = "list", length = ntraits)
    set.seed(123)
    for (i in 1:ntraits) {
       NBanks[[i]] <- cbind(a = rnorm(nitems, mean = dmeans[i], sd =
dsds[i]),
                            b = rnorm(nitems, mean = bmeans[i], sd =
bsds[i]),
                            c = rep(0, nitems))
    }
    dmin <- c(0.507, 0.349, 0.383, 0.467, 0.406, 0.416, 0.450, 0.613,
0.530,
              0.415, 0.585)
    dmax <- c(1.041, 1.153, 1.546, 1.341, 1.299, 1.332, 1.426, 1.672,
1.518,
              1.490, 1.738)
    bmin <- c(-1.869, -1.023, -2.244, -2.691, -0.140, 0.059, -0.703, -
1.757,
              -1.020, -1.116, -0.516)
    bmax <- c(1.310, 1.409, 0.408, -0.890, 2.207, 2.159, 2.563, 0.587,
0.569,
              2.725, 1.943)
    UBanks <- vector(mode = "list", length = ntraits)
    for (i in 1:ntraits) {
       UBanks[[i]] <- cbind(a = runif(nitems, min = dmin[i], max =
dmax[i]),
                            b = runif(nitems, min = bmin[i], max =
bmax[i]),
                            c = rep(0, nitems))
    }
    if (betterbank) {
      dmin <- rep(1.1, ntraits)
      dmax <- rep(2.8, ntraits)
    } else {
      dmin <- c(0.634, 0.592, 0.640, 0.733, 0.577, 0.672, 0.7042,
0.779, 0.803,
                0.513, 0.800)
      dmax <- c(0.735, 0.960, 0.931, 0.948, 0.987, 0.875, 1.014, 0.998,
1.027,
                0.818, 1.151)
    }
    if (equalbank) {
      b1min <- rep(-3, ntraits)
      b1max <- rep(3, ntraits)
    } else {
      bmin <- c(-0.696, -0.200, -1.911, -2.089, 0.152, 0.584, 0.228, -
0.919,
```

```
                -0.660, 0.580, 0.789)
      bmax <- c(0.407, 1.053, -0.885, -1.442, 1.258, 1.363, 1.469,
0.0193,
                0.058, 1.498, 1.497)
    }
    U2Banks <- vector(mode = "list", length = ntraits)
    for (i in 1:ntraits) {
       U2Banks[[i]] <- cbind(a = runif(nitems, min = dmin[i], max =
dmax[i]),
                             b = runif(nitems, min = bmin[i], max =
bmax[i]),
                             c = rep(0, nitems))
    }
    AllBanks <- list(NBanks, UBanks, U2Banks)
    theta_values <- seq(-4, 4, by = 0.05)
    dists <- vector(mode = "list", length = 3)
    for (hh in 1:3) {
       relatives <- matrix(0, nrow = length(theta_values), ncol =
ntraits)
       for (gg in 1:ntraits) {
          storage <- matrix(0, nrow = length(theta_values), ncol =
nitems)
          for (j in 1:nitems) {
             storage[, j] <- sapply(theta_values, item_information_2PL,
                                    a = AllBanks[[hh]][[gg]][j, 1],
                                    b = AllBanks[[hh]][[gg]][j, 2])
          }
          bankinfo <- apply(storage, 1, sum)
          relatives[, gg] <- bankinfo / max(bankinfo)
       }
       dists[[hh]] <- relatives
    }
    relreal <- read.csv(realtestinfos, header = FALSE)
    par(mfrow = c(2, 3))
    for (gg in 1:ntraits) {
       plot(theta_values, dists[[1]][, gg], type = "l", col = "blue",
lwd = 2,
            ylab = "Relative Information", xlab = "Theta",
            main = "Simulated and Real Information")
       lines(theta_values, dists[[2]][, gg], col = "green", lwd = 2)
       lines(theta_values, dists[[3]][, gg], col = "purple", lwd = 2)
       lines(theta_values, relreal[, gg], col = "red", lwd = 2)
       legend("topleft", legend = c("Normal", "MinMax Uniform",
                                    "Robust Uniform", "Real Test
Info."),
              col = c("blue", "green", "purple", "red"), lwd = 2, bty =
"n",
              cex = 0.5)
    }
    traitvals <- matrix(0, nrow = 20, ncol = ntraits)
    traitvals[1, ] <- rep(0, 11)
    traitvals[2, ] <- rep(2, 11)
    traitvals[3, ] <- c(rep(0, 9), rep(0.5, 2))
    traitvals[4, ] <- c(rep(0.75, 9), rep(1.25, 2))
    traitvals[5, ] <- c(rep(1.5, 9), rep(2, 2))
```

```
    traitvals[6, ] <- c(rep(-0.75, 9), rep(0.25, 2))
    traitvals[7, ] <- c(rep(0.75, 9), rep(1.75, 2))
    traitvals[8, ] <- c(rep(1.5, 9), rep(2.5, 2))
    traitvals[9, ] <- c(rep(-0.75, 9), rep(0.75, 2))
    traitvals[10, ] <- c(rep(0.75, 9), rep(2.25, 2))
    traitvals[11, ] <- c(rep(1.5, 9), rep(3, 2))
    traitvals[12, ] <- c(rep(0, 6), rep(0.5, 5))
    traitvals[13, ] <- c(rep(0.75, 6), rep(1.25, 5))
    traitvals[14, ] <- c(rep(1.5, 6), rep(2, 5))
    traitvals[15, ] <- c(rep(-0.75, 6), rep(0.25, 5))
    traitvals[16, ] <- c(rep(0.75, 6), rep(1.75, 5))
    traitvals[17, ] <- c(rep(1.5, 6), rep(2.5, 5))
    traitvals[18, ] <- c(rep(-0.75, 6), rep(0.75, 5))
    traitvals[19, ] <- c(rep(0.75, 6), rep(2.25, 5))
    traitvals[20, ] <- c(rep(1.5, 6), rep(3, 5))
    designs <- rbind(n1, n2, n3, n4, n5, n6, n7, n8, n9, n10, n11, n12,
n13,
                     n14, n15, n16, n17, n18, n19, n20)
    setup <- data.frame(trait1 = traitvals[, 1], trait2 = traitvals[,
2],
                        trait3 = traitvals[, 3], trait4 = traitvals[,
4],
                        trait5 = traitvals[, 5], trait6 = traitvals[,
6],
                        trait7 = traitvals[, 7], trait8 = traitvals[,
8],
                        trait9 = traitvals[, 9], trait10 = traitvals[,
10],
                        trait11 = traitvals[, 11], mag = designs[, 1],
                        info = designs[, 2], ndiff = designs[, 3])
  } else {
    ntraits <- 18
    realtestinfos <- paste0(realtestinfos0, "facets_relative_real.csv")
    if (betterbank) {
      dmin <- rep(1.1, ntraits)
      dmax <- rep(2.8, ntraits)
    } else {
      dmin <- c(0.950, 0.678, 0.766, 0.571, 0.35, 0.567, 0.789, 0.923,
0.737,
                0.741, 0.953, 0.572, 0.931, 1.119, 1.158, 0.858, 0.621,
0.822)
      dmax <- c(1.399, 1.193, 1.084, 1.035, 0.611, 0.812, 1.081, 1.244,
0.991,
                1.307, 1.423, 1.612, 1.194, 1.729, 1.695, 1.015, 0.957,
1.185)
    }
    if (equalbank) {
      b1min <- rep(-3, ntraits)
      b1max <- rep(-1.5001, ntraits)
      b2min <- rep(-1.5, ntraits)
      b2max <- rep(-0.001, ntraits)
      b3min <- rep(0, ntraits)
      b3max <- rep(1.5, ntraits)
      b4min <- rep(1.5001, ntraits)
      b4max <- rep(3, ntraits)
```

```
    } else {
      b1min <- c(-2.316, -3.487, -3.58, -2.97, -4.041, -3.115,
                 -3.532, -2.375, -3.839, -3.392, -1.983, -3.07,
                 -2.13, -1.57, -1.491, -1.877, -3.031, -1.694)
      b1max <- c(-1.782, -2.523, -3.053, -1.981, -2.234, -2.492,
                 -3.149, -1.502, -3.208, -2.601, -1.617, -1.491,
                 -1.512, -1.18, -0.695, -1.267, -1.73, -0.961)
      b2min <- c(-1.171, -1.76, -2.03, -1.217, -1.532, -1.699,
                 -2.089, -1.175, -2.139, -1.692, -0.461, -1.286,
                 -0.572, -0.512, -0.409, -0.415, -1.482, -0.176)
      b2max <- c(-0.473, -1.205, -1.811, -0.495, -0.154, -0.912,
                 -1.626, -0.328, -1.876, -1.372, -0.252, -0.17,
                 -0.186, -0.188, 0.155, -0.153, -0.606, 0.449)
      b3min <- c(-0.382, -1.037, -1.106, -0.472, 0.002, -0.911,
                 -1.087, -0.196, -1.232, -0.947, 0.111, -0.028,
                 -0.185, 0.065, 0.19, 0.094, -0.565, 0.593)
      b3max <- c(0.172, -0.372, -0.903, 0.053, 0.886, 0.405,
                 -0.685, 0.115, -0.982, -0.353, 0.425, 0.517,
                 0.713, 0.258, 0.766, 0.564, 0.093, 1.265)
      b4min <- c(1.277, 0.116, 0.415, 0.633, 1.33, 0.782,
                 0.733, 1.183, 0.384, 0.517, 1.432, 1.206,
                 1.133, 1.263, 1.36, 1.292, 1.232, 1.831)
      b4max <- c(1.852, 1.092, 0.689, 1.77, 3.512, 2.217,
                 1.308, 1.296, 0.896, 1.47, 2.019, 2.287,
                 2.16, 1.844, 1.784, 1.834, 1.706, 2.419)
    }
    U2Banks <- vector(mode = "list", length = ntraits)
    set.seed(123)
    for (i in 1:ntraits) {
       U2Banks[[i]] <- cbind(a = runif(nitems, min = dmin[i], max =
dmax[i]),
                             b1 = runif(nitems, min = b1min[i], max =
b1max[i]),
                             b2 = runif(nitems, min = b2min[i], max =
b2max[i]),
                             b3 = runif(nitems, min = b3min[i], max =
b3max[i]),
                             b4 = runif(nitems, min = b4min[i], max =
b4max[i]))
    }
    AllBanks <- list(U2Banks, U2Banks, U2Banks)
    theta_values <- seq(-4, 4, by = 0.05)
    dists <- vector(mode = "list", length = 3)
    for (hh in 1:3) {
       relatives <- matrix(0, nrow = length(theta_values), ncol =
ntraits)
       for (gg in 1:ntraits) {
          storage <- matrix(0, nrow = length(theta_values), ncol =
nitems)
          for (j in 1:nitems) {
             storage[, j] <- sapply(theta_values, item_information,
                                    a = AllBanks[[hh]][[gg]][j, 1],
                                    b = AllBanks[[hh]][[gg]][j, 2:5])
          }
          bankinfo <- apply(storage, 1, sum)
```

```
            relatives[, gg] <- bankinfo / max(bankinfo)
         }
         dists[[hh]] <- relatives
     }
     relreal <- read.csv(realtestinfos, header = FALSE)
     par(mfrow = c(2, 3))
     for (gg in 1:ntraits) {
         plot(theta_values, dists[[1]][, gg], type = "l", col = "blue",
lwd = 2,
              ylab = "Relative Information", xlab = "Theta", ylim = c(0,
1),
              main = "Simulated and Real Information")
              lines(theta_values, relreal[, gg], col = "red", lwd = 2)
              legend("center", legend = c("Robust Uniform", "Real Test
Info."),
                     col = c("blue", "red"), lwd = 2, bty = "n", cex =
0.5)
     }
     traitvals <- matrix(0, nrow = 20, ncol = ntraits)
     traitvals[1, ] <- rep(0, ntraits)
     traitvals[2, ] <- rep(2, ntraits)
     traitvals[3, ] <- c(rep(0, 14), rep(0.5, 4))
     traitvals[4, ] <- c(rep(0.75, 14), rep(1.25, 4))
     traitvals[5, ] <- c(rep(1.50, 14), rep(2, 4))
     traitvals[6, ] <- c(rep(-0.75, 14), rep(0.25, 4))
     traitvals[7, ] <- c(rep(0.75, 14), rep(1.75, 4))
     traitvals[8, ] <- c(rep(1.50, 14), rep(2.50, 4))
     traitvals[9, ] <- c(rep(-0.75, 14), rep(0.75, 4))
     traitvals[10, ] <- c(rep(0.75, 14), rep(2.25, 4))
     traitvals[11, ] <- c(rep(1.50, 14), rep(3, 4))
     traitvals[12, ] <- c(rep(0, 9), rep(0.5, 9))
     traitvals[13, ] <- c(rep(0.75, 9), rep(1.25, 9))
     traitvals[14, ] <- c(rep(1.50, 9), rep(2, 9))
     traitvals[15, ] <- c(rep(-0.75, 9), rep(0.25, 9))
     traitvals[16, ] <- c(rep(0.75, 9), rep(1.75, 9))
     traitvals[17, ] <- c(rep(1.50, 9), rep(2.50, 9))
     traitvals[18, ] <- c(rep(-0.75, 9), rep(0.75, 9))
     traitvals[19, ] <- c(rep(0.75, 9), rep(2.25, 9))
     traitvals[20, ] <- c(rep(1.50, 9), rep(3, 9))
     designs <- rbind(n1, n2, n3, n4, n5, n6, n7, n8, n9, n10, n11, n12,
n13,
                      n14, n15, n16, n17, n18, n19, n20)
     setup <- data.frame(trait1 = traitvals[, 1], trait2 = traitvals[,
2],
                         trait3 = traitvals[, 3], trait4 = traitvals[,
4],
                         trait5 = traitvals[, 5], trait6 = traitvals[,
6],
                         trait7 = traitvals[, 7], trait8 = traitvals[,
8],
                         trait9 = traitvals[, 9], trait10 = traitvals[,
10],
                         trait11 = traitvals[, 11], trait12 =
traitvals[, 12],
```

```
                        trait13 = traitvals[, 13], trait14 =
traitvals[, 14],
                        trait15 = traitvals[, 15], trait16 =
traitvals[, 16],
                        trait17 = traitvals[, 17], trait18 =
traitvals[, 18],
                        mag = designs[, 1], info = designs[, 2],
                        ndiff = designs[, 3])
  }
  if (vlcat) {
    nitemcond <- length(nSEM)
  } else {
    nitemcond <- length(nitemx)
  }
  if (is.null(corrmat)) {
    corrmat <- matrix(0, nrow = ntraits, ncol = ntraits)
    diag(corrmat) <- 0.1
  }
  npenalties <- length(penalties)
  nboth <- nitemcond * npenalties
  outerfinal <- NULL
  basesigns <- rep(1, nsimulees)
  if (vlcat) {outervl1 <- outervl2 <- NULL}
  if (nsimulees == 1) {
    basesigns <- 1
  } else {
    basesigns[1:(floor(nsimulees / 2))] <- -1
  }
  if (numtraits == "mpq") {
    modtype <- "brm"
  } else {
    modtype <- "grm"
  }
  set.seed(123)
  for (gg in temp.start:temp.end) {
     simtrat <- matrix(0, nrow = nsimulees, ncol = ntraits)
     for (i in 1:nsimulees) {
        signswitch <- basesigns[i]
        vals <- traitvals[gg, ]
        simtrat[i, ] <- signswitch * sample(vals, ntraits)
     }
     catStart1 <- list(init.theta = 0, n.start = 5, select = "UW-FI",
                       at = "theta", n.select = 4, it.range = c(-1, 1),
                       score = "step", range = c(-1, 1), step.size = 3,
                       leave.after.MLE = FALSE)
     catMiddle1 <- list(select = "UW-FI", at = "theta", n.select = 1,
                        it.range = NULL, score = "MLE", range = c(-6,
6),
                        expos = "none")
     storout <- storoutZ <- storoutS <- storoutZ2 <- storoutB <-NULL
     for (hh in 1:nitemcond) {
        if (vlcat) {
          ptval <- list(method = "threshold", crit = nSEM[[hh]])
          catTerm1 <- list(term = "precision", n.min = 5, n.max =
(nitems - 1),
```

```
                            p.term = ptval)
        } else {
          catTerm1 <- list(term = "fixed", n.min = 5, n.max =
nitemx[[hh]])
        }
        Combined <- BankCombine <- final <- finalZ <- finalS <- NULL
        finalZ2 <- finalB <- Combined
        if (vlcat) {vlmeans <- vlsds <- NULL}
        catimplied <- matrix(0, nrow = nsimulees, ncol = ntraits)
        pthetaest <- matrix(0, nrow = ntraits, ncol = nsimulees)
        allparams <- Responses <- vector(mode = "list", length =
ntraits)
        catALL <- allresps <- allparams
        for (i in 1:ntraits) {
           resps <- simIrt(theta = simtrat[, i], params =
AllBanks[[3]][[i]],
                            mod = modtype)
           Responses[[i]] <- resps$resp
           Combined <- cbind(Combined, resps$resp)
           BankCombine <- rbind(BankCombine, AllBanks[[3]][[i]])
             suppressWarnings(caty <- catIrt(params =
AllBanks[[3]][[i]],
                                            mod = modtype, resp =
Responses[[i]],
                                            catStart = catStart1,
                                            catMiddle = catMiddle1,
                                            catTerm = catTerm1))
           catALL[[i]] <- caty
           catimplied[, i] <- caty$cat_theta
           bank <- catALL[[i]]
           indiv <- bank$cat_indiv
           if (vlcat) {
             vlmeans <- c(vlmeans, mean(catALL[[i]]$cat_length))
             vlsds <- c(vlsds, sd(catALL[[i]]$cat_length))
           }
           presps <- vector(mode = "list", length = nsimulees)
           pparams <- vector(mode = "list", length = nsimulees)
           for (j in 1:nsimulees) {
              person <- indiv[[j]]
              pthetaest[i, j] <-
person$cat_theta[length(person$cat_theta)]
              presps[[j]] <- as.numeric(person$cat_resp)
              if (modtype == "brm") {
                params <- person$cat_params[, 2:4]
              } else {
                params <- person$cat_params[, 2:6]
              }
              class(params) <- "matrix"
              pparams[[j]] <- params
           }
           allresps[[i]] <- presps
           allparams[[i]] <- pparams
        }
        if (vlcat) {
          outervl1 <- c(outervl1, mean(vlmeans))
```

```
          outervl2 <- c(outervl2, mean(vlsds))
        }
        for (rr in 1:npenalties) {
           LRTout <- Zout <- Sout <- vector(mode = "list", length =
ntraits)
           Z2out <- Bout <- LRTout
           LRTfinals <- Zfinals <- Sfinals <- matrix(0, nrow =
nsimulees)
           Z2finals <- Bfinals <- LRTfinals
           for (j in 1:nsimulees) {
              presps <- vector(mode = "list", length = ntraits)
              pparams <- vector(mode = "list", length = ntraits)
              for (i in 1:ntraits) {
                 presps[[i]] <- allresps[[i]][[j]]
                 pparams[[i]] <- allparams[[i]][[j]]
                 persontheta <- as.matrix(pthetaest[, j])
              }
              LRTout[[j]] <- amc_LRT(resp.list = presps, param.list =
pparams,
                                     mle.theta.vec = persontheta,
                                     model = modtype, alpha = alphaval,
                                     mod = F, threshold = thresval,
                                     prior = priorname,
                                     mean.trait = simtrat[j, ],
                                     corr.trait = corrmat,
                                     mean.pool = mean(simtrat[j, ]),
                                     var.pool = nullvar)
              Zout[[j]] <- amc_Z(resp.list = presps, param.list =
pparams,
                                 model = modtype, mle.theta.vec =
persontheta,
                                 alpha = alphaval)
              Z2out[[j]] <- amc_Z(resp.list = presps, param.list =
pparams,
                                  model = modtype, mle.theta.vec =
persontheta,
                                  alpha = alphaval, mod = T)
              Sout[[j]] <- amc_score(resp.list = presps, param.list =
pparams,
                                     model = modtype,
                                     mle.theta.vec = persontheta,
                                     alpha = alphaval)
              Bout[[j]] <- amc_LRT(resp.list = presps, param.list =
pparams,
                                   mle.theta.vec = persontheta,
                                   model = modtype, alpha = alphaval,
                                   mod = T, threshold = -6,
                                   prior = priorname,
                                   mean.trait = simtrat[j, ],
                                   corr.trait = corrmat,
                                   mean.pool = mean(simtrat[j, ]),
                                   var.pool = nullvar)
              LRTfinals[j, 1] <- LRTout[[j]]$lrt_decision
              Zfinals[j, 1] <- Zout[[j]]$z_decision
              Z2finals[j, 1] <- Z2out[[j]]$z_decision
```

```
                Sfinals[j, 1] <- Sout[[j]]$s_decision
                Bfinals[j, 1] <- Bout[[j]]$lrt_decision
             }
             outcome <- sum(LRTfinals) / nsimulees
             final <- c(final, outcome)
             outcomeZ <- sum(Zfinals) / nsimulees
             finalZ <- c(finalZ, outcomeZ)
             outcomeZ2 <- sum(Z2finals) / nsimulees
             finalZ2 <- c(finalZ2, outcomeZ2)
             outcomeS <- sum(Sfinals) / nsimulees
             finalS <- c(finalS, outcomeS)
             outcomeB <- sum(Bfinals) / nsimulees
             finalB <- c(finalB, outcomeB)
          }
          storout <- c(storout, final)
          storoutZ <- c(storoutZ, finalZ)
          storoutZ2 <- c(storoutZ2, finalZ2)
          storoutS <- c(storoutS, finalS)
          storoutB <- c(storoutB, finalB)
       }
       hold <- NULL; for (i in 1:nboth) {hold <- rbind(hold, setup[gg,
])}
       hold$outLRT <- storout
       hold$outZ <- storoutZ
       hold$outZ2 <- storoutZ2
       hold$outS <- storoutS
       hold$outB <- storoutB
       hold$penalty <- as.character(rep(penalties, nitemcond))
       if (vlcat) {
         hold$nitems <- as.character(rep(nSEM, each = npenalties))
       } else {
         hold$nitems <- as.character(rep(nitemx, each = npenalties))
       }
       outerfinal <- rbind(outerfinal, hold)
       outname <- paste0("simMID-", gg, ".csv")
       write.csv(outerfinal, file = outname, row.names = FALSE)
    }
    if (vlcat) {
      VLout <- data.frame(outmean = outervl1, outsd = outervl2)
      outerfinal <- cbind(outerfinal, VLout)
      colnames(outerfinal)[which(colnames(outerfinal) == "nitems")] <-
"SEMstop"
      write.csv(outerfinal, file = finalname, row.names = FALSE)
    } else {
      write.csv(outerfinal, file = finalname, row.names = FALSE)
    }
    return("Complete")
}

#######################################################################
# Function: test_information_2PL
# Purpose: calculate test information for a two-parameter logistic
#          (2PL) model
# Arguments:
#     1.) theta: trait score for an examinee
```

```
#     2.) a: vector of item discrimination parameters for items
#     3.) b: vector of item difficulty/location parameters for items
######################################################################
test_information_2PL <- function(theta, a, b) {
  if (ncol(b) != 1) {
    stop("Input 'b' must contain exactly one column.")
  }
  if (ncol(a) != 1) {
    stop("Input 'a' must contain exactly one column.")
  }
  if (nrow(a) != nrow(b)) {
    stop("Inputs 'a' and 'b' must contain the same number of rows.")
  }
  infos <- 0
  for (i in 1:nrow(a)) {
     infos <- infos + item_information_2PL(theta = theta, a = a[i, 1],
                                          b = b[i, 1])
  }
  return(infos)
}

######################################################################
# Function: test_information
# Purpose: calculate test information for the graded response model
#          (GRM)
# Arguments:
#     1.) theta: trait score for an examinee
#     2.) a: vector of item discrimination parameters for items
#     3.) b: matrix of boundary parameters for items; only works for
#            items with five responses
######################################################################
test_information <- function(theta, a, b) {
  if (ncol(b) != 4) {
    stop("Input 'b' must contain exactly four columns.")
  }
  if (ncol(a) != 1) {
    stop("Input 'a' must contain exactly one column.")
  }
  if (nrow(a) != nrow(b)) {
    stop("Inputs 'a' and 'b' must contain the same number of rows.")
  }
  infos <- 0
  for (i in 1:nrow(a)) {
     infos <- infos + item_information(theta = theta, a = a[i, 1],
                                      b = as.matrix(b[i, 1:4]))
  }
  return(infos)
}


######################################################################
# Simulation for Adaptive Profile Difference Analysis
#
# Matthew Snodgress
# Department of Psychology - University of Minnesota, Twin Cities
```

```
#########################################################################

####### Configuration----

# specify working directory
workingd <- "~/Documents/Simulation/"
setwd(workingd)

# specify directory that contains observed relative test information
functions
rtifdir <- paste0(workingd, "TIF/")

# specify whether to use sequential or parallel computing (FALSE for
sequential)
paracomp <- FALSE

# specify number of simulees
nsims <- 1e3 + 1

####### Implementation----

# install and/or load packages
if (!require(catIrt)) {install.packages("catIrt");library(catIrt)}
if (!require(mvtnorm)) {install.packages("mvtnorm");library(mvtnorm)}
if (!require(foreach)) {install.packages("foreach");library(foreach)}
if (!require(doParallel))
{install.packages("doParallel");library(doParallel)}

# define functions
fnames <- c("Pstar", "Pstarp", "item_information",
"item_information_2PL",
            "test_information", "test_information_2PL", "lrtcustom",
"score",
            "score_2PL", "amc_LRT", "amc_Z", "amc_score", "simAPDA")
nfn <- length(fnames)
for(hh in 1:nfn) {source(paste0(workingd, "Functions/", fnames[hh],
".R"))}

# specify over CAT Type, Bank Type, Discrimination, and Number of
Scales
constor <- list()
constor[[1]] <- list("bigfive", F, F, F)  # 5 traits - FL - real -
peaked
constor[[2]] <- list("bigfive", T, F, F)  # 5 traits - VL - real -
peaked
constor[[3]] <- list("bigfive", F, T, F)  # 5 traits - FL - ideal -
peaked
constor[[4]] <- list("bigfive", T, T, F)  # 5 traits - VL - ideal -
peaked
constor[[5]] <- list("mpq", F, F, F)      # 11 traits - FL - real -
peaked
constor[[6]] <- list("mpq", T, F, F)      # 11 traits - VL - real -
peaked
constor[[7]] <- list("mpq", F, T, F)      # 11 traits - FL - ideal -
peaked
```

```
constor[[8]] <- list("mpq", T, T, F)      # 11 traits - VL - ideal -
peaked
constor[[9]] <- list("facets", F, F, F)   # 18 traits - FL - real -
peaked
constor[[10]] <- list("facets", T, F, F)  # 18 traits - VL - real -
peaked
constor[[11]] <- list("facets", F, T, F)  # 18 traits - FL - ideal -
peaked
constor[[12]] <- list("facets", T, T, F)  # 18 traits - VL - ideal -
peaked
constor[[13]] <- list("bigfive", F, F, T) # 5 traits - FL - real -
equal
constor[[14]] <- list("bigfive", T, F, T) # 5 traits - VL - real -
equal
constor[[15]] <- list("bigfive", F, T, T) # 5 traits - FL - ideal -
equal
constor[[16]] <- list("bigfive", T, T, T) # 5 traits - VL - ideal -
equal
constor[[17]] <- list("mpq", F, F, T)     # 11 traits - FL - real -
equal
constor[[18]] <- list("mpq", T, F, T)     # 11 traits - VL - real -
equal
constor[[19]] <- list("mpq", F, T, T)     # 11 traits - FL - ideal -
equal
constor[[20]] <- list("mpq", T, T, T)     # 11 traits - VL - ideal -
equal
constor[[21]] <- list("facets", F, F, T)  # 18 traits - FL - real -
equal
constor[[22]] <- list("facets", T, F, T)  # 18 traits - VL - real -
equal
constor[[23]] <- list("facets", F, T, T)  # 18 traits - FL - ideal -
equal
constor[[24]] <- list("facets", T, T, T)  # 18 traits - VL - ideal -
equal

# specify names for output files
outputnam <- c("real_FL_5_peak.csv", "real_VL_5_peak.csv",
               "ideal_FL_5_peak.csv", "ideal_VL_5_peak.csv",
               "real_FL_11_peak.csv", "real_VL_11_peak.csv",
               "ideal_FL_11_peak.csv", "ideal_VL_11_peak.csv",
               "real_FL_18_peak.csv", "real_VL_18_peak.csv",
               "ideal_FL_18_peak.csv", "ideal_VL_18_peak.csv",
               "real_FL_5_equi.csv", "real_VL_5_equi.csv",
               "ideal_FL_5_equi.csv", "ideal_VL_5_equi.csv",
               "real_FL_11_equi.csv", "real_VL_11_equi.csv",
               "ideal_FL_11_equi.csv", "ideal_VL_11_equi.csv",
               "real_FL_18_equi.csv", "real_VL_18_equi.csv",
               "ideal_FL_18_equi.csv", "ideal_VL_18_equi.csv")
outname <- vector(mode = "list", length = length(constor))
for (tt in 1:length(outputnam)) {outname[[tt]] <- outputnam[tt]}

# loop through conditions
if (paracomp) {
  set.seed(123)
  tic <- Sys.time()
```

```
  numCores <- detectCores()
  cl <- makeCluster(numCores)
  registerDoParallel(cl)
  foreach (i = 1:length(constor), .packages = c('catIrt', 'mvtnorm'))
%dopar% {
         simAPDA(nsimulees = nsims, ncond = 20, nitemx = c(10, 30, 50),
                 alphaval = 0.05, nitems = 100, nSEM = c(0.4, 0.3,
0.2),
                 penalties = c(0), temp.start = 1, temp.end = 20,
                 numtraits = constor[[i]][[1]], vlcat =
constor[[i]][[2]],
                 betterbank = constor[[i]][[3]], equalbank =
constor[[i]][[4]],
                 finalname = outname[[i]], realtestinfos0 = rtifdir)
  }
  stopCluster(cl)
  toc <- Sys.time()
  timer <- toc - tic
} else {
  set.seed(123)
  tic <- Sys.time()
  for (i in 1:length(constor)) {
     simAPDA(nsimulees = nsims, ncond = 20, nitemx = c(10, 30, 50),
             alphaval = 0.05, nitems = 100, nSEM = c(0.4, 0.3, 0.2),
             penalties = c(0), temp.start = 1, temp.end = 20,
             numtraits = constor[[i]][[1]], vlcat = constor[[i]][[2]],
             betterbank = constor[[i]][[3]], equalbank =
constor[[i]][[4]],
             finalname = outname[[i]], realtestinfos0 = rtifdir)
  }
  toc <- Sys.time()
  timer <- toc - tic
}
```

Appendix C. Relative Test Information for Real Scales at Specific Scores

Five Scales – Goldberg's Big Five Measure

Table C1. Observed Relative Test Information for Goldberg's Big Five Measure at Specific Scores.

θ	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
-4	0.0343	0.0615	0.0323	0.0153	0.0487
-3.95	0.0363	0.0657	0.0352	0.0162	0.0519
-3.9	0.0385	0.0701	0.0382	0.0171	0.0552
-3.85	0.0407	0.0749	0.0416	0.0181	0.0588
-3.8	0.0431	0.0800	0.0452	0.0192	0.0627
-3.75	0.0456	0.0854	0.0491	0.0204	0.0668
-3.7	0.0483	0.0912	0.0534	0.0216	0.0711
-3.65	0.0511	0.0974	0.0580	0.0229	0.0758
-3.6	0.0541	0.1040	0.0631	0.0244	0.0807
-3.55	0.0572	0.1110	0.0685	0.0258	0.0860
-3.5	0.0606	0.1185	0.0743	0.0274	0.0916
-3.45	0.0641	0.1265	0.0807	0.0291	0.0977
-3.4	0.0678	0.1350	0.0875	0.0310	0.1041
-3.35	0.0717	0.1440	0.0949	0.0329	0.1109
-3.3	0.0758	0.1535	0.1029	0.0350	0.1183
-3.25	0.0801	0.1637	0.1115	0.0372	0.1261
-3.2	0.0847	0.1744	0.1207	0.0396	0.1345
-3.15	0.0895	0.1857	0.1306	0.0421	0.1434
-3.1	0.0945	0.1976	0.1413	0.0448	0.1530
-3.05	0.0999	0.2103	0.1527	0.0477	0.1632
-3	0.1055	0.2235	0.1649	0.0508	0.1741
-2.95	0.1113	0.2375	0.1780	0.0541	0.1858
-2.9	0.1175	0.2521	0.1919	0.0577	0.1983
-2.85	0.1240	0.2674	0.2067	0.0615	0.2117
-2.8	0.1308	0.2834	0.2225	0.0655	0.2260
-2.75	0.1380	0.3001	0.2392	0.0699	0.2412
-2.7	0.1455	0.3174	0.2569	0.0746	0.2574
-2.65	0.1534	0.3354	0.2756	0.0796	0.2748
-2.6	0.1616	0.3540	0.2953	0.0849	0.2932
-2.55	0.1703	0.3732	0.3159	0.0906	0.3127
-2.5	0.1793	0.3930	0.3375	0.0967	0.3334
-2.45	0.1888	0.4133	0.3601	0.1033	0.3553
-2.4	0.1987	0.4340	0.3835	0.1103	0.3783

-2.35	0.2090	0.4552	0.4078	0.1178	0.4024
-2.3	0.2198	0.4768	0.4328	0.1259	0.4276
-2.25	0.2311	0.4986	0.4585	0.1345	0.4538
-2.2	0.2428	0.5207	0.4848	0.1436	0.4809
-2.15	0.2550	0.5430	0.5116	0.1534	0.5088
-2.1	0.2678	0.5653	0.5387	0.1639	0.5372
-2.05	0.2811	0.5877	0.5660	0.1751	0.5660
-2	0.2949	0.6100	0.5933	0.1870	0.5949
-1.95	0.3093	0.6321	0.6205	0.1997	0.6238
-1.9	0.3241	0.6540	0.6475	0.2131	0.6522
-1.85	0.3396	0.6756	0.6739	0.2275	0.6799
-1.8	0.3556	0.6968	0.6998	0.2426	0.7066
-1.75	0.3722	0.7175	0.7248	0.2587	0.7321
-1.7	0.3893	0.7377	0.7490	0.2758	0.7562
-1.65	0.4070	0.7572	0.7720	0.2937	0.7786
-1.6	0.4252	0.7761	0.7939	0.3126	0.7992
-1.55	0.4440	0.7941	0.8145	0.3325	0.8179
-1.5	0.4633	0.8113	0.8338	0.3534	0.8347
-1.45	0.4831	0.8276	0.8517	0.3753	0.8497
-1.4	0.5034	0.8429	0.8682	0.3980	0.8630
-1.35	0.5240	0.8572	0.8833	0.4217	0.8747
-1.3	0.5451	0.8704	0.8969	0.4463	0.8851
-1.25	0.5666	0.8827	0.9092	0.4716	0.8943
-1.2	0.5884	0.8938	0.9202	0.4978	0.9026
-1.15	0.6103	0.9039	0.9299	0.5245	0.9102
-1.1	0.6325	0.9131	0.9385	0.5518	0.9173
-1.05	0.6547	0.9212	0.9459	0.5795	0.9240
-1	0.6769	0.9284	0.9524	0.6074	0.9305
-0.95	0.6988	0.9348	0.9579	0.6354	0.9369
-0.9	0.7206	0.9405	0.9627	0.6633	0.9432
-0.85	0.7419	0.9455	0.9667	0.6909	0.9494
-0.8	0.7628	0.9500	0.9702	0.7178	0.9554
-0.75	0.7830	0.9540	0.9732	0.7438	0.9612
-0.7	0.8023	0.9576	0.9759	0.7686	0.9668
-0.65	0.8207	0.9609	0.9782	0.7919	0.9721
-0.6	0.8380	0.9641	0.9804	0.8134	0.9770
-0.55	0.8542	0.9671	0.9824	0.8326	0.9815
-0.5	0.8691	0.9700	0.9843	0.8494	0.9855
-0.45	0.8827	0.9729	0.9862	0.8635	0.9891

-0.4	0.8949	0.9757	0.9881	0.8747	0.9922
-0.35	0.9057	0.9785	0.9899	0.8831	0.9948
-0.3	0.9151	0.9812	0.9918	0.8886	0.9969
-0.25	0.9232	0.9839	0.9935	0.8916	0.9984
-0.2	0.9302	0.9865	0.9951	0.8923	0.9995
-0.15	0.9361	0.9888	0.9966	0.8913	1.0000
-0.1	0.9410	0.9911	0.9979	0.8891	1.0000
-0.05	0.9452	0.9931	0.9989	0.8863	0.9995
0	0.9487	0.9949	0.9996	0.8834	0.9986
0.05	0.9518	0.9965	1.0000	0.8812	0.9972
0.1	0.9547	0.9977	1.0000	0.8801	0.9954
0.15	0.9573	0.9987	0.9996	0.8804	0.9933
0.2	0.9598	0.9994	0.9989	0.8825	0.9910
0.25	0.9624	0.9999	0.9977	0.8864	0.9884
0.3	0.9652	1.0000	0.9962	0.8923	0.9857
0.35	0.9679	0.9999	0.9943	0.8998	0.9830
0.4	0.9708	0.9996	0.9920	0.9088	0.9802
0.45	0.9737	0.9990	0.9893	0.9188	0.9775
0.5	0.9766	0.9982	0.9861	0.9295	0.9747
0.55	0.9794	0.9972	0.9824	0.9404	0.9719
0.6	0.9820	0.9960	0.9781	0.9510	0.9691
0.65	0.9844	0.9946	0.9732	0.9609	0.9661
0.7	0.9865	0.9930	0.9676	0.9698	0.9629
0.75	0.9882	0.9911	0.9612	0.9774	0.9593
0.8	0.9895	0.9890	0.9539	0.9838	0.9552
0.85	0.9903	0.9867	0.9458	0.9888	0.9505
0.9	0.9908	0.9840	0.9367	0.9927	0.9450
0.95	0.9909	0.9809	0.9265	0.9955	0.9386
1	0.9908	0.9774	0.9153	0.9975	0.9313
1.05	0.9905	0.9735	0.9031	0.9989	0.9229
1.1	0.9901	0.9691	0.8897	0.9997	0.9133
1.15	0.9896	0.9642	0.8752	1.0000	0.9025
1.2	0.9893	0.9586	0.8595	0.9998	0.8904
1.25	0.9891	0.9525	0.8427	0.9991	0.8771
1.3	0.9892	0.9457	0.8247	0.9978	0.8625
1.35	0.9895	0.9382	0.8056	0.9958	0.8466
1.4	0.9901	0.9301	0.7854	0.9930	0.8295
1.45	0.9910	0.9213	0.7640	0.9896	0.8113
1.5	0.9921	0.9118	0.7417	0.9857	0.7919

1.55	0.9933	0.9016	0.7184	0.9814	0.7715
1.6	0.9947	0.8908	0.6942	0.9770	0.7501
1.65	0.9961	0.8793	0.6692	0.9727	0.7280
1.7	0.9974	0.8673	0.6436	0.9688	0.7051
1.75	0.9984	0.8547	0.6176	0.9653	0.6816
1.8	0.9993	0.8415	0.5912	0.9622	0.6577
1.85	0.9999	0.8278	0.5646	0.9596	0.6335
1.9	1.0000	0.8137	0.5380	0.9570	0.6091
1.95	0.9998	0.7991	0.5116	0.9542	0.5847
2	0.9992	0.7842	0.4854	0.9507	0.5604
2.05	0.9983	0.7688	0.4598	0.9460	0.5364
2.1	0.9971	0.7531	0.4346	0.9396	0.5127
2.15	0.9956	0.7371	0.4102	0.9311	0.4895
2.2	0.9938	0.7209	0.3865	0.9201	0.4669
2.25	0.9919	0.7044	0.3637	0.9064	0.4449
2.3	0.9899	0.6877	0.3418	0.8899	0.4235
2.35	0.9877	0.6708	0.3207	0.8707	0.4030
2.4	0.9854	0.6538	0.3007	0.8489	0.3831
2.45	0.9829	0.6367	0.2816	0.8250	0.3641
2.5	0.9803	0.6196	0.2635	0.7992	0.3459
2.55	0.9773	0.6025	0.2463	0.7719	0.3284
2.6	0.9739	0.5855	0.2300	0.7437	0.3118
2.65	0.9700	0.5685	0.2147	0.7148	0.2959
2.7	0.9654	0.5516	0.2003	0.6856	0.2807
2.75	0.9600	0.5350	0.1867	0.6564	0.2662
2.8	0.9537	0.5185	0.1739	0.6276	0.2525
2.85	0.9463	0.5023	0.1618	0.5992	0.2393
2.9	0.9378	0.4864	0.1506	0.5715	0.2268
2.95	0.9280	0.4707	0.1400	0.5445	0.2149
3	0.9170	0.4555	0.1301	0.5184	0.2035
3.05	0.9046	0.4405	0.1209	0.4932	0.1927
3.1	0.8910	0.4259	0.1122	0.4690	0.1823
3.15	0.8760	0.4116	0.1042	0.4456	0.1725
3.2	0.8599	0.3978	0.0966	0.4233	0.1631
3.25	0.8426	0.3843	0.0896	0.4018	0.1541
3.3	0.8243	0.3712	0.0830	0.3813	0.1456
3.35	0.8051	0.3585	0.0769	0.3616	0.1375
3.4	0.7851	0.3461	0.0712	0.3429	0.1297
3.45	0.7645	0.3342	0.0659	0.3250	0.1224

3.5	0.7434	0.3226	0.0610	0.3079	0.1154
3.55	0.7218	0.3114	0.0564	0.2916	0.1087
3.6	0.7000	0.3006	0.0522	0.2761	0.1024
3.65	0.6780	0.2901	0.0482	0.2614	0.0964
3.7	0.6559	0.2800	0.0446	0.2474	0.0907
3.75	0.6339	0.2702	0.0412	0.2340	0.0853
3.8	0.6120	0.2607	0.0380	0.2214	0.0802
3.85	0.5902	0.2516	0.0351	0.2094	0.0754
3.9	0.5687	0.2428	0.0324	0.1981	0.0708
3.95	0.5476	0.2343	0.0299	0.1873	0.0665
4	0.5268	0.2261	0.0276	0.1772	0.0625

11 Scales – Multidimensional Personality Questionnaire

Table C2. Observed Relative Test Information for the Multidimensional Personality Questionnaire at Specific Scores – Scales 1 through 5.

θ	Absorption	Achievement	Aggression	Alienation	Control
-4	0.0331	0.0270	0.1448	0.1756	0.0192
-3.95	0.0354	0.0284	0.1533	0.1862	0.0202
-3.9	0.0377	0.0298	0.1623	0.1974	0.0212
-3.85	0.0403	0.0313	0.1718	0.2092	0.0223
-3.8	0.0430	0.0329	0.1818	0.2216	0.0234
-3.75	0.0458	0.0346	0.1924	0.2346	0.0246
-3.7	0.0489	0.0364	0.2035	0.2483	0.0258
-3.65	0.0522	0.0382	0.2152	0.2627	0.0271
-3.6	0.0556	0.0402	0.2274	0.2778	0.0285
-3.55	0.0593	0.0423	0.2403	0.2936	0.0299
-3.5	0.0633	0.0445	0.2538	0.3101	0.0315
-3.45	0.0674	0.0468	0.2679	0.3273	0.0330
-3.4	0.0719	0.0493	0.2827	0.3453	0.0347
-3.35	0.0766	0.0519	0.2981	0.3641	0.0365
-3.3	0.0817	0.0546	0.3141	0.3836	0.0383
-3.25	0.0870	0.0575	0.3308	0.4038	0.0403
-3.2	0.0927	0.0606	0.3481	0.4248	0.0423
-3.15	0.0987	0.0639	0.3660	0.4465	0.0445
-3.1	0.1050	0.0673	0.3845	0.4689	0.0468
-3.05	0.1118	0.0709	0.4036	0.4919	0.0492
-3	0.1189	0.0747	0.4232	0.5155	0.0517
-2.95	0.1265	0.0787	0.4432	0.5397	0.0543
-2.9	0.1344	0.0830	0.4637	0.5644	0.0571

-2.85	0.1428	0.0875	0.4846	0.5895	0.0601
-2.8	0.1517	0.0922	0.5058	0.6149	0.0631
-2.75	0.1610	0.0972	0.5272	0.6406	0.0664
-2.7	0.1708	0.1025	0.5488	0.6664	0.0698
-2.65	0.1811	0.1081	0.5704	0.6923	0.0734
-2.6	0.1919	0.1139	0.5920	0.7181	0.0772
-2.55	0.2033	0.1201	0.6135	0.7436	0.0811
-2.5	0.2151	0.1266	0.6348	0.7688	0.0853
-2.45	0.2275	0.1334	0.6557	0.7934	0.0897
-2.4	0.2403	0.1406	0.6763	0.8174	0.0943
-2.35	0.2538	0.1482	0.6963	0.8406	0.0992
-2.3	0.2677	0.1562	0.7158	0.8629	0.1043
-2.25	0.2821	0.1645	0.7346	0.8839	0.1097
-2.2	0.2971	0.1733	0.7527	0.9037	0.1153
-2.15	0.3125	0.1825	0.7701	0.9221	0.1212
-2.1	0.3283	0.1922	0.7867	0.9388	0.1275
-2.05	0.3447	0.2023	0.8025	0.9538	0.1340
-2	0.3614	0.2130	0.8175	0.9669	0.1409
-1.95	0.3785	0.2241	0.8318	0.9779	0.1481
-1.9	0.3959	0.2357	0.8453	0.9869	0.1557
-1.85	0.4137	0.2478	0.8581	0.9936	0.1636
-1.8	0.4317	0.2605	0.8703	0.9980	0.1720
-1.75	0.4500	0.2737	0.8819	1.0000	0.1807
-1.7	0.4684	0.2875	0.8930	0.9996	0.1898
-1.65	0.4870	0.3018	0.9037	0.9968	0.1994
-1.6	0.5056	0.3167	0.9140	0.9915	0.2094
-1.55	0.5244	0.3322	0.9239	0.9839	0.2199
-1.5	0.5431	0.3483	0.9336	0.9739	0.2309
-1.45	0.5618	0.3650	0.9430	0.9617	0.2423
-1.4	0.5805	0.3823	0.9521	0.9475	0.2543
-1.35	0.5992	0.4002	0.9608	0.9312	0.2667
-1.3	0.6177	0.4187	0.9691	0.9132	0.2797
-1.25	0.6362	0.4377	0.9768	0.8934	0.2932
-1.2	0.6546	0.4573	0.9838	0.8723	0.3072
-1.15	0.6729	0.4775	0.9899	0.8498	0.3217
-1.1	0.6911	0.4982	0.9948	0.8262	0.3368
-1.05	0.7094	0.5194	0.9983	0.8017	0.3523
-1	0.7276	0.5410	1.0000	0.7765	0.3684
-0.95	0.7458	0.5631	0.9997	0.7507	0.3849

-0.9	0.7640	0.5855	0.9972	0.7245	0.4019
-0.85	0.7823	0.6083	0.9921	0.6980	0.4193
-0.8	0.8007	0.6313	0.9844	0.6714	0.4372
-0.75	0.8191	0.6545	0.9738	0.6448	0.4554
-0.7	0.8376	0.6777	0.9602	0.6183	0.4739
-0.65	0.8561	0.7010	0.9439	0.5921	0.4928
-0.6	0.8744	0.7242	0.9247	0.5662	0.5118
-0.55	0.8926	0.7471	0.9029	0.5408	0.5311
-0.5	0.9103	0.7698	0.8788	0.5159	0.5504
-0.45	0.9274	0.7920	0.8526	0.4915	0.5699
-0.4	0.9436	0.8136	0.8247	0.4678	0.5894
-0.35	0.9586	0.8346	0.7954	0.4447	0.6088
-0.3	0.9719	0.8548	0.7652	0.4224	0.6282
-0.25	0.9831	0.8740	0.7343	0.4008	0.6474
-0.2	0.9918	0.8923	0.7032	0.3800	0.6665
-0.15	0.9976	0.9093	0.6721	0.3599	0.6855
-0.1	1.0000	0.9251	0.6412	0.3407	0.7042
-0.05	0.9986	0.9396	0.6109	0.3222	0.7227
0	0.9932	0.9526	0.5814	0.3046	0.7410
0.05	0.9836	0.9642	0.5526	0.2877	0.7591
0.1	0.9697	0.9742	0.5248	0.2717	0.7769
0.15	0.9515	0.9826	0.4981	0.2563	0.7945
0.2	0.9294	0.9893	0.4725	0.2418	0.8118
0.25	0.9035	0.9945	0.4480	0.2279	0.8289
0.3	0.8744	0.9979	0.4247	0.2148	0.8457
0.35	0.8425	0.9998	0.4025	0.2023	0.8621
0.4	0.8084	1.0000	0.3814	0.1905	0.8782
0.45	0.7726	0.9986	0.3613	0.1793	0.8939
0.5	0.7358	0.9957	0.3423	0.1687	0.9090
0.55	0.6984	0.9912	0.3243	0.1588	0.9235
0.6	0.6609	0.9852	0.3073	0.1493	0.9372
0.65	0.6238	0.9778	0.2912	0.1404	0.9499
0.7	0.5874	0.9690	0.2759	0.1320	0.9617
0.75	0.5520	0.9589	0.2614	0.1241	0.9721
0.8	0.5179	0.9475	0.2477	0.1166	0.9812
0.85	0.4851	0.9349	0.2347	0.1096	0.9887
0.9	0.4539	0.9211	0.2224	0.1030	0.9944
0.95	0.4242	0.9063	0.2107	0.0968	0.9982
1	0.3961	0.8904	0.1997	0.0909	1.0000

1.05	0.3696	0.8736	0.1892	0.0854	0.9996
1.1	0.3447	0.8558	0.1793	0.0802	0.9971
1.15	0.3214	0.8373	0.1699	0.0754	0.9922
1.2	0.2995	0.8180	0.1609	0.0708	0.9851
1.25	0.2790	0.7981	0.1525	0.0665	0.9758
1.3	0.2599	0.7775	0.1444	0.0625	0.9642
1.35	0.2420	0.7565	0.1368	0.0587	0.9506
1.4	0.2254	0.7350	0.1295	0.0551	0.9351
1.45	0.2098	0.7131	0.1227	0.0518	0.9177
1.5	0.1954	0.6910	0.1161	0.0486	0.8988
1.55	0.1819	0.6687	0.1099	0.0457	0.8784
1.6	0.1694	0.6462	0.1041	0.0429	0.8567
1.65	0.1577	0.6237	0.0985	0.0403	0.8340
1.7	0.1468	0.6013	0.0932	0.0379	0.8105
1.75	0.1367	0.5789	0.0882	0.0356	0.7863
1.8	0.1273	0.5568	0.0834	0.0334	0.7616
1.85	0.1186	0.5349	0.0789	0.0314	0.7366
1.9	0.1104	0.5132	0.0746	0.0295	0.7114
1.95	0.1028	0.4919	0.0706	0.0277	0.6862
2	0.0958	0.4711	0.0667	0.0261	0.6611
2.05	0.0892	0.4506	0.0631	0.0245	0.6362
2.1	0.0831	0.4307	0.0597	0.0230	0.6116
2.15	0.0774	0.4112	0.0564	0.0217	0.5873
2.2	0.0721	0.3924	0.0533	0.0204	0.5635
2.25	0.0672	0.3740	0.0504	0.0192	0.5401
2.3	0.0626	0.3563	0.0476	0.0180	0.5173
2.35	0.0583	0.3392	0.0450	0.0169	0.4950
2.4	0.0543	0.3227	0.0425	0.0159	0.4733
2.45	0.0506	0.3068	0.0402	0.0150	0.4523
2.5	0.0472	0.2915	0.0380	0.0141	0.4318
2.55	0.0440	0.2768	0.0359	0.0133	0.4120
2.6	0.0410	0.2627	0.0339	0.0125	0.3929
2.65	0.0382	0.2493	0.0320	0.0118	0.3744
2.7	0.0356	0.2364	0.0303	0.0111	0.3566
2.75	0.0332	0.2241	0.0286	0.0104	0.3395
2.8	0.0310	0.2124	0.0270	0.0098	0.3230
2.85	0.0289	0.2012	0.0255	0.0092	0.3071
2.9	0.0269	0.1906	0.0241	0.0087	0.2919
2.95	0.0251	0.1805	0.0228	0.0082	0.2773

3	0.0234	0.1708	0.0216	0.0077	0.2634
3.05	0.0219	0.1617	0.0204	0.0073	0.2500
3.1	0.0204	0.1530	0.0192	0.0068	0.2373
3.15	0.0190	0.1448	0.0182	0.0065	0.2251
3.2	0.0178	0.1370	0.0172	0.0061	0.2135
3.25	0.0166	0.1296	0.0162	0.0057	0.2024
3.3	0.0155	0.1226	0.0154	0.0054	0.1919
3.35	0.0144	0.1160	0.0145	0.0051	0.1819
3.4	0.0135	0.1097	0.0137	0.0048	0.1723
3.45	0.0126	0.1038	0.0130	0.0045	0.1633
3.5	0.0117	0.0982	0.0123	0.0043	0.1546
3.55	0.0110	0.0929	0.0116	0.0040	0.1465
3.6	0.0102	0.0879	0.0110	0.0038	0.1387
3.65	0.0096	0.0831	0.0104	0.0036	0.1314
3.7	0.0089	0.0787	0.0098	0.0034	0.1244
3.75	0.0084	0.0744	0.0093	0.0032	0.1178
3.8	0.0078	0.0704	0.0088	0.0030	0.1115
3.85	0.0073	0.0667	0.0083	0.0029	0.1056
3.9	0.0068	0.0631	0.0079	0.0027	0.0999
3.95	0.0064	0.0597	0.0074	0.0025	0.0946
4	0.0059	0.0566	0.0070	0.0024	0.0896

Table C3. Observed Relative Test Information for the Multidimensional Personality Questionnaire at Specific Scores – Scales 6 through 10.

θ	Harm Avoidance	Social Closeness	Social Potency	Stress Reaction	Traditionalism
-4	0.0098	0.0092	0.0331	0.0178	0.0230
-3.95	0.0103	0.0098	0.0354	0.0190	0.0240
-3.9	0.0108	0.0105	0.0377	0.0204	0.0252
-3.85	0.0115	0.0111	0.0403	0.0218	0.0263
-3.8	0.0121	0.0118	0.0430	0.0234	0.0276
-3.75	0.0127	0.0126	0.0458	0.0250	0.0288
-3.7	0.0134	0.0134	0.0489	0.0268	0.0302
-3.65	0.0142	0.0143	0.0522	0.0287	0.0316
-3.6	0.0150	0.0152	0.0556	0.0308	0.0331
-3.55	0.0158	0.0161	0.0593	0.0330	0.0346
-3.5	0.0167	0.0172	0.0633	0.0354	0.0362
-3.45	0.0176	0.0183	0.0674	0.0379	0.0379
-3.4	0.0186	0.0194	0.0719	0.0406	0.0397
-3.35	0.0196	0.0207	0.0766	0.0436	0.0415

-3.3	0.0207	0.0220	0.0817	0.0467	0.0435
-3.25	0.0219	0.0234	0.0870	0.0501	0.0455
-3.2	0.0231	0.0249	0.0927	0.0538	0.0476
-3.15	0.0244	0.0265	0.0987	0.0577	0.0498
-3.1	0.0258	0.0282	0.1050	0.0619	0.0521
-3.05	0.0273	0.0300	0.1118	0.0664	0.0546
-3	0.0288	0.0319	0.1189	0.0712	0.0571
-2.95	0.0305	0.0339	0.1265	0.0764	0.0598
-2.9	0.0322	0.0360	0.1344	0.0820	0.0625
-2.85	0.0341	0.0383	0.1428	0.0880	0.0654
-2.8	0.0360	0.0407	0.1517	0.0944	0.0685
-2.75	0.0381	0.0433	0.1610	0.1013	0.0717
-2.7	0.0403	0.0461	0.1708	0.1087	0.0750
-2.65	0.0426	0.0490	0.1811	0.1167	0.0785
-2.6	0.0451	0.0520	0.1919	0.1252	0.0821
-2.55	0.0477	0.0553	0.2033	0.1344	0.0859
-2.5	0.0504	0.0588	0.2151	0.1442	0.0899
-2.45	0.0534	0.0625	0.2275	0.1547	0.0940
-2.4	0.0565	0.0664	0.2403	0.1659	0.0984
-2.35	0.0598	0.0705	0.2538	0.1780	0.1029
-2.3	0.0632	0.0749	0.2677	0.1909	0.1077
-2.25	0.0669	0.0795	0.2821	0.2046	0.1126
-2.2	0.0709	0.0844	0.2971	0.2193	0.1178
-2.15	0.0750	0.0896	0.3125	0.2350	0.1233
-2.1	0.0794	0.0951	0.3283	0.2517	0.1289
-2.05	0.0841	0.1010	0.3447	0.2694	0.1349
-2	0.0891	0.1071	0.3614	0.2883	0.1411
-1.95	0.0943	0.1136	0.3785	0.3083	0.1476
-1.9	0.0999	0.1205	0.3959	0.3295	0.1544
-1.85	0.1058	0.1278	0.4137	0.3519	0.1614
-1.8	0.1120	0.1354	0.4317	0.3755	0.1688
-1.75	0.1187	0.1435	0.4500	0.4003	0.1766
-1.7	0.1257	0.1520	0.4684	0.4263	0.1847
-1.65	0.1332	0.1611	0.4870	0.4535	0.1931
-1.6	0.1411	0.1705	0.5056	0.4819	0.2019
-1.55	0.1494	0.1805	0.5244	0.5112	0.2111
-1.5	0.1583	0.1911	0.5431	0.5415	0.2207
-1.45	0.1677	0.2022	0.5618	0.5726	0.2307
-1.4	0.1776	0.2138	0.5805	0.6044	0.2412

-1.35	0.1881	0.2261	0.5992	0.6366	0.2521
-1.3	0.1992	0.2390	0.6177	0.6691	0.2634
-1.25	0.2109	0.2526	0.6362	0.7015	0.2752
-1.2	0.2233	0.2668	0.6546	0.7337	0.2875
-1.15	0.2364	0.2817	0.6729	0.7652	0.3003
-1.1	0.2502	0.2974	0.6911	0.7959	0.3136
-1.05	0.2648	0.3138	0.7094	0.8253	0.3274
-1	0.2801	0.3310	0.7276	0.8532	0.3417
-0.95	0.2963	0.3490	0.7458	0.8793	0.3565
-0.9	0.3132	0.3678	0.7640	0.9033	0.3719
-0.85	0.3310	0.3875	0.7823	0.9249	0.3877
-0.8	0.3497	0.4079	0.8007	0.9440	0.4041
-0.75	0.3692	0.4293	0.8191	0.9604	0.4210
-0.7	0.3896	0.4515	0.8376	0.9740	0.4384
-0.65	0.4108	0.4745	0.8561	0.9848	0.4563
-0.6	0.4329	0.4984	0.8744	0.9926	0.4747
-0.55	0.4557	0.5231	0.8926	0.9977	0.4935
-0.5	0.4794	0.5485	0.9103	1.0000	0.5127
-0.45	0.5037	0.5747	0.9274	0.9996	0.5324
-0.4	0.5287	0.6015	0.9436	0.9967	0.5524
-0.35	0.5543	0.6288	0.9586	0.9913	0.5727
-0.3	0.5803	0.6566	0.9719	0.9837	0.5933
-0.25	0.6067	0.6847	0.9831	0.9739	0.6142
-0.2	0.6333	0.7130	0.9918	0.9622	0.6352
-0.15	0.6599	0.7412	0.9976	0.9486	0.6564
-0.1	0.6865	0.7691	1.0000	0.9334	0.6777
-0.05	0.7128	0.7966	0.9986	0.9167	0.6990
0	0.7388	0.8234	0.9932	0.8986	0.7203
0.05	0.7642	0.8492	0.9836	0.8793	0.7415
0.1	0.7888	0.8738	0.9697	0.8590	0.7625
0.15	0.8126	0.8968	0.9515	0.8376	0.7833
0.2	0.8353	0.9181	0.9294	0.8154	0.8038
0.25	0.8568	0.9373	0.9035	0.7925	0.8240
0.3	0.8771	0.9542	0.8744	0.7690	0.8436
0.35	0.8960	0.9687	0.8425	0.7450	0.8627
0.4	0.9134	0.9806	0.8084	0.7206	0.8812
0.45	0.9293	0.9897	0.7726	0.6960	0.8988
0.5	0.9436	0.9960	0.7358	0.6711	0.9156
0.55	0.9563	0.9994	0.6984	0.6461	0.9313

0.6	0.9674	1.0000	0.6609	0.6212	0.9459
0.65	0.9769	0.9978	0.6238	0.5963	0.9590
0.7	0.9848	0.9930	0.5874	0.5716	0.9707
0.75	0.9910	0.9856	0.5520	0.5472	0.9806
0.8	0.9956	0.9759	0.5179	0.5230	0.9887
0.85	0.9986	0.9639	0.4851	0.4993	0.9947
0.9	1.0000	0.9499	0.4539	0.4760	0.9985
0.95	0.9998	0.9340	0.4242	0.4533	1.0000
1	0.9981	0.9165	0.3961	0.4311	0.9991
1.05	0.9949	0.8975	0.3696	0.4095	0.9957
1.1	0.9901	0.8772	0.3447	0.3885	0.9898
1.15	0.9839	0.8558	0.3214	0.3683	0.9815
1.2	0.9762	0.8335	0.2995	0.3487	0.9708
1.25	0.9671	0.8104	0.2790	0.3299	0.9580
1.3	0.9567	0.7867	0.2599	0.3118	0.9431
1.35	0.9450	0.7626	0.2420	0.2944	0.9264
1.4	0.9320	0.7383	0.2254	0.2778	0.9082
1.45	0.9178	0.7137	0.2098	0.2619	0.8886
1.5	0.9024	0.6892	0.1954	0.2467	0.8679
1.55	0.8861	0.6648	0.1819	0.2323	0.8464
1.6	0.8687	0.6406	0.1694	0.2186	0.8243
1.65	0.8505	0.6167	0.1577	0.2056	0.8018
1.7	0.8315	0.5932	0.1468	0.1932	0.7791
1.75	0.8118	0.5701	0.1367	0.1816	0.7563
1.8	0.7914	0.5475	0.1273	0.1705	0.7337
1.85	0.7705	0.5255	0.1186	0.1600	0.7113
1.9	0.7492	0.5041	0.1104	0.1502	0.6892
1.95	0.7274	0.4832	0.1028	0.1408	0.6675
2	0.7054	0.4630	0.0958	0.1321	0.6462
2.05	0.6832	0.4434	0.0892	0.1238	0.6254
2.1	0.6609	0.4244	0.0831	0.1160	0.6051
2.15	0.6385	0.4061	0.0774	0.1087	0.5853
2.2	0.6161	0.3884	0.0721	0.1019	0.5661
2.25	0.5938	0.3713	0.0672	0.0954	0.5474
2.3	0.5717	0.3549	0.0626	0.0893	0.5292
2.35	0.5497	0.3390	0.0583	0.0837	0.5116
2.4	0.5281	0.3238	0.0543	0.0783	0.4944
2.45	0.5067	0.3091	0.0506	0.0733	0.4778
2.5	0.4857	0.2950	0.0472	0.0686	0.4616

2.55	0.4651	0.2814	0.0440	0.0642	0.4459
2.6	0.4450	0.2684	0.0410	0.0601	0.4307
2.65	0.4253	0.2559	0.0382	0.0563	0.4159
2.7	0.4061	0.2440	0.0356	0.0527	0.4015
2.75	0.3874	0.2325	0.0332	0.0493	0.3876
2.8	0.3693	0.2215	0.0310	0.0461	0.3741
2.85	0.3518	0.2110	0.0289	0.0432	0.3609
2.9	0.3348	0.2009	0.0269	0.0404	0.3481
2.95	0.3185	0.1913	0.0251	0.0378	0.3357
3	0.3027	0.1820	0.0234	0.0354	0.3237
3.05	0.2876	0.1733	0.0219	0.0331	0.3120
3.1	0.2730	0.1649	0.0204	0.0310	0.3007
3.15	0.2590	0.1568	0.0190	0.0290	0.2897
3.2	0.2456	0.1492	0.0178	0.0272	0.2790
3.25	0.2328	0.1419	0.0166	0.0255	0.2686
3.3	0.2206	0.1350	0.0155	0.0238	0.2586
3.35	0.2089	0.1284	0.0144	0.0223	0.2488
3.4	0.1978	0.1220	0.0135	0.0209	0.2393
3.45	0.1872	0.1160	0.0126	0.0196	0.2302
3.5	0.1771	0.1103	0.0117	0.0184	0.2213
3.55	0.1676	0.1049	0.0110	0.0172	0.2127
3.6	0.1585	0.0997	0.0102	0.0161	0.2044
3.65	0.1498	0.0948	0.0096	0.0151	0.1964
3.7	0.1416	0.0901	0.0089	0.0142	0.1886
3.75	0.1339	0.0857	0.0084	0.0133	0.1811
3.8	0.1265	0.0815	0.0078	0.0124	0.1738
3.85	0.1196	0.0775	0.0073	0.0117	0.1668
3.9	0.1130	0.0736	0.0068	0.0109	0.1600
3.95	0.1067	0.0700	0.0064	0.0103	0.1535
4	0.1009	0.0666	0.0059	0.0096	0.1472

Table C4. Observed Relative Test Information for the Multidimensional Personality Questionnaire at Specific Scores – Scale 11.

θ	Wellbeing
-4	0.0028
-3.95	0.0030
-3.9	0.0032
-3.85	0.0034
-3.8	0.0036

-3.75	0.0039
-3.7	0.0042
-3.65	0.0045
-3.6	0.0048
-3.55	0.0052
-3.5	0.0055
-3.45	0.0059
-3.4	0.0064
-3.35	0.0068
-3.3	0.0073
-3.25	0.0078
-3.2	0.0084
-3.15	0.0090
-3.1	0.0097
-3.05	0.0104
-3	0.0111
-2.95	0.0119
-2.9	0.0128
-2.85	0.0137
-2.8	0.0147
-2.75	0.0157
-2.7	0.0168
-2.65	0.0180
-2.6	0.0193
-2.55	0.0207
-2.5	0.0222
-2.45	0.0238
-2.4	0.0254
-2.35	0.0272
-2.3	0.0292
-2.25	0.0312
-2.2	0.0334
-2.15	0.0358
-2.1	0.0382
-2.05	0.0409
-2	0.0437
-1.95	0.0467
-1.9	0.0499
-1.85	0.0533

-1.8	0.0570
-1.75	0.0608
-1.7	0.0649
-1.65	0.0692
-1.6	0.0738
-1.55	0.0786
-1.5	0.0837
-1.45	0.0891
-1.4	0.0948
-1.35	0.1009
-1.3	0.1072
-1.25	0.1139
-1.2	0.1209
-1.15	0.1283
-1.1	0.1361
-1.05	0.1443
-1	0.1529
-0.95	0.1619
-0.9	0.1713
-0.85	0.1812
-0.8	0.1915
-0.75	0.2024
-0.7	0.2137
-0.65	0.2256
-0.6	0.2381
-0.55	0.2511
-0.5	0.2647
-0.45	0.2790
-0.4	0.2939
-0.35	0.3095
-0.3	0.3258
-0.25	0.3429
-0.2	0.3607
-0.15	0.3794
-0.1	0.3988
-0.05	0.4192
0	0.4404
0.05	0.4626
0.1	0.4856

0.15	0.5096
0.2	0.5345
0.25	0.5603
0.3	0.5870
0.35	0.6144
0.4	0.6426
0.45	0.6714
0.5	0.7006
0.55	0.7301
0.6	0.7597
0.65	0.7890
0.7	0.8178
0.75	0.8457
0.8	0.8724
0.85	0.8974
0.9	0.9204
0.95	0.9411
1	0.9590
1.05	0.9740
1.1	0.9857
1.15	0.9940
1.2	0.9988
1.25	1.0000
1.3	0.9977
1.35	0.9920
1.4	0.9828
1.45	0.9704
1.5	0.9549
1.55	0.9364
1.6	0.9151
1.65	0.8913
1.7	0.8650
1.75	0.8367
1.8	0.8065
1.85	0.7749
1.9	0.7421
1.95	0.7086
2	0.6745
2.05	0.6404

2.1	0.6065
2.15	0.5730
2.2	0.5403
2.25	0.5085
2.3	0.4778
2.35	0.4483
2.4	0.4200
2.45	0.3932
2.5	0.3677
2.55	0.3437
2.6	0.3210
2.65	0.2996
2.7	0.2795
2.75	0.2608
2.8	0.2432
2.85	0.2267
2.9	0.2113
2.95	0.1970
3	0.1836
3.05	0.1711
3.1	0.1595
3.15	0.1487
3.2	0.1386
3.25	0.1292
3.3	0.1204
3.35	0.1123
3.4	0.1047
3.45	0.0976
3.5	0.0910
3.55	0.0849
3.6	0.0792
3.65	0.0739
3.7	0.0689
3.75	0.0643
3.8	0.0600
3.85	0.0560
3.9	0.0523
3.95	0.0488
4	0.0456

18 Scales – Stability Facets of the Big Five

Table C5. Observed Relative Test Information for the Stability Facets of the International Pool of Personality Items at Specific Scores – Scales 1 through 5.

θ	Trust	Morality	Altruism	Cooperation	Modesty
-4	0.0860	0.2767	0.5020	0.1607	0.0576
-3.95	0.0918	0.2866	0.5238	0.1681	0.0585
-3.9	0.0978	0.2968	0.5460	0.1759	0.0593
-3.85	0.1044	0.3074	0.5684	0.1843	0.0601
-3.8	0.1113	0.3185	0.5911	0.1933	0.0609
-3.75	0.1187	0.3302	0.6140	0.2030	0.0617
-3.7	0.1266	0.3425	0.6369	0.2133	0.0625
-3.65	0.1350	0.3556	0.6599	0.2243	0.0633
-3.6	0.1439	0.3694	0.6827	0.2361	0.0641
-3.55	0.1534	0.3841	0.7053	0.2487	0.0648
-3.5	0.1636	0.3999	0.7275	0.2621	0.0656
-3.45	0.1743	0.4166	0.7493	0.2763	0.0663
-3.4	0.1857	0.4346	0.7705	0.2915	0.0670
-3.35	0.1979	0.4538	0.7909	0.3076	0.0677
-3.3	0.2107	0.4743	0.8103	0.3247	0.0684
-3.25	0.2244	0.4961	0.8287	0.3428	0.0690
-3.2	0.2388	0.5193	0.8459	0.3618	0.0697
-3.15	0.2541	0.5439	0.8616	0.3818	0.0703
-3.1	0.2703	0.5698	0.8759	0.4028	0.0709
-3.05	0.2874	0.5969	0.8887	0.4248	0.0714
-3	0.3054	0.6250	0.8998	0.4476	0.0720
-2.95	0.3243	0.6539	0.9093	0.4712	0.0725
-2.9	0.3443	0.6833	0.9173	0.4955	0.0731
-2.85	0.3652	0.7126	0.9238	0.5205	0.0736
-2.8	0.3872	0.7415	0.9289	0.5459	0.0742
-2.75	0.4101	0.7694	0.9328	0.5716	0.0748
-2.7	0.4341	0.7956	0.9358	0.5974	0.0754
-2.65	0.4590	0.8194	0.9380	0.6232	0.0760
-2.6	0.4848	0.8405	0.9398	0.6486	0.0768
-2.55	0.5114	0.8582	0.9414	0.6734	0.0776
-2.5	0.5389	0.8723	0.9430	0.6975	0.0786
-2.45	0.5671	0.8825	0.9448	0.7205	0.0798
-2.4	0.5958	0.8891	0.9469	0.7423	0.0813
-2.35	0.6249	0.8923	0.9495	0.7626	0.0831

-2.3	0.6542	0.8927	0.9527	0.7814	0.0853
-2.25	0.6836	0.8908	0.9563	0.7984	0.0881
-2.2	0.7126	0.8875	0.9605	0.8137	0.0917
-2.15	0.7411	0.8837	0.9650	0.8272	0.0962
-2.1	0.7687	0.8800	0.9698	0.8390	0.1020
-2.05	0.7951	0.8773	0.9746	0.8491	0.1093
-2	0.8197	0.8761	0.9794	0.8578	0.1186
-1.95	0.8423	0.8768	0.9840	0.8653	0.1304
-1.9	0.8623	0.8798	0.9881	0.8717	0.1453
-1.85	0.8795	0.8851	0.9917	0.8773	0.1640
-1.8	0.8935	0.8925	0.9948	0.8823	0.1876
-1.75	0.9042	0.9018	0.9971	0.8871	0.2168
-1.7	0.9113	0.9125	0.9988	0.8917	0.2529
-1.65	0.9152	0.9243	0.9997	0.8964	0.2968
-1.6	0.9158	0.9364	1.0000	0.9014	0.3494
-1.55	0.9137	0.9483	0.9996	0.9066	0.4112
-1.5	0.9094	0.9596	0.9986	0.9123	0.4820
-1.45	0.9035	0.9698	0.9969	0.9183	0.5606
-1.4	0.8966	0.9786	0.9945	0.9247	0.6443
-1.35	0.8895	0.9860	0.9915	0.9313	0.7288
-1.3	0.8830	0.9917	0.9879	0.9382	0.8082
-1.25	0.8775	0.9959	0.9834	0.9451	0.8756
-1.2	0.8736	0.9987	0.9782	0.9519	0.9241
-1.15	0.8717	1.0000	0.9722	0.9586	0.9481
-1.1	0.8721	0.9999	0.9654	0.9650	0.9448
-1.05	0.8748	0.9985	0.9578	0.9710	0.9150
-1	0.8799	0.9957	0.9494	0.9766	0.8628
-0.95	0.8872	0.9915	0.9403	0.9817	0.7949
-0.9	0.8964	0.9860	0.9307	0.9862	0.7190
-0.85	0.9070	0.9795	0.9206	0.9900	0.6425
-0.8	0.9188	0.9722	0.9104	0.9933	0.5717
-0.75	0.9311	0.9644	0.9001	0.9959	0.5110
-0.7	0.9435	0.9567	0.8901	0.9978	0.4636
-0.65	0.9555	0.9495	0.8805	0.9991	0.4311
-0.6	0.9665	0.9432	0.8716	0.9998	0.4145
-0.55	0.9763	0.9381	0.8636	1.0000	0.4141
-0.5	0.9845	0.9345	0.8567	0.9997	0.4300
-0.45	0.9910	0.9323	0.8511	0.9989	0.4618
-0.4	0.9957	0.9313	0.8469	0.9978	0.5087

-0.35	0.9987	0.9312	0.8441	0.9964	0.5691
-0.3	1.0000	0.9314	0.8428	0.9949	0.6403
-0.25	0.9997	0.9314	0.8429	0.9932	0.7182
-0.2	0.9979	0.9304	0.8444	0.9914	0.7971
-0.15	0.9945	0.9277	0.8471	0.9895	0.8703
-0.1	0.9895	0.9226	0.8508	0.9875	0.9311
-0.05	0.9828	0.9148	0.8553	0.9853	0.9741
0	0.9743	0.9039	0.8603	0.9829	0.9966
0.05	0.9638	0.8897	0.8654	0.9802	0.9996
0.1	0.9515	0.8723	0.8703	0.9769	0.9879
0.15	0.9372	0.8521	0.8746	0.9730	0.9688
0.2	0.9212	0.8293	0.8779	0.9683	0.9504
0.25	0.9037	0.8046	0.8800	0.9627	0.9391
0.3	0.8854	0.7785	0.8804	0.9560	0.9388
0.35	0.8666	0.7513	0.8788	0.9481	0.9495
0.4	0.8480	0.7238	0.8752	0.9389	0.9675
0.45	0.8302	0.6962	0.8692	0.9284	0.9865
0.5	0.8139	0.6688	0.8608	0.9166	0.9994
0.55	0.7996	0.6420	0.8500	0.9035	1.0000
0.6	0.7878	0.6159	0.8367	0.8891	0.9847
0.65	0.7788	0.5906	0.8212	0.8736	0.9536
0.7	0.7729	0.5661	0.8035	0.8570	0.9102
0.75	0.7701	0.5426	0.7839	0.8395	0.8605
0.8	0.7703	0.5199	0.7626	0.8212	0.8117
0.85	0.7734	0.4982	0.7398	0.8022	0.7708
0.9	0.7789	0.4773	0.7159	0.7826	0.7434
0.95	0.7863	0.4572	0.6911	0.7627	0.7331
1	0.7949	0.4380	0.6656	0.7425	0.7413
1.05	0.8041	0.4196	0.6398	0.7222	0.7666
1.1	0.8132	0.4020	0.6138	0.7018	0.8051
1.15	0.8213	0.3852	0.5878	0.6816	0.8508
1.2	0.8278	0.3691	0.5621	0.6614	0.8958
1.25	0.8321	0.3538	0.5367	0.6416	0.9318
1.3	0.8336	0.3393	0.5118	0.6221	0.9514
1.35	0.8319	0.3254	0.4876	0.6030	0.9495
1.4	0.8268	0.3122	0.4641	0.5843	0.9242
1.45	0.8182	0.2997	0.4413	0.5661	0.8770
1.5	0.8061	0.2879	0.4193	0.5484	0.8121
1.55	0.7907	0.2766	0.3981	0.5313	0.7354

1.6	0.7721	0.2659	0.3778	0.5146	0.6529
1.65	0.7507	0.2558	0.3584	0.4985	0.5702
1.7	0.7268	0.2463	0.3397	0.4828	0.4915
1.75	0.7009	0.2372	0.3220	0.4677	0.4197
1.8	0.6733	0.2286	0.3050	0.4531	0.3563
1.85	0.6444	0.2204	0.2889	0.4389	0.3017
1.9	0.6147	0.2126	0.2735	0.4251	0.2557
1.95	0.5845	0.2053	0.2588	0.4117	0.2174
2	0.5542	0.1982	0.2449	0.3988	0.1861
2.05	0.5242	0.1916	0.2317	0.3861	0.1606
2.1	0.4946	0.1852	0.2192	0.3738	0.1401
2.15	0.4657	0.1791	0.2073	0.3618	0.1236
2.2	0.4378	0.1733	0.1960	0.3501	0.1104
2.25	0.4109	0.1677	0.1853	0.3386	0.0998
2.3	0.3852	0.1623	0.1752	0.3273	0.0913
2.35	0.3608	0.1572	0.1655	0.3163	0.0845
2.4	0.3375	0.1522	0.1564	0.3055	0.0790
2.45	0.3156	0.1474	0.1478	0.2950	0.0746
2.5	0.2950	0.1428	0.1396	0.2846	0.0709
2.55	0.2756	0.1383	0.1318	0.2744	0.0679
2.6	0.2574	0.1340	0.1245	0.2644	0.0655
2.65	0.2403	0.1297	0.1175	0.2547	0.0634
2.7	0.2244	0.1256	0.1109	0.2451	0.0616
2.75	0.2095	0.1216	0.1047	0.2357	0.0600
2.8	0.1955	0.1177	0.0988	0.2266	0.0587
2.85	0.1825	0.1138	0.0932	0.2176	0.0574
2.9	0.1704	0.1101	0.0879	0.2089	0.0563
2.95	0.1591	0.1064	0.0829	0.2004	0.0553
3	0.1485	0.1028	0.0782	0.1922	0.0544
3.05	0.1386	0.0993	0.0737	0.1841	0.0535
3.1	0.1294	0.0959	0.0695	0.1763	0.0527
3.15	0.1208	0.0925	0.0655	0.1688	0.0519
3.2	0.1128	0.0892	0.0617	0.1615	0.0511
3.25	0.1053	0.0860	0.0582	0.1544	0.0504
3.3	0.0983	0.0828	0.0548	0.1476	0.0497
3.35	0.0917	0.0797	0.0516	0.1410	0.0490
3.4	0.0856	0.0767	0.0486	0.1347	0.0483
3.45	0.0799	0.0738	0.0458	0.1286	0.0476
3.5	0.0746	0.0709	0.0431	0.1227	0.0470

3.55	0.0696	0.0681	0.0406	0.1171	0.0463
3.6	0.0650	0.0653	0.0382	0.1117	0.0457
3.65	0.0607	0.0627	0.0359	0.1065	0.0451
3.7	0.0566	0.0601	0.0338	0.1016	0.0444
3.75	0.0528	0.0576	0.0318	0.0969	0.0438
3.8	0.0493	0.0551	0.0300	0.0923	0.0432
3.85	0.0460	0.0528	0.0282	0.0880	0.0426
3.9	0.0430	0.0505	0.0265	0.0839	0.0420
3.95	0.0401	0.0482	0.0249	0.0800	0.0414
4	0.0374	0.0461	0.0235	0.0762	0.0408

Table C6. Observed Relative Test Information for the Stability Facets of the International Pool of Personality Items at Specific Scores – Scales 6 through 10.

θ	Sympathy	Self-Efficacy	Orderliness	Dutifulness	Achievement-Striving
-4	0.3771	0.6111	0.1025	0.6417	0.2844
-3.95	0.3925	0.6350	0.1084	0.6607	0.2984
-3.9	0.4083	0.6587	0.1147	0.6793	0.3131
-3.85	0.4245	0.6822	0.1215	0.6975	0.3285
-3.8	0.4411	0.7054	0.1287	0.7152	0.3446
-3.75	0.4581	0.7280	0.1364	0.7324	0.3615
-3.7	0.4755	0.7500	0.1445	0.7490	0.3791
-3.65	0.4933	0.7711	0.1533	0.7649	0.3976
-3.6	0.5115	0.7913	0.1625	0.7801	0.4167
-3.55	0.5300	0.8104	0.1724	0.7946	0.4367
-3.5	0.5489	0.8283	0.1829	0.8083	0.4574
-3.45	0.5680	0.8448	0.1941	0.8212	0.4788
-3.4	0.5874	0.8598	0.2059	0.8332	0.5008
-3.35	0.6071	0.8732	0.2184	0.8444	0.5234
-3.3	0.6269	0.8850	0.2316	0.8548	0.5465
-3.25	0.6468	0.8951	0.2455	0.8644	0.5699
-3.2	0.6668	0.9035	0.2602	0.8731	0.5934
-3.15	0.6867	0.9103	0.2755	0.8811	0.6171
-3.1	0.7065	0.9156	0.2915	0.8884	0.6406
-3.05	0.7261	0.9195	0.3081	0.8951	0.6638
-3	0.7453	0.9220	0.3254	0.9012	0.6865
-2.95	0.7641	0.9235	0.3432	0.9068	0.7085
-2.9	0.7823	0.9241	0.3616	0.9119	0.7297
-2.85	0.7999	0.9241	0.3806	0.9168	0.7498
-2.8	0.8167	0.9236	0.3999	0.9214	0.7688

-2.75	0.8327	0.9229	0.4196	0.9258	0.7865
-2.7	0.8477	0.9223	0.4396	0.9301	0.8029
-2.65	0.8617	0.9219	0.4599	0.9344	0.8180
-2.6	0.8747	0.9219	0.4804	0.9387	0.8317
-2.55	0.8867	0.9225	0.5011	0.9429	0.8442
-2.5	0.8976	0.9238	0.5219	0.9473	0.8555
-2.45	0.9074	0.9258	0.5427	0.9517	0.8657
-2.4	0.9163	0.9286	0.5635	0.9562	0.8751
-2.35	0.9243	0.9321	0.5843	0.9606	0.8836
-2.3	0.9314	0.9363	0.6050	0.9651	0.8916
-2.25	0.9378	0.9412	0.6256	0.9695	0.8991
-2.2	0.9435	0.9465	0.6460	0.9738	0.9063
-2.15	0.9487	0.9523	0.6662	0.9779	0.9132
-2.1	0.9534	0.9582	0.6861	0.9818	0.9199
-2.05	0.9578	0.9642	0.7056	0.9854	0.9266
-2	0.9618	0.9701	0.7247	0.9887	0.9330
-1.95	0.9656	0.9757	0.7432	0.9917	0.9394
-1.9	0.9692	0.9810	0.7611	0.9943	0.9456
-1.85	0.9726	0.9857	0.7782	0.9964	0.9516
-1.8	0.9759	0.9898	0.7946	0.9980	0.9573
-1.75	0.9791	0.9932	0.8100	0.9992	0.9628
-1.7	0.9821	0.9960	0.8246	0.9999	0.9679
-1.65	0.9850	0.9980	0.8382	1.0000	0.9727
-1.6	0.9877	0.9994	0.8509	0.9996	0.9771
-1.55	0.9903	1.0000	0.8626	0.9987	0.9811
-1.5	0.9926	1.0000	0.8734	0.9972	0.9847
-1.45	0.9946	0.9994	0.8835	0.9951	0.9880
-1.4	0.9964	0.9981	0.8928	0.9924	0.9909
-1.35	0.9979	0.9963	0.9016	0.9891	0.9935
-1.3	0.9990	0.9939	0.9098	0.9853	0.9956
-1.25	0.9997	0.9909	0.9177	0.9808	0.9974
-1.2	1.0000	0.9872	0.9253	0.9758	0.9988
-1.15	0.9999	0.9828	0.9326	0.9703	0.9996
-1.1	0.9992	0.9777	0.9398	0.9643	1.0000
-1.05	0.9981	0.9718	0.9468	0.9579	0.9998
-1	0.9965	0.9651	0.9536	0.9512	0.9989
-0.95	0.9943	0.9574	0.9602	0.9443	0.9973
-0.9	0.9917	0.9490	0.9665	0.9373	0.9950
-0.85	0.9885	0.9396	0.9725	0.9303	0.9919

-0.8	0.9850	0.9295	0.9780	0.9235	0.9881
-0.75	0.9811	0.9187	0.9829	0.9169	0.9836
-0.7	0.9769	0.9073	0.9873	0.9107	0.9785
-0.65	0.9725	0.8954	0.9911	0.9050	0.9729
-0.6	0.9678	0.8832	0.9942	0.8999	0.9669
-0.55	0.9632	0.8710	0.9966	0.8955	0.9606
-0.5	0.9585	0.8588	0.9983	0.8918	0.9542
-0.45	0.9540	0.8470	0.9995	0.8889	0.9479
-0.4	0.9496	0.8357	1.0000	0.8868	0.9417
-0.35	0.9454	0.8252	1.0000	0.8853	0.9358
-0.3	0.9416	0.8156	0.9994	0.8846	0.9304
-0.25	0.9381	0.8071	0.9984	0.8845	0.9255
-0.2	0.9350	0.7998	0.9969	0.8848	0.9213
-0.15	0.9322	0.7939	0.9950	0.8856	0.9177
-0.1	0.9298	0.7895	0.9927	0.8865	0.9149
-0.05	0.9279	0.7866	0.9901	0.8876	0.9127
0	0.9262	0.7851	0.9871	0.8885	0.9112
0.05	0.9248	0.7852	0.9839	0.8893	0.9101
0.1	0.9236	0.7866	0.9805	0.8895	0.9095
0.15	0.9226	0.7894	0.9769	0.8892	0.9090
0.2	0.9216	0.7933	0.9733	0.8881	0.9085
0.25	0.9204	0.7983	0.9696	0.8862	0.9076
0.3	0.9191	0.8040	0.9661	0.8832	0.9063
0.35	0.9175	0.8103	0.9626	0.8791	0.9041
0.4	0.9154	0.8169	0.9594	0.8737	0.9009
0.45	0.9127	0.8234	0.9564	0.8671	0.8965
0.5	0.9093	0.8298	0.9537	0.8592	0.8906
0.55	0.9052	0.8356	0.9512	0.8499	0.8830
0.6	0.9002	0.8406	0.9489	0.8392	0.8738
0.65	0.8943	0.8446	0.9467	0.8272	0.8628
0.7	0.8873	0.8473	0.9446	0.8138	0.8499
0.75	0.8794	0.8486	0.9422	0.7992	0.8352
0.8	0.8704	0.8482	0.9396	0.7834	0.8189
0.85	0.8603	0.8461	0.9365	0.7666	0.8009
0.9	0.8493	0.8421	0.9326	0.7487	0.7814
0.95	0.8372	0.8363	0.9279	0.7299	0.7605
1	0.8242	0.8286	0.9219	0.7103	0.7386
1.05	0.8103	0.8190	0.9146	0.6900	0.7156
1.1	0.7955	0.8076	0.9057	0.6691	0.6919

1.15	0.7800	0.7945	0.8950	0.6479	0.6676
1.2	0.7639	0.7798	0.8825	0.6262	0.6430
1.25	0.7471	0.7635	0.8680	0.6044	0.6181
1.3	0.7299	0.7459	0.8516	0.5824	0.5932
1.35	0.7122	0.7271	0.8331	0.5604	0.5685
1.4	0.6942	0.7072	0.8127	0.5385	0.5441
1.45	0.6760	0.6863	0.7905	0.5168	0.5201
1.5	0.6577	0.6647	0.7667	0.4953	0.4966
1.55	0.6393	0.6424	0.7414	0.4741	0.4737
1.6	0.6209	0.6196	0.7148	0.4533	0.4514
1.65	0.6027	0.5965	0.6872	0.4329	0.4300
1.7	0.5845	0.5732	0.6589	0.4130	0.4092
1.75	0.5666	0.5497	0.6301	0.3936	0.3893
1.8	0.5490	0.5264	0.6010	0.3747	0.3702
1.85	0.5317	0.5031	0.5719	0.3565	0.3518
1.9	0.5147	0.4802	0.5430	0.3388	0.3343
1.95	0.4982	0.4576	0.5145	0.3218	0.3176
2	0.4820	0.4355	0.4865	0.3053	0.3016
2.05	0.4662	0.4139	0.4593	0.2895	0.2864
2.1	0.4509	0.3929	0.4328	0.2744	0.2719
2.15	0.4361	0.3725	0.4072	0.2598	0.2581
2.2	0.4216	0.3529	0.3827	0.2459	0.2450
2.25	0.4076	0.3339	0.3592	0.2326	0.2325
2.3	0.3941	0.3157	0.3367	0.2199	0.2207
2.35	0.3810	0.2983	0.3154	0.2078	0.2094
2.4	0.3683	0.2816	0.2952	0.1962	0.1986
2.45	0.3560	0.2657	0.2760	0.1852	0.1884
2.5	0.3442	0.2506	0.2579	0.1748	0.1787
2.55	0.3327	0.2362	0.2409	0.1649	0.1695
2.6	0.3216	0.2225	0.2249	0.1555	0.1607
2.65	0.3109	0.2096	0.2098	0.1466	0.1523
2.7	0.3005	0.1973	0.1957	0.1382	0.1444
2.75	0.2905	0.1857	0.1825	0.1302	0.1369
2.8	0.2808	0.1747	0.1701	0.1226	0.1297
2.85	0.2714	0.1644	0.1585	0.1155	0.1229
2.9	0.2623	0.1546	0.1477	0.1088	0.1164
2.95	0.2535	0.1454	0.1376	0.1024	0.1103
3	0.2450	0.1367	0.1282	0.0964	0.1044
3.05	0.2367	0.1285	0.1195	0.0908	0.0989

3.1	0.2287	0.1208	0.1113	0.0854	0.0936
3.15	0.2210	0.1135	0.1037	0.0804	0.0886
3.2	0.2135	0.1067	0.0966	0.0757	0.0839
3.25	0.2062	0.1002	0.0899	0.0712	0.0794
3.3	0.1991	0.0942	0.0838	0.0670	0.0751
3.35	0.1922	0.0885	0.0781	0.0630	0.0711
3.4	0.1856	0.0831	0.0727	0.0593	0.0672
3.45	0.1791	0.0781	0.0678	0.0558	0.0636
3.5	0.1729	0.0734	0.0631	0.0525	0.0602
3.55	0.1668	0.0689	0.0588	0.0494	0.0569
3.6	0.1609	0.0647	0.0548	0.0465	0.0538
3.65	0.1552	0.0608	0.0511	0.0438	0.0509
3.7	0.1496	0.0571	0.0476	0.0412	0.0481
3.75	0.1442	0.0536	0.0444	0.0388	0.0455
3.8	0.1390	0.0504	0.0414	0.0365	0.0430
3.85	0.1340	0.0473	0.0386	0.0344	0.0406
3.9	0.1291	0.0444	0.0360	0.0324	0.0384
3.95	0.1243	0.0417	0.0336	0.0305	0.0363
4	0.1197	0.0392	0.0313	0.0287	0.0343

Table C7. Observed Relative Test Information for the Stability Facets of the International Pool of Personality Items at Specific Scores – Scales 11 through 15.

θ	Self-Discipline	Cautiousness	Anxiety	Anger	Depression
-4	0.0743	0.0553	0.0647	0.0145	0.0115
-3.95	0.0786	0.0569	0.0701	0.0155	0.0125
-3.9	0.0832	0.0587	0.0758	0.0166	0.0135
-3.85	0.0880	0.0605	0.0821	0.0177	0.0146
-3.8	0.0929	0.0625	0.0889	0.0189	0.0159
-3.75	0.0981	0.0646	0.0962	0.0203	0.0172
-3.7	0.1036	0.0669	0.1040	0.0217	0.0187
-3.65	0.1092	0.0694	0.1125	0.0233	0.0203
-3.6	0.1152	0.0721	0.1217	0.0251	0.0220
-3.55	0.1214	0.0750	0.1315	0.0270	0.0239
-3.5	0.1279	0.0782	0.1421	0.0290	0.0260
-3.45	0.1347	0.0816	0.1534	0.0313	0.0282
-3.4	0.1419	0.0854	0.1656	0.0339	0.0307
-3.35	0.1495	0.0895	0.1785	0.0367	0.0333
-3.3	0.1575	0.0940	0.1924	0.0398	0.0362
-3.25	0.1660	0.0989	0.2071	0.0432	0.0394

-3.2	0.1750	0.1043	0.2228	0.0470	0.0428
-3.15	0.1846	0.1102	0.2394	0.0512	0.0466
-3.1	0.1949	0.1167	0.2569	0.0560	0.0506
-3.05	0.2059	0.1238	0.2753	0.0612	0.0551
-3	0.2177	0.1316	0.2947	0.0671	0.0599
-2.95	0.2304	0.1402	0.3149	0.0737	0.0651
-2.9	0.2440	0.1496	0.3360	0.0811	0.0708
-2.85	0.2587	0.1599	0.3578	0.0893	0.0770
-2.8	0.2746	0.1712	0.3803	0.0985	0.0837
-2.75	0.2917	0.1836	0.4035	0.1088	0.0909
-2.7	0.3101	0.1971	0.4271	0.1203	0.0988
-2.65	0.3300	0.2119	0.4511	0.1331	0.1073
-2.6	0.3512	0.2281	0.4754	0.1473	0.1165
-2.55	0.3741	0.2456	0.4999	0.1632	0.1264
-2.5	0.3984	0.2647	0.5244	0.1807	0.1371
-2.45	0.4243	0.2854	0.5488	0.2000	0.1485
-2.4	0.4517	0.3078	0.5730	0.2212	0.1608
-2.35	0.4806	0.3320	0.5968	0.2445	0.1739
-2.3	0.5108	0.3579	0.6203	0.2697	0.1880
-2.25	0.5421	0.3857	0.6433	0.2971	0.2029
-2.2	0.5742	0.4154	0.6657	0.3265	0.2187
-2.15	0.6069	0.4468	0.6876	0.3578	0.2355
-2.1	0.6398	0.4799	0.7089	0.3909	0.2532
-2.05	0.6725	0.5145	0.7296	0.4256	0.2717
-2	0.7044	0.5504	0.7497	0.4617	0.2912
-1.95	0.7349	0.5872	0.7692	0.4986	0.3116
-1.9	0.7637	0.6244	0.7881	0.5360	0.3328
-1.85	0.7900	0.6615	0.8063	0.5735	0.3548
-1.8	0.8135	0.6976	0.8239	0.6104	0.3776
-1.75	0.8338	0.7320	0.8408	0.6463	0.4012
-1.7	0.8504	0.7639	0.8568	0.6807	0.4255
-1.65	0.8634	0.7923	0.8720	0.7130	0.4505
-1.6	0.8725	0.8165	0.8861	0.7429	0.4761
-1.55	0.8781	0.8359	0.8993	0.7700	0.5024
-1.5	0.8803	0.8499	0.9113	0.7940	0.5292
-1.45	0.8796	0.8586	0.9221	0.8148	0.5566
-1.4	0.8766	0.8620	0.9317	0.8323	0.5844
-1.35	0.8719	0.8607	0.9401	0.8467	0.6125
-1.3	0.8661	0.8554	0.9474	0.8582	0.6407

-1.25	0.8601	0.8471	0.9538	0.8670	0.6689
-1.2	0.8543	0.8370	0.9592	0.8736	0.6969
-1.15	0.8495	0.8262	0.9639	0.8783	0.7243
-1.1	0.8461	0.8160	0.9680	0.8817	0.7509
-1.05	0.8446	0.8076	0.9717	0.8843	0.7764
-1	0.8453	0.8017	0.9752	0.8866	0.8006
-0.95	0.8483	0.7992	0.9785	0.8891	0.8230
-0.9	0.8536	0.8006	0.9817	0.8923	0.8436
-0.85	0.8611	0.8061	0.9848	0.8963	0.8622
-0.8	0.8706	0.8155	0.9879	0.9016	0.8787
-0.75	0.8818	0.8285	0.9908	0.9081	0.8930
-0.7	0.8942	0.8446	0.9935	0.9159	0.9053
-0.65	0.9073	0.8630	0.9958	0.9249	0.9156
-0.6	0.9208	0.8827	0.9977	0.9349	0.9242
-0.55	0.9341	0.9027	0.9991	0.9455	0.9313
-0.5	0.9469	0.9223	0.9999	0.9563	0.9373
-0.45	0.9587	0.9405	1.0000	0.9667	0.9425
-0.4	0.9693	0.9568	0.9994	0.9765	0.9472
-0.35	0.9786	0.9706	0.9981	0.9850	0.9516
-0.3	0.9862	0.9819	0.9961	0.9919	0.9561
-0.25	0.9923	0.9904	0.9934	0.9969	0.9607
-0.2	0.9966	0.9962	0.9902	0.9996	0.9655
-0.15	0.9992	0.9994	0.9866	1.0000	0.9705
-0.1	1.0000	1.0000	0.9825	0.9980	0.9756
-0.05	0.9990	0.9981	0.9782	0.9936	0.9807
0	0.9963	0.9939	0.9737	0.9870	0.9855
0.05	0.9919	0.9873	0.9691	0.9782	0.9899
0.1	0.9858	0.9786	0.9647	0.9675	0.9938
0.15	0.9782	0.9679	0.9606	0.9552	0.9968
0.2	0.9692	0.9557	0.9568	0.9414	0.9989
0.25	0.9591	0.9423	0.9536	0.9266	1.0000
0.3	0.9483	0.9285	0.9510	0.9113	0.9999
0.35	0.9371	0.9148	0.9490	0.8958	0.9987
0.4	0.9259	0.9019	0.9478	0.8808	0.9962
0.45	0.9150	0.8905	0.9473	0.8669	0.9926
0.5	0.9050	0.8813	0.9473	0.8548	0.9879
0.55	0.8960	0.8746	0.9477	0.8449	0.9823
0.6	0.8884	0.8707	0.9484	0.8378	0.9760
0.65	0.8825	0.8696	0.9490	0.8338	0.9691

0.7	0.8782	0.8710	0.9492	0.8330	0.9619
0.75	0.8756	0.8742	0.9488	0.8353	0.9546
0.8	0.8747	0.8786	0.9474	0.8404	0.9476
0.85	0.8751	0.8830	0.9448	0.8477	0.9410
0.9	0.8766	0.8864	0.9405	0.8565	0.9351
0.95	0.8787	0.8878	0.9345	0.8658	0.9300
1	0.8810	0.8861	0.9266	0.8747	0.9258
1.05	0.8830	0.8806	0.9167	0.8821	0.9223
1.1	0.8841	0.8709	0.9047	0.8869	0.9195
1.15	0.8840	0.8566	0.8907	0.8882	0.9171
1.2	0.8822	0.8380	0.8749	0.8852	0.9148
1.25	0.8784	0.8152	0.8573	0.8774	0.9121
1.3	0.8724	0.7889	0.8381	0.8643	0.9085
1.35	0.8639	0.7595	0.8176	0.8459	0.9037
1.4	0.8530	0.7278	0.7961	0.8225	0.8971
1.45	0.8397	0.6944	0.7736	0.7943	0.8885
1.5	0.8241	0.6599	0.7505	0.7621	0.8776
1.55	0.8063	0.6250	0.7270	0.7265	0.8640
1.6	0.7866	0.5901	0.7032	0.6884	0.8479
1.65	0.7651	0.5557	0.6794	0.6487	0.8290
1.7	0.7422	0.5221	0.6557	0.6082	0.8076
1.75	0.7179	0.4896	0.6322	0.5677	0.7837
1.8	0.6927	0.4584	0.6091	0.5278	0.7575
1.85	0.6667	0.4288	0.5864	0.4891	0.7292
1.9	0.6402	0.4009	0.5642	0.4521	0.6991
1.95	0.6133	0.3746	0.5425	0.4172	0.6676
2	0.5862	0.3501	0.5214	0.3844	0.6350
2.05	0.5592	0.3274	0.5010	0.3540	0.6016
2.1	0.5325	0.3064	0.4811	0.3261	0.5678
2.15	0.5061	0.2871	0.4618	0.3005	0.5340
2.2	0.4802	0.2693	0.4432	0.2772	0.5004
2.25	0.4549	0.2530	0.4251	0.2561	0.4674
2.3	0.4303	0.2381	0.4076	0.2371	0.4353
2.35	0.4066	0.2245	0.3907	0.2199	0.4043
2.4	0.3836	0.2120	0.3743	0.2045	0.3745
2.45	0.3616	0.2006	0.3584	0.1907	0.3460
2.5	0.3405	0.1902	0.3430	0.1783	0.3191
2.55	0.3204	0.1806	0.3281	0.1671	0.2936
2.6	0.3011	0.1718	0.3137	0.1570	0.2698

2.65	0.2828	0.1637	0.2997	0.1479	0.2475
2.7	0.2655	0.1563	0.2862	0.1396	0.2267
2.75	0.2490	0.1493	0.2732	0.1320	0.2074
2.8	0.2334	0.1429	0.2606	0.1250	0.1896
2.85	0.2187	0.1369	0.2484	0.1186	0.1731
2.9	0.2047	0.1313	0.2367	0.1126	0.1579
2.95	0.1916	0.1261	0.2254	0.1070	0.1440
3	0.1792	0.1211	0.2146	0.1017	0.1312
3.05	0.1676	0.1165	0.2041	0.0967	0.1195
3.1	0.1566	0.1121	0.1941	0.0920	0.1088
3.15	0.1463	0.1079	0.1845	0.0875	0.0991
3.2	0.1366	0.1039	0.1753	0.0832	0.0901
3.25	0.1275	0.1001	0.1665	0.0790	0.0820
3.3	0.1190	0.0965	0.1581	0.0751	0.0746
3.35	0.1110	0.0931	0.1500	0.0712	0.0679
3.4	0.1035	0.0897	0.1424	0.0675	0.0618
3.45	0.0965	0.0865	0.1351	0.0640	0.0562
3.5	0.0899	0.0835	0.1281	0.0606	0.0511
3.55	0.0838	0.0805	0.1215	0.0573	0.0465
3.6	0.0780	0.0777	0.1152	0.0541	0.0423
3.65	0.0726	0.0750	0.1092	0.0511	0.0386
3.7	0.0676	0.0723	0.1035	0.0482	0.0351
3.75	0.0630	0.0698	0.0981	0.0454	0.0320
3.8	0.0586	0.0673	0.0930	0.0427	0.0291
3.85	0.0545	0.0649	0.0882	0.0402	0.0266
3.9	0.0507	0.0626	0.0836	0.0377	0.0242
3.95	0.0472	0.0604	0.0792	0.0354	0.0221
4	0.0439	0.0582	0.0751	0.0332	0.0202

Table C8. Observed Relative Test Information for the Stability Facets of the International Pool of Personality Items at Specific Scores – Scales 16 through 18.

θ	Self-Consciousness	Immoderation	Vulnerability
-4	0.0819	0.1895	0.0427
-3.95	0.0872	0.1966	0.0455
-3.9	0.0928	0.2040	0.0485
-3.85	0.0987	0.2116	0.0516
-3.8	0.1050	0.2196	0.0550
-3.75	0.1116	0.2279	0.0586
-3.7	0.1187	0.2365	0.0624

-3.65	0.1261	0.2455	0.0665
-3.6	0.1340	0.2549	0.0708
-3.55	0.1422	0.2646	0.0754
-3.5	0.1510	0.2748	0.0803
-3.45	0.1602	0.2855	0.0855
-3.4	0.1698	0.2966	0.0911
-3.35	0.1800	0.3083	0.0970
-3.3	0.1907	0.3205	0.1032
-3.25	0.2019	0.3332	0.1099
-3.2	0.2137	0.3466	0.1170
-3.15	0.2260	0.3606	0.1245
-3.1	0.2389	0.3752	0.1324
-3.05	0.2524	0.3906	0.1409
-3	0.2664	0.4067	0.1499
-2.95	0.2810	0.4235	0.1594
-2.9	0.2962	0.4410	0.1695
-2.85	0.3120	0.4593	0.1801
-2.8	0.3284	0.4784	0.1914
-2.75	0.3454	0.4982	0.2033
-2.7	0.3629	0.5187	0.2159
-2.65	0.3810	0.5399	0.2292
-2.6	0.3996	0.5618	0.2432
-2.55	0.4187	0.5842	0.2580
-2.5	0.4383	0.6070	0.2735
-2.45	0.4584	0.6302	0.2898
-2.4	0.4788	0.6536	0.3069
-2.35	0.4996	0.6771	0.3248
-2.3	0.5207	0.7004	0.3435
-2.25	0.5421	0.7233	0.3630
-2.2	0.5636	0.7457	0.3834
-2.15	0.5853	0.7673	0.4046
-2.1	0.6070	0.7880	0.4265
-2.05	0.6286	0.8075	0.4492
-2	0.6501	0.8256	0.4726
-1.95	0.6714	0.8423	0.4966
-1.9	0.6924	0.8573	0.5213
-1.85	0.7130	0.8706	0.5464
-1.8	0.7330	0.8823	0.5718
-1.75	0.7523	0.8922	0.5976

-1.7	0.7710	0.9006	0.6234
-1.65	0.7888	0.9074	0.6492
-1.6	0.8056	0.9129	0.6747
-1.55	0.8214	0.9173	0.6998
-1.5	0.8361	0.9208	0.7242
-1.45	0.8496	0.9235	0.7478
-1.4	0.8619	0.9258	0.7702
-1.35	0.8731	0.9279	0.7913
-1.3	0.8832	0.9299	0.8109
-1.25	0.8921	0.9319	0.8287
-1.2	0.9001	0.9342	0.8446
-1.15	0.9071	0.9369	0.8585
-1.1	0.9134	0.9399	0.8702
-1.05	0.9191	0.9433	0.8799
-1	0.9244	0.9470	0.8874
-0.95	0.9293	0.9511	0.8929
-0.9	0.9341	0.9555	0.8965
-0.85	0.9388	0.9600	0.8984
-0.8	0.9435	0.9646	0.8988
-0.75	0.9483	0.9692	0.8980
-0.7	0.9532	0.9737	0.8964
-0.65	0.9583	0.9780	0.8942
-0.6	0.9633	0.9820	0.8918
-0.55	0.9684	0.9857	0.8895
-0.5	0.9735	0.9891	0.8875
-0.45	0.9783	0.9921	0.8862
-0.4	0.9829	0.9947	0.8857
-0.35	0.9871	0.9968	0.8861
-0.3	0.9908	0.9984	0.8877
-0.25	0.9940	0.9995	0.8905
-0.2	0.9965	1.0000	0.8943
-0.15	0.9984	0.9999	0.8993
-0.1	0.9996	0.9991	0.9053
-0.05	1.0000	0.9976	0.9121
0	0.9997	0.9954	0.9197
0.05	0.9988	0.9923	0.9278
0.1	0.9972	0.9885	0.9361
0.15	0.9950	0.9839	0.9446
0.2	0.9922	0.9786	0.9530

0.25	0.9890	0.9727	0.9611
0.3	0.9854	0.9662	0.9687
0.35	0.9816	0.9594	0.9757
0.4	0.9776	0.9523	0.9820
0.45	0.9736	0.9452	0.9875
0.5	0.9696	0.9382	0.9920
0.55	0.9658	0.9314	0.9955
0.6	0.9623	0.9250	0.9980
0.65	0.9590	0.9192	0.9995
0.7	0.9560	0.9140	1.0000
0.75	0.9534	0.9095	0.9995
0.8	0.9510	0.9057	0.9981
0.85	0.9489	0.9025	0.9958
0.9	0.9469	0.8999	0.9927
0.95	0.9449	0.8978	0.9889
1	0.9427	0.8960	0.9845
1.05	0.9402	0.8943	0.9798
1.1	0.9373	0.8924	0.9747
1.15	0.9336	0.8902	0.9694
1.2	0.9291	0.8874	0.9642
1.25	0.9235	0.8836	0.9591
1.3	0.9169	0.8788	0.9542
1.35	0.9089	0.8726	0.9496
1.4	0.8995	0.8649	0.9454
1.45	0.8888	0.8555	0.9415
1.5	0.8766	0.8444	0.9378
1.55	0.8631	0.8314	0.9343
1.6	0.8481	0.8167	0.9308
1.65	0.8319	0.8002	0.9271
1.7	0.8144	0.7821	0.9229
1.75	0.7959	0.7624	0.9181
1.8	0.7764	0.7413	0.9123
1.85	0.7560	0.7190	0.9054
1.9	0.7349	0.6957	0.8970
1.95	0.7132	0.6716	0.8870
2	0.6911	0.6469	0.8753
2.05	0.6686	0.6217	0.8617
2.1	0.6460	0.5964	0.8462
2.15	0.6232	0.5710	0.8288

2.2	0.6006	0.5457	0.8096
2.25	0.5780	0.5207	0.7887
2.3	0.5557	0.4961	0.7663
2.35	0.5337	0.4720	0.7425
2.4	0.5120	0.4486	0.7177
2.45	0.4908	0.4258	0.6919
2.5	0.4701	0.4038	0.6655
2.55	0.4499	0.3826	0.6387
2.6	0.4302	0.3622	0.6117
2.65	0.4112	0.3427	0.5848
2.7	0.3927	0.3241	0.5580
2.75	0.3748	0.3063	0.5317
2.8	0.3575	0.2894	0.5058
2.85	0.3409	0.2733	0.4806
2.9	0.3248	0.2581	0.4561
2.95	0.3094	0.2437	0.4324
3	0.2945	0.2300	0.4096
3.05	0.2802	0.2172	0.3877
3.1	0.2664	0.2050	0.3668
3.15	0.2532	0.1935	0.3467
3.2	0.2406	0.1827	0.3277
3.25	0.2284	0.1725	0.3095
3.3	0.2168	0.1629	0.2923
3.35	0.2056	0.1539	0.2760
3.4	0.1950	0.1454	0.2606
3.45	0.1847	0.1374	0.2461
3.5	0.1750	0.1299	0.2323
3.55	0.1657	0.1228	0.2193
3.6	0.1568	0.1162	0.2070
3.65	0.1483	0.1099	0.1955
3.7	0.1402	0.1040	0.1846
3.75	0.1325	0.0985	0.1743
3.8	0.1252	0.0933	0.1647
3.85	0.1182	0.0884	0.1555
3.9	0.1116	0.0837	0.1469
3.95	0.1053	0.0794	0.1388
4	0.0994	0.0753	0.1312

Appendix D. Parameter Estimates from Informative Real Data Analyses

Goldberg's Big Five Measure

Table D1. Item Parameter Estimates for Unidimensional Graded Response Models for the Five Scales of Goldberg's Big Five Measure for a Sample from the Open Source Psychometrics Project (*N* = 10,000). Note: *D* = 1.7.

Scale	Item	Discrimination	Boundary 1	Boundary 2	Boundary 3	Boundary 4
Openness	1	0.7495	-0.9209	0.5801	1.9137	3.1054
	2	0.8741	-0.3351	0.9272	1.8901	3.0209
	3	0.8677	-0.2478	0.9684	1.9885	3.1339
	4	0.7922	-0.2238	1.0517	2.1916	3.1819
	5	0.9275	-0.8475	0.7143	2.3234	3.6612
	6	0.927	-0.0573	1.1899	1.8829	2.7858
	7	0.6164	-0.634	1.3947	2.975	4.7837
	8	0.7013	-1.6425	-0.066	1.0512	2.3191
	9	0.4814	-0.0306	2.2449	3.6949	5.6628
	10	1.3049	-0.3082	0.7857	1.868	2.9049
Conscientiousness	1	0.9939	-1.4788	0.0484	1.0125	2.2159
	2	0.9291	-1.2933	-0.3944	0.2686	1.4054
	3	0.4992	-0.6418	1.598	3.1376	5.2743
	4	0.9795	-1.05	0.101	0.9348	2.0401
	5	1.0276	-1.9525	-0.9012	-0.0672	1.0167
	6	1.1454	-0.9901	-0.0644	0.4668	1.3672
	7	0.7969	-0.9217	0.6702	1.9373	3.0872
	8	0.7114	-1.0483	0.2408	1.7108	3.1797
	9	0.8515	-1.46	-0.0227	0.8882	2.0225
	10	0.5698	-1.5419	0.4258	2.4227	4.2597
Extraversion	1	1.1246	-2.0561	-0.8667	0.1293	0.8778
	2	1.2495	-1.1056	-0.2095	0.6071	1.523
	3	1.1176	-1.2273	-0.1202	0.7814	1.9453
	4	1.485	-1.5942	-0.6085	0.228	1.1672
	5	1.4692	-1.0557	-0.0517	0.6817	1.5588
	6	0.9152	-0.8939	0.4181	1.2406	2.3914
	7	1.3299	-1.3377	-0.5361	0.0496	0.8752
	8	0.8937	-2.3183	-1.1331	-0.2531	0.8941
	9	1.0051	-1.5472	-0.441	0.2903	1.2775
	10	1.2851	-1.8761	-0.9457	-0.272	0.6125
Agreeableness	1	0.9532	-0.0849	0.9396	1.4655	2.3062
	2	0.9053	-0.5679	0.7863	1.7853	2.8795

	3	0.4684	-0.6991	0.7374	1.9431	3.8972
	4	1.8131	-0.2841	0.7757	1.3898	2.1808
	5	1.1955	-0.6848	0.5728	1.2702	2.1995
	6	0.9197	-0.6497	0.6283	1.4991	2.4646
	7	1.1232	-0.6073	0.6594	1.4788	2.5678
	8	0.9434	-0.8965	0.683	1.6556	2.9207
	9	1.371	-0.5415	0.6176	1.2626	2.1009
	10	0.628	-1.356	0.37	2.1432	3.722
Neuroticism	1	1.1905	-0.9297	0.0364	0.7308	1.6272
	2	0.7908	-2.2143	-0.7882	0.2834	1.7042
	3	0.9622	-0.4876	0.8191	1.5654	2.6348
	4	0.6409	-1.4602	0.0585	1.1219	2.459
	5	0.7082	-2.3684	-0.8943	0.0947	1.4591
	6	1.3902	-1.4998	-0.5251	0.1294	1.0966
	7	1.2738	-1.3894	-0.4049	0.3639	1.4744
	8	1.436	-1.5774	-0.7299	-0.0956	0.811
	9	1.1419	-1.4042	-0.2538	0.445	1.4994
	10	0.9915	-1.6483	-0.5956	0.0964	1.1458

Multidimensional Personality Questionnaire

Table D2. Item Parameter Estimates for Unidimensional Two-parameter Logistic Item Response Models for 11 Primary Scales of the Multidimensional Personality Questionnaire for a Sample from the Minnesota Center for Twin and Family Research (N = 5,432). Note: D = 1.7.

Scale	Item	Discrimination	Difficulty
Absorption	1	0.5849	0.2985
	2	0.5154	-0.3368
	3	0.5515	-0.2893
	4	0.6448	0.2589
	5	0.6575	-1.7871
	6	0.5177	1.3096
	7	0.7272	0.0481
	8	0.631	-0.7573
	9	0.6941	-1.4797
	10	0.7011	-1.1003
	11	0.6915	-0.3636
	12	0.6467	-0.547
	13	0.7081	-1.8692
	14	0.6663	0.062

	15	0.735	0.9997
	16	0.9885	-0.3375
	17	0.7541	0.5072
	18	0.677	0.0187
	19	0.7361	0.7633
	20	0.5327	1.3067
	21	0.6377	-0.5278
	22	0.5361	0.3928
	23	0.6346	0.6423
	24	0.7528	-0.6755
	25	0.7334	-0.9391
	26	0.6492	-0.3977
	27	1.041	-0.5049
	28	0.5075	-0.0285
	29	0.7093	-0.8485
	30	0.7253	-0.2147
	31	0.9306	-0.6705
	32	0.8713	0.4497
	33	0.6894	-0.8721
	34	0.756	0.9655
Achievement	1	0.6912	0.575
	2	0.3556	0.1919
	3	0.7401	1.285
	4	0.7508	-0.4212
	5	0.3495	-0.5477
	6	1.1533	-0.0073
	7	0.9706	0.3354
	8	0.9493	1.1798
	9	0.753	0.7081
	10	0.9724	1.0382
	11	0.4302	0.1793
	12	0.5941	-0.1555
	13	0.7875	-0.4668
	14	0.8301	0.5193
	15	0.9855	-0.2451
	16	0.7786	0.1604
	17	0.5906	1.0683
	18	0.7817	-1.0232
	19	0.5805	1.3063

	20	0.8802	1.4095
	21	1.0349	0.512
Aggression	1	0.7067	-1.2637
	2	1.3142	-0.7589
	3	0.6349	-1.3485
	4	0.3831	-2.2058
	5	0.686	-1.6429
	6	0.7626	-1.5832
	7	1.1239	-2.1511
	8	0.7676	0.4076
	9	0.9841	-0.4405
	10	0.5818	-0.5886
	11	0.7974	-1.2839
	12	0.907	-1.952
	13	0.7683	-1.9421
	14	0.5955	-1.4727
	15	0.7976	-2.2442
	16	0.9389	-1.6555
	17	0.6544	-1.5561
	18	1.5463	-0.7055
	19	0.5998	-1.3612
	20	0.7962	-1.8174
Alienation	1	0.8627	-1.8683
	2	0.9545	-1.5745
	3	0.4669	-1.7501
	4	1.1164	-1.0327
	5	1.3413	-1.5699
	6	0.6594	-2.0651
	7	0.6448	-1.6853
	8	0.7312	-2.6294
	9	0.7209	-1.7355
	10	1.0692	-2.1513
	11	0.7484	-2.6913
	12	0.7448	-1.5678
	13	0.7384	-1.3996
	14	0.7462	-1.2536
	15	0.8119	-2.0967
	16	1.0873	-2.0536
	17	0.7738	-0.8896

	18	0.9248	-2.1291
	19	0.9267	-1.954
	20	0.7996	-1.3215
Control	1	0.5625	1.132
	2	0.4061	-0.14
	3	1.0853	1.2399
	4	0.9142	0.6586
	5	1.2893	1.2643
	6	0.7752	0.1341
	7	0.5106	0.7628
	8	0.7482	1.6426
	9	0.6556	0.2076
	10	0.6597	0.5764
	11	0.5072	0.6187
	12	1.0119	2.0029
	13	1.1565	-0.0948
	14	1.2995	0.9715
	15	0.8274	-0.0464
	16	0.63	-0.0772
	17	0.6328	0.9804
	18	0.5894	-0.0205
	19	1.0249	1.2761
	20	0.6264	0.5426
	21	0.4942	1.7965
	22	0.5725	0.3943
	23	0.7551	2.2066
	24	0.6241	0.3229
Harm Avoidance	1	0.5156	0.0588
	2	0.6651	0.971
	3	0.9528	0.7925
	4	0.5008	1.2323
	5	0.6969	1.6877
	6	0.8416	1.2899
	7	1.0175	0.5343
	8	0.4845	1.9126
	9	0.757	0.9999
	10	0.7584	0.7437
	11	0.8116	1.6293
	12	0.7353	0.5992

	13	1.0553	1.3264
	14	0.8748	1.2263
	15	1.3322	0.2397
	16	0.8088	0.5788
	17	0.7494	0.5431
	18	0.4161	1.0713
	19	0.8365	0.5575
	20	0.6934	0.0718
	21	0.7981	1.3755
	22	0.8743	0.9753
	23	0.8806	1.1068
	24	0.5902	1.1615
	25	0.9756	2.1591
	26	0.7759	1.5708
	27	0.7413	1.3919
	28	0.5676	1.2933
Social Closeness	1	1.0235	0.2691
	2	0.7305	0.6712
	3	0.8666	-0.1106
	4	0.7392	-0.044
	5	0.926	1.4161
	6	1.3519	1.0763
	7	0.7133	0.4666
	8	0.7861	0.6639
	9	1.341	0.3537
	10	0.4501	2.5628
	11	0.6461	0.8631
	12	0.9441	1.6274
	13	0.9834	0.1057
	14	0.7467	1.3859
	15	0.6782	-0.3917
	16	0.6315	1.7002
	17	1.0934	0.6936
	18	1.0106	0.5154
	19	0.7539	-0.7035
	20	1.4259	0.4876
	21	0.7794	2.1477
	22	0.4697	2.1289
Social Potency	1	1.1615	0.1768

	2	0.7271	-0.201
	3	0.9553	-0.3623
	4	0.9299	-1.757
	5	0.9916	-0.7935
	6	0.6134	-1.3661
	7	0.9219	-0.2589
	8	0.6741	-1.5981
	9	1.6586	0.1403
	10	1.6716	0.0558
	11	0.8977	0.0177
	12	0.8822	-0.911
	13	0.7797	-1.1803
	14	0.9362	-0.7056
	15	0.6741	0.2311
	16	0.8543	-0.6592
	17	0.7769	-0.7963
	18	0.6148	-0.9069
	19	0.8541	-0.6085
	20	0.8774	-0.8314
	21	1.5143	-0.0119
	22	0.8934	-0.9445
	23	1.017	-0.1513
	24	0.8677	0.5866
	25	0.8634	0.0243
	26	1.1337	-1.3677
Stress Reaction	1	1.1328	0.4633
	2	0.8099	0.5689
	3	0.7546	-0.5935
	4	0.7818	0.5446
	5	1.1045	-0.6527
	6	0.9302	-0.7761
	7	0.584	-0.7736
	8	0.7122	-1.0197
	9	1.0256	0.0851
	10	0.948	-0.2229
	11	0.9622	-0.5897
	12	0.8759	-0.3967
	13	0.6473	-0.092
	14	1.5185	-0.8564

	15	0.9737	-0.5162
	16	0.961	-0.1621
	17	0.9399	-0.6339
	18	1.2128	-0.6829
	19	0.816	0.509
	20	0.53	0.2409
	21	1.0668	-0.5064
	22	1.0037	-0.5091
	23	0.914	-0.6912
	24	1.032	-0.0314
	25	0.8221	0.0484
	26	0.9726	-0.509
Traditionalism	1	0.8707	0.5799
	2	0.6291	1.6352
	3	0.445	0.4593
	4	0.5548	0.8019
	5	0.5774	-1.1161
	6	0.8334	0.6049
	7	0.4664	0.8377
	8	0.935	0.6022
	9	0.6934	0.208
	10	0.688	1.9548
	11	0.8149	1.2807
	12	0.5133	1.0152
	13	0.6905	2.7249
	14	0.543	0.769
	15	1.4898	1.087
	16	0.818	0.3921
	17	0.4379	1.0445
	18	0.5263	1.8006
	19	0.4253	0.931
	20	0.5818	1.042
	21	0.6844	1.4979
	22	0.4146	1.5693
	23	0.9084	0.3249
	24	0.7933	0.5074
	25	0.5786	1.3295
	26	0.8726	1.228
	27	0.4698	2.2586

Wellbeing	1	0.7934	0.6513
	2	0.9125	1.1522
	3	0.7525	1.9427
	4	0.8193	1.1559
	5	0.9114	0.2329
	6	0.8729	0.553
	7	0.8357	1.5727
	8	0.9677	1.021
	9	0.9345	0.9739
	10	1.6882	1.6864
	11	0.9769	1.124
	12	0.5851	1.2271
	13	1.1533	1.2533
	14	1.1435	1.4532
	15	1.3433	1.2897
	16	0.6433	0.7292
	17	0.9024	-0.516
	18	0.7416	1.4222
	19	0.66	1.5132
	20	0.8599	1.5115
	21	0.9195	0.6737
	22	1.1935	1.691
	23	1.7376	1.0731
	24	1.2275	0.9683

Stability Facets of the Big Five

Table D3. Item Parameter Estimates for Unidimensional Graded Response Models for 18 Facet Scales from the International Personality Item Pool for Stability Facets of the Big Five for a Sample from Johnson, 2014 (N = 10,000). Note: D = 1.7.

Scale	Item	Discrimination	Boundary 1	Boundary 2	Boundary 3	Boundary 4
Trust	1	1.3647	-1.7904	-0.599	0.0243	1.4801
	2	1.411	-2.1509	-0.9746	-0.1714	1.3628
	3	1.6051	-1.7241	-0.5422	0.1393	1.6552
	4	0.9896	-2.0232	-0.7945	0.1832	1.9215
	5	1.06	-2.3712	-1.236	-0.4763	1.1475
	6	0.5908	-3.351	-1.5705	-0.4518	1.7207
	7	1.8307	-1.6337	-0.4498	0.0809	1.2478
	8	0.9465	-1.7788	-0.1285	0.6109	1.8962
	9	0.8275	-2.1263	-0.1994	0.7447	2.3198

	10	0.9589	-2.7672	-1.6754	-0.8706	0.2276
Morality	1	0.5166	-3.4159	-1.8479	-0.7738	0.425
	2	0.5563	-3.6867	-1.7086	-0.5255	2.0042
	3	0.6583	-2.8503	-1.1358	-0.0846	1.1586
	4	1.3222	-2.2715	-1.1242	-0.3208	0.5417
	5	0.7376	-1.4689	0.2498	1.1945	2.5663
	6	1.2815	-2.7256	-1.7037	-1.1268	-0.331
	7	0.744	-3.407	-1.4786	-0.5381	0.8914
	8	0.7758	-3.511	-1.7769	-1.0713	0.2907
	9	1.9119	-2.4558	-1.4123	-0.9347	-0.0322
	10	0.9287	-3.9519	-2.4768	-1.3337	0.0579
Altruism	1	0.8901	-3.5853	-2.0427	-0.9795	0.6954
	2	0.8361	-3.5625	-2.1203	-0.9452	1.1147
	3	1.2519	-2.9138	-1.8118	-0.8956	0.5076
	4	1.4258	-2.8239	-1.8107	-1.1488	0.4097
	5	0.6876	-3.0406	-1.3061	-0.1919	1.7496
	6	0.7426	-3.5384	-1.7025	-0.869	0.5711
	7	1.025	-3.089	-1.822	-0.9238	0.4294
	8	0.7158	-3.6438	-1.9921	-0.9451	0.6687
	9	0.8956	-3.7448	-2.3697	-1.6006	0.0233
	10	1.1035	-3.2107	-1.9764	-1.3222	0.3529
Cooperation	1	0.4108	-4.3823	-1.7389	-0.2993	2.894
	2	0.4779	-3.2069	-1.2565	-0.1077	1.4126
	3	0.5088	-4.6789	-2.5872	-1.4601	0.596
	4	0.8978	-1.0653	0.1394	0.8906	1.889
	5	0.8089	-2.0753	-0.453	0.7333	2.278
	6	0.8583	-1.7271	-0.6199	-0.0571	0.7448
	7	1.0813	-2.2594	-1.0459	-0.53	0.3501
	8	1.3964	-2.1717	-1.0973	-0.5805	0.2886
	9	1.1299	-1.9722	-0.7572	-0.2432	0.7688
	10	0.7564	-2.0074	-0.4187	0.0892	1.3961
Modesty	1	0.2851	-3.4745	-0.5605	1.3968	4.1159
	2	0.4264	-3.1983	-0.6459	0.9766	3.1116
	3	0.3789	-2.726	-0.5626	0.2622	2.9302
	4	0.2773	-6.9665	-2.6101	-0.0209	3.6452
	5	0.6719	-2.773	-0.8772	0.0705	1.1614
	6	2.887	-1.1617	-0.0043	0.614	1.3977
	7	3.1362	-1.1139	-0.0191	0.5286	1.3
	8	0.3503	-2.0703	1.4734	3.6417	6.5642

	9	0.3506	-6.539	-3.1852	-1.1965	1.307
	10	0.4206	-4.2293	-1.7506	-0.5644	1.5424
Sympathy	1	1.1336	-2.1718	-1.139	-0.3117	1.0606
	2	1.2409	-2.5718	-1.5869	-0.9239	0.5216
	3	0.464	-4.2912	-2.2282	-0.8728	1.2433
	4	0.7396	-2.2773	-0.9819	0.0231	1.841
	5	0.8025	-3.1547	-1.7358	-1.0068	0.6891
	6	0.8047	-3.5058	-2.0688	-1.08	0.2442
	7	0.4077	-2.4834	-0.3346	0.9265	2.7026
	8	0.8148	-2.9966	-1.2964	-0.2443	1.401
	9	0.6301	-2.5704	-0.335	0.8076	2.8045
	10	0.5466	-2.5176	-0.5961	0.5324	2.3428
Self-Efficacy	1	0.9032	-3.5572	-1.9828	-1.0405	0.8227
	2	1.0993	-3.3136	-2.1828	-1.0908	0.652
	3	1.204	-3.2956	-1.7509	-0.6493	1.4443
	4	0.7803	-3.1003	-1.5049	-0.4601	1.4495
	5	1.0249	-3.9898	-2.4536	-1.4036	0.8998
	6	1.3304	-3.0375	-1.7642	-0.9721	0.7947
	7	0.7063	-3.4402	-1.295	-0.1211	1.9859
	8	0.8429	-3.0545	-1.5841	-0.7901	0.7619
	9	0.8135	-3.4555	-1.9112	-1.0767	0.7232
	10	0.5724	-4.1398	-2.1242	-1.3538	0.7038
Orderliness	1	1.2749	-2.5084	-1.4477	-0.5865	0.7254
	2	1.3662	-1.5566	-0.6703	-0.0536	1.15
	3	0.585	-3.7533	-1.8654	-0.6647	1.206
	4	1.1531	-1.9736	-0.8184	-0.0296	1.233
	5	0.7621	-3.0094	-1.294	-0.2428	1.9949
	6	1.0379	-1.4509	-0.2942	0.1504	1.1748
	7	1.3024	-0.7544	0.1505	0.5142	1.3063
	8	1.0742	-1.4838	-0.1571	0.2955	1.2562
	9	0.8847	-1.7754	-0.431	0.0106	1.3968
	10	1.076	-1.8976	-0.527	-0.0373	1.2656
Dutifulness	1	0.7745	-3.4135	-1.9893	-1.0193	1.0188
	2	1.0264	-3.8934	-2.4546	-1.4919	0.4076
	3	0.4999	-3.1871	-1.7327	-0.6762	0.9276
	4	1.1161	-2.9661	-1.8531	-1.0941	0.6789
	5	0.7249	-3.6474	-2.1889	-1.233	0.8008
	6	0.8842	-2.4457	-0.9341	-0.2086	1.0854
	7	1.1682	-3.2718	-1.9464	-1.2301	0.1413

	8	0.7094	-3.9286	-1.9661	-1.0844	0.376
	9	0.8107	-3.9644	-2.4306	-1.356	0.1616
	10	0.8208	-3.6772	-1.9756	-0.9701	0.5042
Achievement-Striving	1	0.7017	-3.6419	-1.5885	-0.2583	1.6089
	2	1.3688	-2.7621	-1.7269	-0.9574	0.4004
	3	0.76	-3.2776	-1.5085	-0.355	1.6608
	4	0.9016	-3.1037	-1.3865	-0.2269	1.3093
	5	1.2825	-2.5895	-1.3451	-0.4491	0.9733
	6	0.7346	-3.4298	-2.0486	-1.0949	0.5385
	7	0.5234	-4.1768	-2.4442	-1.0503	1.5239
	8	0.7793	-2.6348	-1.3674	-0.7301	0.5094
	9	1.3469	-1.861	-0.7819	-0.352	0.669
	10	1.3157	-2.4238	-1.3937	-0.9143	0.3699
Self-Discipline	1	1.0268	-1.2366	0.0886	0.8067	2.1431
	2	0.7689	-2.8152	-0.9318	0.1844	2.3409
	3	1.4525	-1.4113	-0.1903	0.4472	1.7228
	4	1.161	-1.8223	-0.4606	0.262	1.5312
	5	0.9279	-3.171	-1.4387	-0.4478	1.7623
	6	1.5677	-1.6136	-0.4602	0.0869	1.2159
	7	1.0958	-1.6817	-0.2938	0.359	1.5442
	8	1.3352	-1.6272	-0.2879	0.2516	1.3996
	9	1.6912	-1.6726	-0.4386	0.0493	1.166
	10	0.8332	-2.036	-0.2405	0.4839	2.1048
Cautiousness	1	0.4472	-5.3337	-2.445	-0.6107	2.3729
	2	0.5395	-3.4514	-1.4699	-0.3461	2.0012
	3	0.3954	-4.7028	-1.6708	0.1603	3.37
	4	1.8684	-1.2845	-0.2598	0.2705	1.1151
	5	1.4465	-1.9268	-0.7328	-0.0413	0.9843
	6	0.8839	-1.6857	-0.1407	0.8292	2.0289
	7	1.6674	-1.5421	-0.286	0.2983	1.4787
	8	0.7795	-1.4859	-0.1226	0.5904	1.7165
	9	2.0439	-1.5056	-0.4357	0.0119	0.9884
	10	0.668	-1.0277	0.7017	1.4253	3.0031
Anxiety	1	1.3044	-2.2165	-1.1901	-0.6801	0.7728
	2	1.0861	-1.4927	-0.4236	0.2184	1.3838
	3	1.0184	-1.0347	0.1964	0.8699	2.1062
	4	1.5381	-1.3679	-0.4767	-0.0143	0.9316
	5	1.2143	-2.1942	-0.9642	-0.3611	1.1177
	6	1.1319	-1.6152	-0.4375	0.0509	1.308

	7	0.9025	-1.7062	0.1764	0.8518	2.3565
	8	0.8367	-2.3046	-0.4889	0.2963	2.1785
	9	1.0196	-1.9388	-0.5991	-0.2418	1.1796
	10	0.64	-1.5694	0.8051	1.5984	3.3985
Anger	1	1.7815	-1.2544	-0.3471	0.2346	1.3195
	2	1.6868	-1.6932	-0.7373	-0.1804	1.0136
	3	1.5055	-1.3645	-0.3299	0.2475	1.3218
	4	0.8756	-0.9028	0.4075	1.3111	2.6155
	5	1.823	-1.0041	-0.1736	0.2613	1.2553
	6	1.7424	-1.5657	-0.559	-0.1039	1.0823
	7	1.6878	-1.3228	-0.2302	0.2194	1.3852
	8	1.4732	-1.5712	-0.3715	0.0475	1.2844
	9	1.0003	-1.1546	0.6634	1.3277	2.7014
	10	0.7031	-2.5261	-0.6954	0.1187	1.9973
Depression	1	1.4153	-1.5159	-0.5315	0.139	1.2947
	2	1.7709	-0.575	0.173	0.8031	1.8065
	3	1.8269	-0.984	-0.1632	0.4058	1.3527
	4	1.7677	-0.6349	0.1007	0.5706	1.4479
	5	0.8583	-1.6641	-0.5293	0.122	1.3582
	6	1.1593	-0.5184	0.308	0.9644	2.1855
	7	0.7595	-1.4176	-0.2707	0.3448	1.7105
	8	1.158	-1.6229	-0.4552	0.0841	1.3655
	9	1.4779	-0.8755	0.466	0.9766	1.9612
	10	1.3756	-1.409	-0.0808	0.6544	1.7162
Self-Consciousness	1	1.0262	-1.2291	-0.139	0.6007	1.8602
	2	0.7543	-2.5519	-1.2013	-0.4263	1.0837
	3	1.3476	-1.2357	-0.2668	0.1925	1.2616
	4	1.0734	-1.3597	-0.1962	0.454	1.6313
	5	0.8845	-1.8916	-0.4898	0.1296	1.3932
	6	0.6001	-1.7061	-0.0042	0.933	2.9137
	7	0.9711	-1.6489	-0.4203	0.0025	1.1904
	8	0.8491	-1.9683	-0.2449	0.3735	1.7572
	9	0.9804	-1.835	-0.3993	0.0819	1.3832
	10	0.909	-0.4967	1.1156	1.5896	2.7499
Immoderation	1	0.9559	-1.5588	-0.6675	0.0388	1.185
	2	0.4223	-3.5855	-1.8188	-0.7301	1.4877
	3	0.4675	-3.1326	-1.0418	0.0978	2.2493
	4	0.9578	-1.2204	-0.4558	0.2569	1.4406
	5	0.7984	-2.7257	-1.6291	-0.736	0.622

	6	1.3488	-1.8381	-0.6738	0.0798	1.3739
	7	0.843	-2.2272	-0.6351	0.0463	1.7327
	8	1.179	-1.6936	-0.2058	0.3283	1.6665
	9	0.6002	-1.9883	-0.5959	-0.0711	1.7197
	10	0.6831	-3.7302	-1.848	-0.9157	1.0881
Vulnerability	1	1.2095	-0.888	0.1151	0.7557	1.8091
	2	1.0573	-1.5645	-0.2725	0.539	1.9066
	3	1.1132	-0.8201	0.2643	0.8935	2.0409
	4	0.6552	-2.2461	-0.6426	0.1355	1.8101
	5	0.9558	-1.7367	-0.6162	0.0097	1.1709
	6	1.4056	-0.9576	0.4675	1.0012	2.0612
	7	0.7207	-0.9755	1.1889	2.0783	3.5494
	8	1.1113	-0.9727	0.7977	1.3908	2.5388
	9	0.7773	-1.887	0.3954	1.3535	2.9541
	10	1.2991	-1.2435	0.2297	0.7767	1.8926

Appendix E. Bank Information Functions for Simulated CAT Banks

Figure E1 displays in red the observed, relative test information function for each scale of Goldberg's Big Five Measure for the informative real sample. The simulated, relative bank information function for each scale is shown in purple; these bank functions correspond to realistic item discriminations and middle-peaked information functions. In addition, Figures E2, E3, and E4 display similar plots for each scale of the Multidimensional Personality Questionnaire. Finally, Figures E5, E6, and E7 show similar plots for each of the 18 facets associated with the metatrait of Stability. Note that relative information functions are used to show that measurement precision for simulated banks matched observed precision for real scales across the latent trait continuum.

Figure E1.

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of Goldberg's Big Five Measure – Realistic Locations

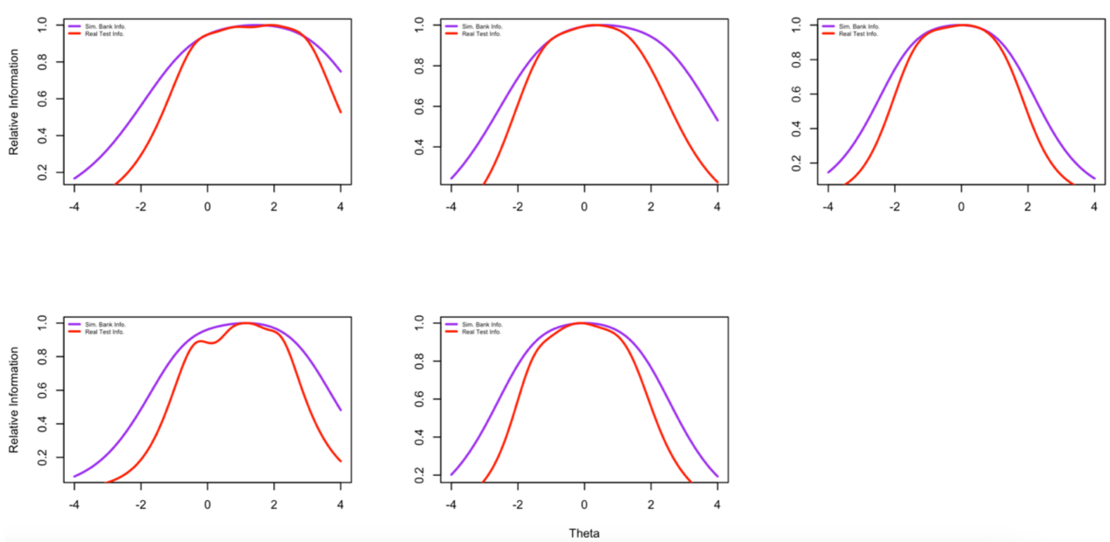


Figure E2

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of the Multidimensional Personality Questionnaire – Scales 1 to 6

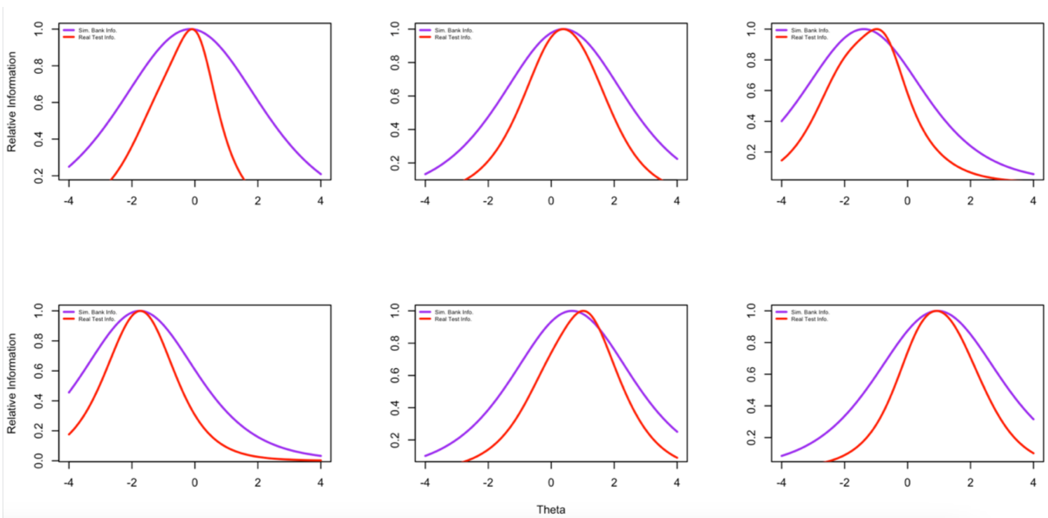


Figure E3

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of the Multidimensional Personality Questionnaire – Scales 7 to 11

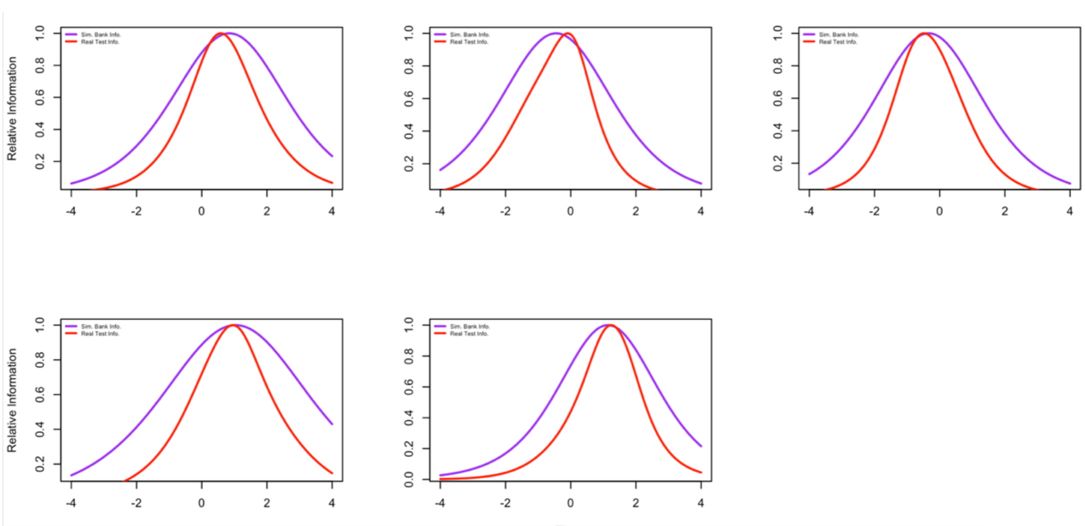


Figure E4

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of the Big Five Stability Facets – Scales 1 to 6

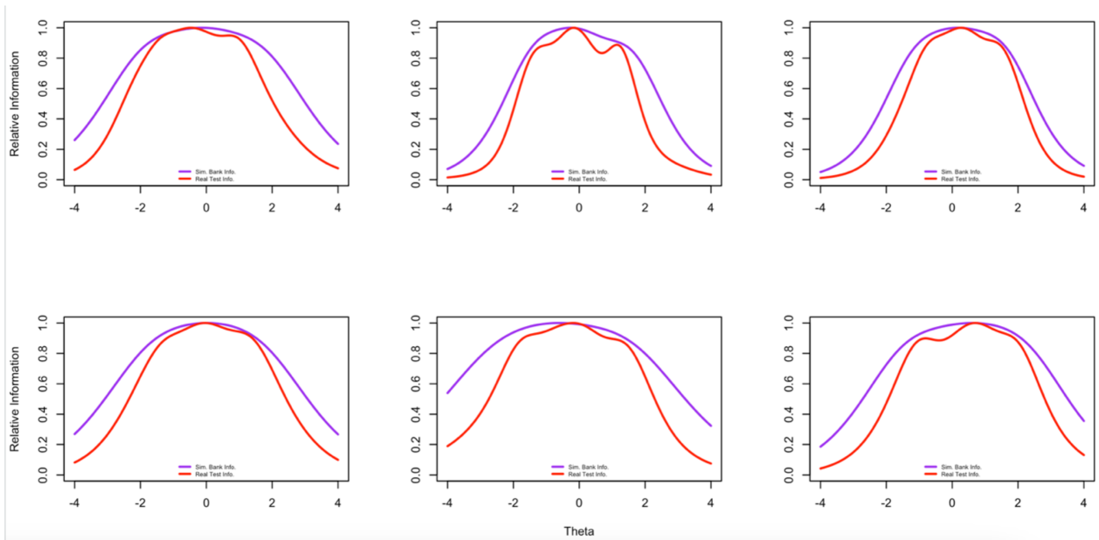


Figure E5

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of the Big Five Stability Facets – Scales 7 to 12

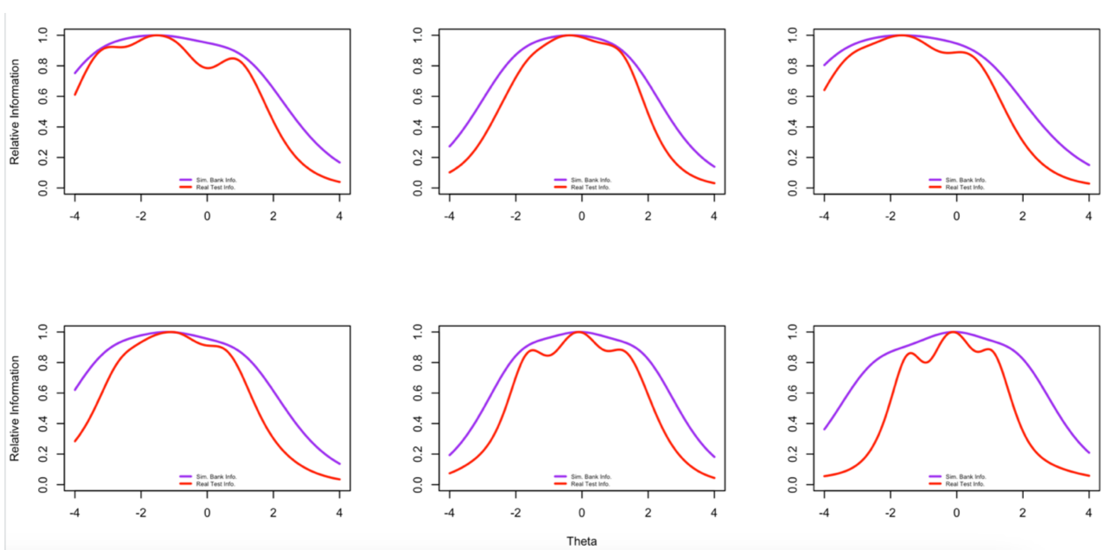


Figure E6

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of the Big Five Stability Facets – Scales 13 to 18

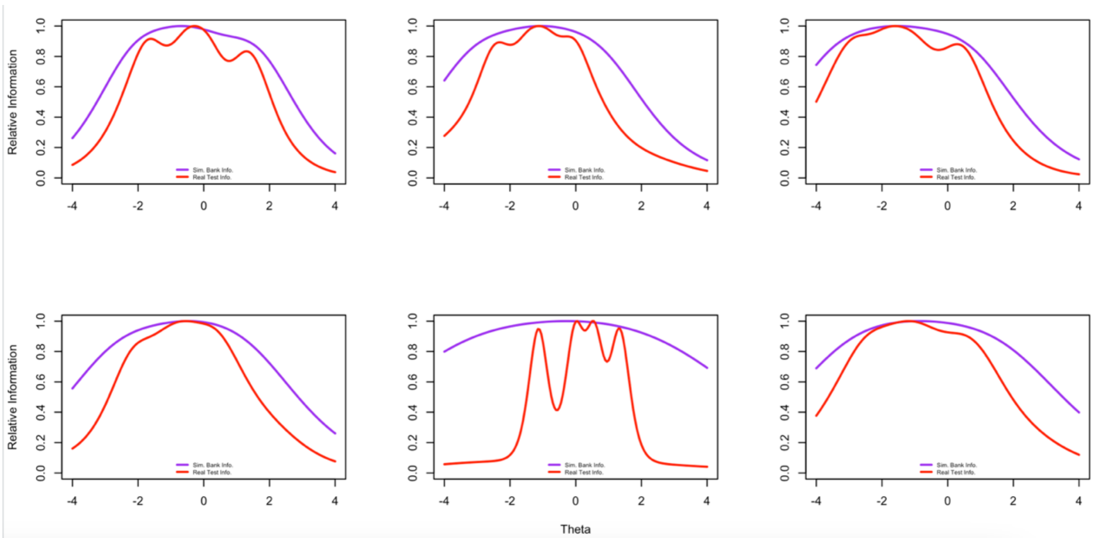
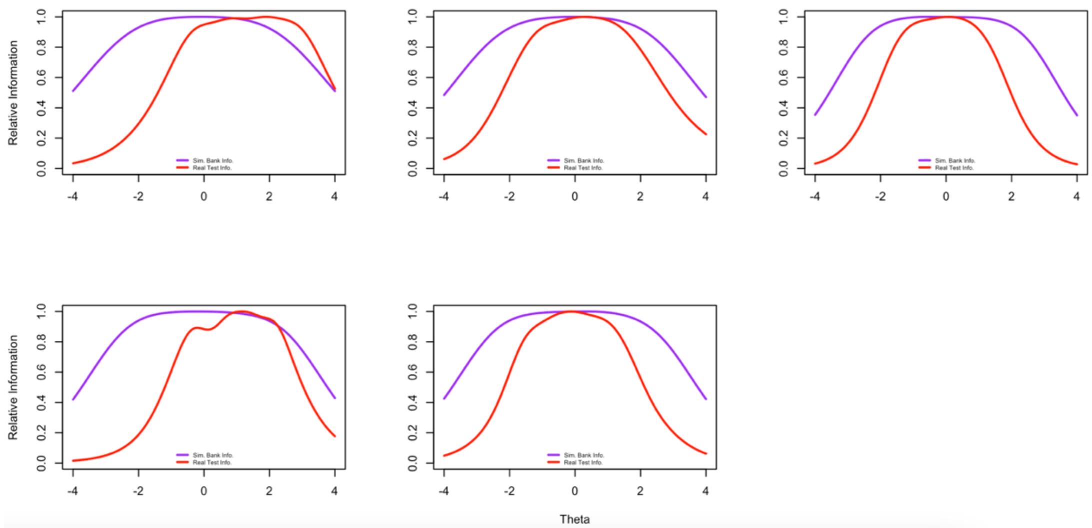


Figure E7

Observed Relative Test Information Functions and Simulated Relative Bank Information Functions for Each Scale of Goldberg's Big Five Measure – Flatter Locations



Appendix F. Simulation Results

Table F1. Classical Effect Sizes (η^2) from an Eight-way ANOVA on TPR for each of Three AMC Hypothesis Tests.

Source of Variation	η^2 for LR	η^2 for Z	η^2 for Score
Main Effects			
Number of Scales	0.081	0.121	0.069
Magnitude	0.467	0.39	0.474
Scale Location	0.012	0.001	0
Number of Differences	0.031	0.025	0.027
Discrimination	0.056	0.056	0.077
Boundaries	0	0	0.002
Type of CAT	0	0.002	0.003
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	0.119	0.08	0.075
Two-way Interactions			
Magnitude x Number of Scales	0.016	0.051	0.017
Scale Location x Number of Scales	0.005	0.008	0.007
Number of Scales x Number of Differences	0.003	0.006	0.003
Discrimination x Number of Scales	0.002	0.007	0.006
Boundaries x Number of Scales	0	0	0
Type of CAT x Number of Scales	0.011	0.002	0.002
Magnitude x Scale Location	0.003	0.006	0.006
Magnitude x Number of Differences	0.005	0.007	0.006
Magnitude x Discrimination	0.014	0.026	0.029
Boundaries x Magnitude	0	0	0.001
Magnitude x Type of CAT	0.014	0.012	0.016
Scale Location x Number of Differences	0.002	0.002	0.001
Discrimination x Scale Location	0	0	0
Boundaries x Scale Location	0.001	0	0
Type of CAT x Scale Location	0.001	0	0
Discrimination x Number of Differences	0	0.001	0.001
Boundaries x Number of Differences	0	0	0
Type of CAT x Number of Differences	0	0	0
Boundaries x Discrimination	0	0	0
Discrimination x Type of CAT	0.033	0.034	0.031
Boundaries x Type of CAT	0	0	0
Precision x Number of Scales	0.005	0.016	0.015

Precision x Magnitude	0.021	0.031	0.022
Precision x Scale Location	0.001	0.002	0.003
Precision x Number of Differences	0	0.001	0
Precision x Discrimination	0.004	0.006	0.004
Precision x Boundaries	0	0.001	0.001
Three-way Interactions			
Number of Scales x Magnitude x Scale Location	0.004	0.008	0.006
Number of Scales x Magnitude x Number of Differences	0.004	0.006	0.005
Number of Scales x Magnitude x Discrimination	0.002	0.001	0.003
Number of Scales x Magnitude x Boundaries	0	0	0
Number of Scales x Magnitude x Type of CAT	0.001	0.002	0.001
Number of Scales x Scale Location x Number of Differences	0.002	0.004	0.003
Number of Scales x Scale Location x Discrimination	0.001	0.001	0.001
Number of Scales x Scale Location x Boundaries	0	0	0
Number of Scales x Scale Location x Type of CAT	0.001	0	0
Number of Scales x Number of Differences x Discrimination	0	0	0
Number of Scales x Number of Differences x Boundaries	0	0	0
Number of Scales x Number of Differences x Type of CAT	0	0	0
Number of Scales x Discrimination x Boundaries	0	0	0
Number of Scales x Discrimination x Type of CAT	0.001	0.005	0.002
Number of Scales x Boundaries x Type of CAT	0	0	0
Magnitude x Scale Location x Number of Differences	0.003	0.004	0.003
Magnitude x Scale Location x Discrimination	0	0	0
Magnitude x Scale Location x Boundaries	0	0	0
Magnitude x Scale Location x Type of CAT	0	0	0
Magnitude x Number of Differences x Discrimination	0.001	0	0.001
Magnitude x Number of Differences x Boundaries	0	0	0
Magnitude x Number of Differences x Type of CAT	0	0	0
Magnitude x Discrimination x Boundaries	0	0	0
Magnitude x Discrimination x Type of CAT	0.006	0.012	0.008
Magnitude x Boundaries x Type of CAT	0	0	0
Scale Location x Number of Differences x Discrimination	0	0	0
Scale Location x Number of Differences x Boundaries	0	0	0
Scale Location x Number of Differences x Type of CAT	0	0	0
Scale Location x Discrimination x Boundaries	0	0	0
Scale Location x Discrimination x Type of CAT	0	0	0
Scale Location x Boundaries x Type of CAT	0	0	0
Number of Differences x Discrimination x Boundaries	0	0	0

Number of Differences x Discrimination x Type of CAT	0	0	0
Number of Differences x Boundaries x Type of CAT	0	0	0
Discrimination x Boundaries x Type of CAT	0	0	0
Precision x Magnitude x Number of Scales	0.015	0.011	0.019
Number of Scales x Scale Location x Precision	0.001	0.003	0.002
Number of Scales x Number of Differences x Precision	0	0	0
Number of Scales x Discrimination x Precision	0.002	0.001	0.001
Number of Scales x Boundaries x Precision	0	0	0
Magnitude x Scale Location x Precision	0.001	0.002	0.001
Magnitude x Number of Differences x Precision	0.002	0.001	0.002
Magnitude x Discrimination x Precision	0.011	0.004	0.01
Magnitude x Boundaries x Precision	0	0	0
Scale Location x Number of Differences x Precision	0	0	0
Scale Location x Discrimination x Precision	0.001	0.001	0.001
Scale Location x Boundaries x Precision	0	0	0
Number of Differences x Discrimination x Precision	0	0	0
Number of Differences x Boundaries x Precision	0	0	0
Discrimination x Boundaries x Precision	0	0	0
Residuals	0.028	0.036	0.03

Table F2. Classical Effect Sizes (η^2) from a Six-way ANOVA on FPR for each of Three AMC Hypothesis Tests.

Source of Variation	η^2 for LR	η^2 for Z	η^2 for Score
Main Effects			
Number of Scales	0.037	0.313	0.265
Scale Location	0.027	0.122	0.135
Discrimination	0.002	0.053	0.073
Boundaries	0.002	0.007	0.01
Type of CAT	0.059	0.03	0.019
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	0.087	0.027	0.034
Two-way Interactions			
Scale Location x Number of Scales	0.001	0.153	0.145
Discrimination x Number of Scales	0.004	0.079	0.099
Boundaries x Number of Scales	0.004	0.003	0.002
Type of CAT x Number of Scales	0.013	0.005	0.008
Discrimination x Scale Location	0	0.045	0.05
Boundaries x Scale Location	0	0.001	0.003

Scale Location x Type of CAT	0.002	0.006	0.006
Boundaries x Discrimination	0.003	0	0.001
Discrimination x Type of CAT	0.02	0	0
Boundaries x Type of CAT	0.002	0.006	0.004
Precision x Number of Scales	0.016	0.006	0.009
Scale Location x Precision	0.007	0.008	0.009
Discrimination x Precision	0.005	0.001	0
Boundaries x Precision	0.007	0.004	0.005
Three-way Interactions			
Discrimination x Scale Location x Number of Scales	0.004	0.077	0.071
Number of Scales x Scale Location x Boundaries	0.002	0.001	0.001
Scale Location x Type of CAT x Number of Scales	0.081	0.003	0
Number of Scales x Discrimination x Boundaries	0.007	0.001	0
Number of Scales x Discrimination x Type of CAT	0.003	0	0.001
Number of Scales x Boundaries x Type of CAT	0.009	0.001	0
Scale Location x Discrimination x Boundaries	0.004	0	0
Scale Location x Discrimination x Type of CAT	0.003	0	0
Scale Location x Boundaries x Type of CAT	0.006	0.001	0
Discrimination x Boundaries x Type of CAT	0.002	0	0
Precision x Scale Location x Number of Scales	0.076	0.002	0
Precision x Discrimination x Number of Scales	0.033	0.004	0.005
Precision x Boundaries x Number of Scales	0.028	0.001	0.001
Precision x Discrimination x Scale Location	0.022	0.001	0.001
Precision x Boundaries x Scale Location	0.027	0	0
Discrimination x Boundaries x Precision	0.002	0	0
Residuals	0.395	0.039	0.04

Table F3. Classical Effect Sizes (η^2) from a Seven-way ANOVA on Mean Number of Items for Variable-length Computerized Adaptive Tests.

Source of Variation	η^2 for Mean Number of Items
Main Effects	
Number of Scales	0.237
Magnitude	0.002
Scale Location	0.025
Number of Differences	0.001
Discrimination	0.354
Bound	0.004

Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	0.257
Two-way Interactions	
Magnitude x Number of Scales	0
Scale Location x Number of Scales	0.011
Number of Scales x Number of Differences	0
Discrimination x Number of Scales	0
Boundaries x Number of Scales	0.001
Precision x Number of Scales	0.005
Magnitude x Scale Location	0
Magnitude x Number of Differences	0
Discrimination x Magnitude	0
Boundaries x Magnitude	0
Precision x Magnitude	0
Scale Location x Number of Differences	0.001
Discrimination x Scale Location	0.003
Boundaries x Scale Location	0.002
Precision x Scale Location	0
Discrimination x Number of Differences	0
Boundaries x Number of Differences	0
Precision x Number of Differences	0
Boundaries x Discrimination	0.002
Precision x Discrimination	0.065
Precision x Boundaries	0
Three-way Interactions	
Number of Scales x Magnitude x Scale Location	0
Number of Scales x Magnitude x Number of Differences	0
Number of Scales x Magnitude x Discrimination	0
Number of Scales x Magnitude x Boundaries	0
Number of Scales x Magnitude x Precision	0
Number of Scales x Scale Location x Number of Differences	0
Number of Scales x Scale Location x Discrimination	0.005
Number of Scales x Scale Location x Boundaries	0.001
Number of Scales x Scale Location x Precision	0.002
Number of Scales x Number of Differences x Discrimination	0
Number of Scales x Number of Differences x Boundaries	0
Number of Scales x Discrimination x Precision	0

Number of Scales by Discrimination x Boundaries	0.001
Precision x Discrimination x Number of Scales	0.014
Number of Scales x Boundaries x Precision	0
Magnitude x Scale Location x Number of Differences	0
Magnitude x Scale Location x Discrimination	0
Magnitude x Scale Location x Boundaries	0
Magnitude x Scale Location x Precision	0
Magnitude x Number of Differences x Discrimination	0
Magnitude x Number of Differences x Boundaries	0
Magnitude x Number of Differences x Precision	0
Magnitude x Discrimination x Boundaries	0
Magnitude x Discrimination x Precision	0
Magnitude x Boundaries x Precision	0
Scale Location x Number of Differences x Discrimination	0
Scale Location x Number of Differences x Boundaries	0
Scale Location x Number of Differences x Precision	0
Scale Location x Discrimination x Boundaries	0
Scale Location x Discrimination x Precision	0.001
Scale Location x Boundaries x Precision	0
Number of Differences x Discrimination x Boundaries	0
Number of Differences x Discrimination x Precision	0
Number of Differences x Boundaries x Precision	0
Discrimination x Boundaries x Precision	0
Residuals	0.002

Table F4. Sums of Squares and Degrees of Freedom from an Eight-way ANOVA on TPR for each of Three AMC Hypothesis Tests.

Source of Variation	SS for LR	SS for Z	SS for Score	DF
Main Effects				
Number of Scales	10.27	7.58	7.17	2
Magnitude	59.07	24.47	49.58	2
Scale Location	1.48	0.08	0.01	2
Number of Differences	3.98	1.58	2.8	1
Discrimination	7.07	3.53	8.02	1
Boundaries	0.05	0.01	0.16	1
Type of CAT	0.03	0.12	0.31	1
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	15.08	5	7.81	4

Two-way Interactions				
Magnitude x Number of Scales	1.97	3.19	1.82	4
Scale Location x Number of Scales	0.6	0.52	0.75	4
Number of Scales x Number of Differences	0.43	0.36	0.34	2
Discrimination x Number of Scales	0.24	0.46	0.64	2
Boundaries x Number of Scales	0.06	0.01	0.02	2
Type of CAT x Number of Scales	1.41	0.13	0.17	2
Magnitude x Scale Location	0.42	0.38	0.61	4
Magnitude x Number of Differences	0.64	0.45	0.67	2
Magnitude x Discrimination	1.8	1.6	3.03	2
Boundaries x Magnitude	0.01	0.02	0.09	2
Magnitude x Type of CAT	1.81	0.76	1.7	2
Scale Location x Number of Differences	0.26	0.09	0.09	2
Discrimination x Scale Location	0.06	0.01	0.03	2
Boundaries x Scale Location	0.07	0	0.02	2
Type of CAT x Scale Location	0.16	0	0	2
Discrimination x Number of Differences	0.02	0.05	0.08	1
Boundaries x Number of Differences	0	0	0	1
Type of CAT x Number of Differences	0	0.02	0.02	1
Boundaries x Discrimination	0.01	0	0	1
Discrimination x Type of CAT	4.12	2.1	3.24	1
Boundaries x Type of CAT	0.03	0	0	1
Precision x Number of Scales	0.62	1.03	1.54	8
Precision x Magnitude	2.66	1.92	2.29	8
Precision x Scale Location	0.1	0.11	0.3	8
Precision x Number of Differences	0.03	0.05	0.03	4
Precision x Discrimination	0.49	0.39	0.37	4
Precision x Boundaries	0.01	0.05	0.15	4
Three-way Interactions				
Number of Scales x Magnitude x Scale Location	0.52	0.48	0.64	8
Number of Scales x Magnitude x Number of Differences	0.55	0.39	0.56	4
Number of Scales x Magnitude x Discrimination	0.22	0.09	0.34	4
Number of Scales x Magnitude x Boundaries	0	0.01	0.02	4
Number of Scales x Magnitude x Type of CAT	0.15	0.12	0.13	4
Number of Scales x Scale Location x Number of Differences	0.25	0.27	0.34	4
Number of Scales x Scale Location x Discrimination	0.11	0.03	0.08	4

Number of Scales x Scale Location x Boundaries	0.02	0.03	0.01	4
Number of Scales x Scale Location x Type of CAT	0.12	0	0	4
Number of Scales x Number of Differences x Discrimination	0.02	0	0.01	2
Number of Scales x Number of Differences x Boundaries	0	0	0	2
Number of Scales x Number of Differences x Type of CAT	0.02	0	0	2
Number of Scales x Discrimination x Boundaries	0.03	0	0	2
Number of Scales x Discrimination x Type of CAT	0.19	0.32	0.2	2
Number of Scales x Boundaries x Type of CAT	0.01	0	0.01	2
Magnitude x Scale Location x Number of Differences	0.36	0.25	0.28	4
Magnitude x Scale Location x Discrimination	0.01	0.01	0.03	4
Magnitude x Scale Location x Boundaries	0	0	0.02	4
Magnitude x Scale Location x Type of CAT	0.01	0.01	0.01	4
Magnitude x Number of Differences x Discrimination	0.07	0.01	0.06	2
Magnitude x Number of Differences x Boundaries	0	0	0	2
Magnitude x Number of Differences x Type of CAT	0.06	0.03	0.03	2
Magnitude x Discrimination x Boundaries	0	0	0.02	2
Magnitude x Discrimination x Type of CAT	0.77	0.73	0.82	2
Magnitude x Boundaries x Type of CAT	0	0	0	2
Scale Location x Number of Differences x Discrimination	0.01	0	0	2
Scale Location x Number of Differences x Boundaries	0	0	0	2
Scale Location x Number of Differences x Type of CAT	0	0	0	2
Scale Location x Discrimination x Boundaries	0.01	0	0.01	2
Scale Location x Discrimination x Type of CAT	0.02	0.01	0.05	2
Scale Location x Boundaries x Type of CAT	0.03	0	0.01	2
Number of Differences x Discrimination x Boundaries	0	0	0	1
Number of Differences x Discrimination x Type of CAT	0.01	0.03	0.04	1
Number of Differences x Boundaries x Type of CAT	0	0	0	1
Discrimination x Boundaries x Type of CAT	0.01	0.01	0.04	1
Precision x Magnitude x Number of Scales	1.85	0.69	2	16
Number of Scales x Scale Location x Precision	0.07	0.19	0.22	16
Number of Scales x Number of Differences x Precision	0.03	0.01	0.03	8
Number of Scales x Discrimination x Precision	0.19	0.07	0.11	8
Number of Scales x Boundaries x Precision	0	0.02	0.02	8
Magnitude x Scale Location x Precision	0.14	0.1	0.11	16
Magnitude x Number of Differences x Precision	0.29	0.06	0.19	8
Magnitude x Discrimination x Precision	1.45	0.28	1.03	8
Magnitude x Boundaries x Precision	0.02	0.03	0.04	8
Scale Location x Number of Differences x Precision	0.03	0	0.01	8
Scale Location x Discrimination x Precision	0.12	0.04	0.08	8

Scale Location x Boundaries x Precision	0.01	0.01	0.02	8
Number of Differences x Discrimination x Precision	0.04	0.01	0.02	4
Number of Differences x Boundaries x Precision	0	0	0	4
Discrimination x Boundaries x Precision	0.01	0.01	0.01	4
Residuals	3.54	2.28	3.14	975

Table F5. Sums of Squares and Degrees of Freedom from a Six-way ANOVA on FPR for each of Three AMC Hypothesis Tests.

Source of Variation	SS for LR	SS for Z	SS for Score	DF
Main Effects				
Number of Scales	0.01	18.33	20.76	2
Scale Location	0	7.13	10.54	1
Discrimination	0	3.11	5.72	1
Boundaries	0	0.41	0.77	1
Type of CAT	0.01	1.77	1.5	1
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	0.01	1.55	2.68	4
Two-way Interactions				
Scale Location x Number of Scales	0	8.97	11.33	2
Discrimination x Number of Scales	0	4.62	7.72	2
Boundaries x Number of Scales	0	0.17	0.15	2
Type of CAT x Number of Scales	0	0.3	0.62	2
Discrimination x Scale Location	0	2.66	3.93	1
Boundaries x Scale Location	0	0.07	0.2	1
Scale Location x Type of CAT	0	0.36	0.47	1
Boundaries x Discrimination	0	0.01	0.06	1
Discrimination x Type of CAT	0	0	0.02	1
Boundaries x Type of CAT	0	0.34	0.29	1
Precision x Number of Scales	0	0.36	0.67	8
Scale Location x Precision	0	0.49	0.74	4
Discrimination x Precision	0	0.03	0.04	4
Boundaries x Precision	0	0.25	0.41	4
Three-way Interactions				
Discrimination x Scale Location x Number of Scales	0	4.5	5.54	2
Number of Scales x Scale Location x Boundaries	0	0.03	0.1	2
Scale Location x Type of CAT x Number of Scales	0.01	0.2	0.03	2
Number of Scales x Discrimination x Boundaries	0	0.04	0.02	2

Number of Scales x Discrimination x Type of CAT	0	0.01	0.11	2
Number of Scales x Boundaries x Type of CAT	0	0.06	0.03	2
Scale Location x Discrimination x Boundaries	0	0	0	1
Scale Location x Discrimination x Type of CAT	0	0.02	0.03	1
Scale Location x Boundaries x Type of CAT	0	0.04	0.02	1
Discrimination x Boundaries x Type of CAT	0	0	0	1
Precision x Scale Location x Number of Scales	0.01	0.09	0.03	8
Precision x Discrimination x Number of Scales	0.01	0.26	0.36	8
Precision x Boundaries x Number of Scales	0	0.03	0.04	8
Precision x Discrimination x Scale Location	0	0.05	0.1	4
Precision x Boundaries x Scale Location	0	0.01	0.02	4
Discrimination x Boundaries x Precision	0	0.01	0.01	4
Residuals	0.06	2.31	3.14	1199

Table F6. Sums of Squares and Degrees of Freedom from a Seven-way ANOVA on Mean Number of Items for Variable-length Computerized Adaptive Tests.

Source of Variation	SS	DF
Main Effects		
Number of Scales	148784.98	2
Magnitude	1411.28	2
Scale Location	15903.36	2
Number of Differences	610.85	1
Discrimination	222657.03	1
Bound	2826.28	1
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	161692.7	2
Two-way Interactions		
Magnitude x Number of Scales	176	4
Scale Location x Number of Scales	6836.08	4
Number of Scales x Number of Differences	127.7	2
Discrimination x Number of Scales	191.29	2
Boundaries x Number of Scales	503.06	2
Precision x Number of Scales	3438.64	4
Magnitude x Scale Location	245.87	4
Magnitude x Number of Differences	142.44	2
Discrimination x Magnitude	135.21	2
Boundaries x Magnitude	128.6	2

Precision x Magnitude	2.93	4
Scale Location x Number of Differences	382.16	2
Discrimination x Scale Location	1778.73	2
Boundaries x Scale Location	953.81	2
Precision x Scale Location	54.56	4
Discrimination x Number of Differences	63.14	1
Boundaries x Number of Differences	48.86	1
Precision x Number of Differences	1.53	2
Boundaries x Discrimination	1050.76	1
Precision x Discrimination	41048.57	2
Precision x Boundaries	8.1	2
Three-way Interactions		
Number of Scales x Magnitude x Scale Location	78.12	8
Number of Scales x Magnitude x Number of Differences	12.58	4
Number of Scales x Magnitude x Discrimination	127.1	4
Number of Scales x Magnitude x Boundaries	56.74	4
Number of Scales x Magnitude x Precision	69.33	8
Number of Scales x Scale Location x Number of Differences	95.63	4
Number of Scales x Scale Location x Discrimination	3181.62	4
Number of Scales x Scale Location x Boundaries	504.58	4
Number of Scales x Scale Location x Precision	967.84	8
Number of Scales x Number of Differences x Discrimination	72.74	2
Number of Scales x Number of Differences x Boundaries	33.15	2
Number of Scales x Discrimination x Precision	45.82	4
Number of Scales by Discrimination x Boundaries	740.26	2
Precision x Discrimination x Number of Scales	9006.48	4
Number of Scales x Boundaries x Precision	61.91	4
Magnitude x Scale Location x Number of Differences	94.96	4
Magnitude x Scale Location x Discrimination	10.68	4
Magnitude x Scale Location x Boundaries	39.41	4
Magnitude x Scale Location x Precision	3.39	8
Magnitude x Number of Differences x Discrimination	12.46	2
Magnitude x Number of Differences x Boundaries	13.42	2
Magnitude x Number of Differences x Precision	0.63	4
Magnitude x Discrimination x Boundaries	20.8	2
Magnitude x Discrimination x Precision	32.54	4
Magnitude x Boundaries x Precision	3.54	4
Scale Location x Number of Differences x Discrimination	38.87	2

Scale Location x Number of Differences x Boundaries	35.86	2
Scale Location x Number of Differences x Precision	0.86	4
Scale Location x Discrimination x Boundaries	66.62	2
Scale Location x Discrimination x Precision	418.83	4
Scale Location x Boundaries x Precision	4.12	4
Number of Differences x Discrimination x Boundaries	7.59	1
Number of Differences x Discrimination x Precision	13.06	2
Number of Differences x Boundaries x Precision	1.59	2
Discrimination x Boundaries x Precision	82.62	2
Residuals	1486.73	501

Appendix G. Bayes Factor Index

Bayes Factor Index. Originally proposed by Jeffreys (1961), the Bayes Factor is defined by Kass and Raftery (1995) as

$$B = \frac{Y_1}{Y_2} = \frac{P(D|H_1)}{P(D|H_2)} = \frac{\int P(D|\theta, H_1)\pi(\theta|H_1)d\theta}{\int P(D|\theta, H_2)\pi(\theta|H_2)d\theta}, \tag{G1}$$

where $P(D|H)$ is the probability of obtaining observed data D, given that a hypothesis H is true; where $\pi(\theta|H)$ is the prior distribution of θ, given that H is true; and where $P(D|\theta, H)$ is the probability of obtaining D, given some parameter θ for hypothesis H. The Bayes factor B directly describes how much evidence exists for H_1 compared to H_2. The integrals on the right-hand side of the above equation are often intractable. As Kass and Raftery (1995) further describe, a variant of Laplace's method approximates this integral such that

$$\hat{Y}^{\text{MLE}} = (2\pi)^{d/2} I\left(\hat{\theta}_{\text{MLE}}\right)^{-1/2} P\left(D|\hat{\theta}_{\text{MLE}}, H\right) \pi\left(\hat{\theta}_{\text{MLE}}|H\right), \tag{G2}$$

where d is the number of parameters for H; where $\hat{\theta}_{\text{MLE}}$ is the maximum likelihood estimate (MLE) of θ; where $P\left(D|\hat{\theta}_{\text{MLE}}, H\right)$ is the likelihood function evaluated at the MLE; where $I\left(\hat{\theta}_{\text{MLE}}\right)$ is the observed information at the MLE; and where $\pi\left(\hat{\theta}_{\text{MLE}}|H\right)$ evaluates the prior distribution at the MLE. Then, dropping 'MLE' for simplicity, the approximation is written

$$\hat{B} = \frac{\hat{Y}_1}{\hat{Y}_2} = \frac{(2\pi)^{d_1/2} I\left(\hat{\theta}_1\right)^{-1/2} P\left(D|\hat{\theta}_1, H_1\right)\pi\left(\hat{\theta}_1|H_1\right)}{(2\pi)^{d_2/2} I\left(\hat{\theta}_2\right)^{-1/2} P\left(D|\hat{\theta}_2, H_2\right)\pi\left(\hat{\theta}_2|H_2\right)}. \tag{G3}$$

Incorporating $\hat{B}$ into the adaptive measurement of change (AMC) hypothesis testing framework provides the following:

$$\hat{B} = \frac{\hat{P}(D|H_{\mathrm{A}})}{\hat{P}(D|H_0)} = \frac{(2\pi)^{T/2} \prod_{t=1}^{T} \left(I_t(\hat{\theta}_t)^{-1/2} L(\hat{\theta}_t|\mathbf{u}_t) P(\hat{\theta}_t) \right)}{(2\pi)^{1/2} I_p(\hat{\theta}_p)^{-1/2} L(\hat{\theta}_p|\mathbf{u}_p) P(\hat{\theta}_p)}, \tag{G4}$$

where $P(\hat{\theta})$ is a prior distribution like the uniform distribution (i.e., an uninformative prior) or the standard normal distribution (i.e., an informative prior). For example, if using the standard normal,

$$\hat{B} = \frac{\hat{P}(D|H_{\mathrm{A}})}{\hat{P}(D|H_0)} = \frac{(2\pi)^{T/2} \prod_{t=1}^{T} \left(I_t(\hat{\theta}_t)^{-1/2} L(\hat{\theta}_t|\mathbf{u}_t) \Phi(\hat{\theta}_t) \right)}{(2\pi)^{1/2} I_p(\hat{\theta}_p)^{-1/2} L(\hat{\theta}_p|\mathbf{u}_p) \Phi(\hat{\theta}_p)}, \tag{G5}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution. Further, applying a logarithm transformation, and using probabilities to denote priors,

$$\begin{aligned} \hat{C} &= \log(\hat{B}) \\ &= \left(\frac{T-1}{2}\right)\log(2\pi) + \textstyle\sum_{t=1}^{T} \log\left(P(\hat{\theta}_t)\right) - \frac{1}{2}\sum_{t=1}^{T} \log\left(I_t(\hat{\theta}_t)\right) + \sum_{t=1}^{T} \ell\left(\hat{\theta}_t|\mathbf{u}_t\right) \\ &\quad -\log\left(P(\hat{\theta}_p)\right) + \tfrac{1}{2}\log\left(I_p(\hat{\theta}_p)\right) - \ell(\hat{\theta}_p|\mathbf{u}_p). \end{aligned} \tag{G6}$$

$\hat{C}$ represents a new possible AMC statistic, but this statistic is not easily interpretable. To improve interpretability, $\hat{C}$ could be multiplied by -2 to produce:

$$\begin{aligned} -2\hat{C} &= -2\log(\hat{B}) \\ &= (1-T)\log(2\pi) - 2\textstyle\sum_{t=1}^{T} \log\left(P(\hat{\theta}_t)\right) + \sum_{t=1}^{T} \log\left(I_t(\hat{\theta}_t)\right) \\ &\quad -2\textstyle\sum_{t=1}^{T} \ell\left(\hat{\theta}_t|\mathbf{u}_t\right) + 2\log\left(P(\hat{\theta}_p)\right) - \log\left(I_p(\hat{\theta}_p)\right) + 2\ell(\hat{\theta}_p|\mathbf{u}_p). \end{aligned} \tag{G7}$$

As Kass and Raftery (1995) describe, the likelihood-ratio statistic is found within a structure like $-2\hat{C}$ by combining the fourth and seventh terms of the last equation, allowing the simplification:

$$
\begin{aligned}
-2\hat{C} \quad &= (1-T)\log(2\pi) - 2\sum_{t=1}^{T}\log\left(P(\hat{\theta}_t)\right) + \sum_{t=1}^{T}\log\left(I_t(\hat{\theta}_t)\right) \\
&\quad -\lambda_{\text{obs}} + 2\log\left(P(\hat{\theta}_p)\right) - \log\left(I_p(\hat{\theta}_p)\right).
\end{aligned}
\tag{G8}
$$

Then, the statistic $-2\hat{C}$ represents a penalized likelihood-ratio statistic whose penalties depend on five quantities: (1) observed information for the pooled MLE; (2) observed test information for the separate tests; (3) a prior distribution evaluated at the pooled MLE; (4) a prior distribution evaluated at each trait estimate; and (5) the number of scales.

This statistic might provide a new path to explore for AMC hypothesis testing. Specifically, prior distributions might be modified to explore their impact on the false positive rate or true positive rate. For example, a standard normal distribution might be useful for modeling the pooled MLE. In addition, normal distributions with mean zero and a small variance might be used to model each trait score. Measurement is intra-individual; so correlations need not exist in this scenario (e.g., such as correlations that might exist using a multivariate normal distribution). To be careful, a strong threshold could be used for identifying meaningful evidence for the alternative hypothesis. For example, using $-2\hat{C} = -6$ corresponds to $\hat{B} \approx 20.086$. This threshold corresponds to strong evidence for the alternative hypothesis, according to Jeffreys (1998).

Results. In addition to the likelihood-ratio test, *Z*-test, and score test, the Bayes Factor Index (BFI) $-2\hat{C}$ was calculated for all conditions of the Monte Carlo simulation. A threshold of $-2\hat{C} = -6$ was used to identify psychometrically significant differences. This threshold corresponds to a Bayes Factor of about 20 and to strong evidence for the alternative hypothesis that at least one psychometrically significant difference exists among a set of scales for a single individual. Then, $-2\hat{C} < -6$ indicates significance. A normal distribution was used as the prior distribution of the pooled MLE. For any given simulation condition, the mean among the trait scores for that condition was used as the mean for this normal prior, and 0.85 was used as the variance. For each trait score, a normal distribution was used as the prior distribution. The trait scores for the simulation for a given condition served as the means for these prior normal distributions. The variance for each of these prior normal distributions was set at 0.1.

Table G1 displays ANOVA results for the TPR for the BFI. Unlike the three typical AMC hypothesis tests, magnitude was not the largest source of variation among the main effects. Instead, the number of scales was the largest source of variation, accounting for 23.4% of observed TPR variance. The second largest source of variation was precision at 17.4%, which was larger than the other three tests. In addition, magnitude and scale location accounted for 9.2% and 6.0% of TPR variance, respectively. Together, main effects accounted for 61.9% of observed TPR variance for the BFI.

For the number of scales, the mean TPRs for the BFI included: 0.728 for 5 scales, 0.517 for 11 scales, and 0.897 for 18 scales. While worst for 11 scales, the BFI had the second-best performance for both 5 and 18 scales, compared to the other three hypothesis tests. For test length for FL-CATs, mean TPRs included: 0.396 for 10 items, 0.724 for 30 items, and 0.818 for 50 items. TPRs improved as test length increased although the improvement between 10 and 30 items was largest. For standard error termination values for VL-CATs, mean TPRs included: 0.707 for a standard error of 0.4, 0.799 for a standard error of 0.3, and 0.842 for a standard error of 0.2. As standard error decreased (i.e., as precision increased), performance improved. Finally, mean TPRs for levels of magnitude included: 0.580 for a magnitude of 0.5, 0.755 for a magnitude of 1.0, and 0.809 for a magnitude of 1.5. TPRs improved as magnitude increased.

Table G1. Classical Effect Sizes (η^2) from an Eight-way ANOVA on TPR for the Bayes Factor Index.

Source of Variation	η^2 for BFI
Main Effects	
Number of Scales	0.234
Magnitude	0.092
Scale Location	0.060
Number of Differences	0.006
Discrimination	0.002
Boundaries	0.006
Type of CAT	0.045
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	0.174
Two-way Interactions	
Magnitude x Number of Scales	0.029
Scale Location x Number of Scales	0.041
Number of Scales x Number of Differences	0.005
Discrimination x Number of Scales	0.046
Boundaries x Number of Scales	0.003
Type of CAT x Number of Scales	0.027
Magnitude x Scale Location	0.003
Magnitude x Number of Differences	0.004
Magnitude x Discrimination	0.018
Boundaries x Magnitude	0
Magnitude x Type of CAT	0.003
Scale Location x Number of Differences	0.001
Discrimination x Scale Location	0.012
Boundaries x Scale Location	0.001
Type of CAT x Scale Location	0.001
Discrimination x Number of Differences	0.001
Boundaries x Number of Differences	0
Type of CAT x Number of Differences	0
Boundaries x Discrimination	0.002
Discrimination x Type of CAT	0.037

Boundaries x Type of CAT	0.002
Precision x Number of Scales	0.007
Precision x Magnitude	0.006
Precision x Scale Location	0
Precision x Number of Differences	0
Precision x Discrimination	0.033
Precision x Boundaries	0.001
Three-way Interactions	
Number of Scales x Magnitude x Scale Location	0.004
Number of Scales x Magnitude x Number of Differences	0.002
Number of Scales x Magnitude x Discrimination	0.004
Number of Scales x Magnitude x Boundaries	0
Number of Scales x Magnitude x Type of CAT	0.001
Number of Scales x Scale Location x Number of Differences	0.003
Number of Scales x Scale Location x Discrimination	0.018
Number of Scales x Scale Location x Boundaries	0.001
Number of Scales x Scale Location x Type of CAT	0.001
Number of Scales x Number of Differences x Discrimination	0.001
Number of Scales x Number of Differences x Boundaries	0
Number of Scales x Number of Differences x Type of CAT	0
Number of Scales x Discrimination x Boundaries	0.001
Number of Scales x Discrimination x Type of CAT	0.002
Number of Scales x Boundaries x Type of CAT	0
Magnitude x Scale Location x Number of Differences	0.002
Magnitude x Scale Location x Discrimination	0
Magnitude x Scale Location x Boundaries	0
Magnitude x Scale Location x Type of CAT	0
Magnitude x Number of Differences x Discrimination	0
Magnitude x Number of Differences x Boundaries	0
Magnitude x Number of Differences x Type of CAT	0
Magnitude x Discrimination x Boundaries	0
Magnitude x Discrimination x Type of CAT	0.002
Magnitude x Boundaries x Type of CAT	0
Scale Location x Number of Differences x Discrimination	0
Scale Location x Number of Differences x Boundaries	0
Scale Location x Number of Differences x Type of CAT	0
Scale Location x Discrimination x Boundaries	0
Scale Location x Discrimination x Type of CAT	0

Scale Location x Boundaries x Type of CAT	0
Number of Differences x Discrimination x Boundaries	0
Number of Differences x Discrimination x Type of CAT	0
Number of Differences x Boundaries x Type of CAT	0
Discrimination x Boundaries x Type of CAT	0
Precision x Magnitude x Number of Scales	0.004
Number of Scales x Scale Location x Precision	0.010
Number of Scales x Number of Differences x Precision	0
Number of Scales x Discrimination x Precision	0.004
Number of Scales x Boundaries x Precision	0.002
Magnitude x Scale Location x Precision	0.001
Magnitude x Number of Differences x Precision	0.001
Magnitude x Discrimination x Precision	0.002
Magnitude x Boundaries x Precision	0
Scale Location x Number of Differences x Precision	0
Scale Location x Discrimination x Precision	0
Scale Location x Boundaries x Precision	0
Number of Differences x Discrimination x Precision	0
Number of Differences x Boundaries x Precision	0
Discrimination x Boundaries x Precision	0.001
Residuals	0.029

Table G2 displays the ANOVA results for the FPR for the BFI. Like the *Z*-test and score test, the number of scales was the largest source of variation among the main effects, accounting for 31.4% of observed FPR variance. Unlike the *Z*-test or score test, however, scale location and discrimination did not account for large amounts of variance for the BFI. Instead, similar to the LRT, the type of CAT and precision were the second and third most important main effects, accounting for 5.7% and 4.5% of observed FPR variance, respectively. Together, main effects accounted for 43.4% of variance. Across all simulation conditions, the mean FPR for the BFI was 0.039, lower than all other hypothesis tests.

For the number of scales, the mean FPRs for the BFI included: 0.016 for 5 scales, 0.034 for 11 scales, and 0.066 for 18 scales. As the number of scales increased, mean FPRs increased. For the type of CAT, the mean FPR was 0.030 for FL-CATs and was 0.047 for VL-CATs. FPRs were somewhat smaller for FL-CATs. Moreover, for test length for FL-CATs, mean FPRs included: 0.031 for 10 items, 0.033 for 30 items, and 0.026 for 50 items. For standard error termination values for VL-CATs, mean FPRs included: 0.063 for a 0.4 standard error, 0.040 for a 0.3 standard error, and 0.039 for a 0.2 standard error.

Table G2. Classical Effect Sizes (η^2) from a Six-way ANOVA on FPR for the Bayes Factor Index.

Source of Variation	η^2 for BFI
Main Effects	
Number of Scales	0.314
Scale Location	0
Discrimination	0.017
Boundaries	0.001
Type of CAT	0.057
Precision (Test Length for FL-CAT or Standard Error for VL-CAT)	0.045
Two-way Interactions	
Scale Location x Number of Scales	0.008
Discrimination x Number of Scales	0.008
Boundaries x Number of Scales	0.001
Type of CAT x Number of Scales	0.072
Discrimination x Scale Location	0.004
Boundaries x Scale Location	0.002
Scale Location x Type of CAT	0.005
Boundaries x Discrimination	0.001
Discrimination x Type of CAT	0.008
Boundaries x Type of CAT	0
Precision x Number of Scales	0.170
Scale Location x Precision	0.009

Discrimination x Precision	0.047
Boundaries x Precision	0.002
Three-way Interactions	
Discrimination x Scale Location x Number of Scales	0.024
Number of Scales x Scale Location x Boundaries	0.001
Scale Location x Type of CAT x Number of Scales	0.002
Number of Scales x Discrimination x Boundaries	0.001
Number of Scales x Discrimination x Type of CAT	0.026
Number of Scales x Boundaries x Type of CAT	0.001
Scale Location x Discrimination x Boundaries	0.003
Scale Location x Discrimination x Type of CAT	0
Scale Location x Boundaries x Type of CAT	0
Discrimination x Boundaries x Type of CAT	0.002
Precision x Scale Location x Number of Scales	0.048
Precision x Discrimination x Number of Scales	0.048
Precision x Boundaries x Number of Scales	0.003
Precision x Discrimination x Scale Location	0.015
Precision x Boundaries x Scale Location	0.002
Discrimination x Boundaries x Precision	0.002
Residuals	0.051

Discussion. The BFI performed better than the *Z*-test and score test in terms of the FPR, but it performed worse in terms of the TPR. Nevertheless, the BFI had the second-best performance for low magnitudes of 0.5 but maintained a consistently good FPR across simulation conditions. This result was the best balance between FPR and TPR for low magnitudes among all hypothesis tests explored under the conditions of the current study. Future work might examine this index in greater detail to determine if it is viable. For example, BFI mean FPRs were consistently lower than 0.05 and could be increased slightly if doing so produces higher mean TPRs. To achieve this outcome, the BFI threshold might be increased slightly (i.e., made closer to zero, such as using a value of -5 instead of -6) to be less conservative. A future study might be conducted to systematically examine different threshold values or different prior distributions to determine if the tradeoff between FPR and TPR could be optimized further. If further research finds it viable, the BFI might serve as a supplement to the LRT when detecting differences of small magnitudes in some real-world applications.

References

Jeffreys, H. (1998) (1961). *The theory of probability*. Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

Appendix H. Additional Equations

Graded Response Model and the Score Function

For $g = 0,1,\ldots,m$ integers, specifying $k = 1,2,\ldots,m+1$ categories, for some i-th item, the graded response model has an item information function given by

$$I_i(\theta) = \sum_{g=0}^{m} \frac{\left(P_{i_g}^{*\prime} - P_{i_{g+1}}^{*\prime}\right)^2}{P_{i_g}^{*} - P_{i_{g+1}}^{*}}, \tag{H1}$$

where the probability boundary function is given by

$$P_{i_g}^{*} = \left(1 + \exp\left(-a\left(\theta - b_{i_g}\right)\right)\right)^{-1}, \tag{H2}$$

with the exception of $g = 0$, in which case $P_{i_0}^{*} = 1$. The first derivative of the probability boundary function is given by

$$P_{i_g}^{*\prime} = (a)\exp\left(a\left(\theta - b_{i_g}\right)\right)\left(1 + \exp\left(a\left(\theta - b_{i_g}\right)\right)\right)^{-2}, \tag{H3}$$

with the exception of $g = 0$, in which case $P_{i_0}^{*\prime} = 0$. Then, the total test information function is given by

$$I(\theta) = \sum_{i=1}^{I} I_i\,(\theta), \tag{H4}$$

assuming local independence, for $i = 1,\ldots,I$ items (Ostini and Nering, 2006, p. 69). Furthermore, category probabilities can be calculated via the following:

$$\begin{aligned} P_{i_0} &= 1 - P_{i_1}^{*} \\ P_{i_g} &= P_{i_g}^{*} - P_{i_{g+1}}^{*} \\ P_{i_m} &= P_{i_m}^{*} - 0. \end{aligned} \tag{H5}$$

The likelihood function then has the form

$$L(\theta|\mathbf{u}) \quad = \prod_{i=1}^{I}\left(\prod_{k=1}^{m+1} P_{i_g}\,(U_i = u_k)^{\delta_k}\right), \tag{H6}$$

where δ_k = 1 when the response indicates the k-th category and where δ_k = 0 otherwise. The log-likelihood function has the form

$$\ell(\theta|\mathbf{u}) \quad = \sum_{i=1}^{I}\left(\sum_{k=1}^{m+1} \delta_k \log\left(P_{i_g}(U_i = u_k)\right)\right). \tag{H7}$$

The score function is the first derivative of the log-likelihood function. For the graded response model, the score function has a form that depends on the type of category probability. Specifically, the following three forms exist:

$$\begin{aligned}\frac{d}{d\theta}\log\left(P_{i_0}\right) &= \frac{d}{d\theta}\log\left(1 - P^*_{i_0}\right)\\ \frac{d}{d\theta}\log\left(P_{i_g}\right) &= \frac{d}{d\theta}\log\left(P^*_{i_g} - P^*_{i_{g+1}}\right)\\ \frac{d}{d\theta}\log\left(P_{i_m}\right) &= \frac{d}{d\theta}\log\left(P^*_{i_m}\right).\end{aligned} \tag{H8}$$

Then, three cases exist with the following derivatives:

$$\begin{aligned}\frac{d}{d\theta}\log\left(P_{i_0}\right) &= -a\exp\left(a\left(\theta - b_{i_0}\right)\right)\left(1 + \exp\left(a\left(\theta - b_{i_0}\right)\right)\right)^{-1}\\ \frac{d}{d\theta}\log\left(P_{i_g}\right) &= \frac{-a\left(\exp\left(a\left(\theta - b_{i_g}\right)\right) + \exp\left(a\left(\theta - b_{i_{g+1}}\right)\right)\right)}{\left(1+\exp\left(a\left(\theta - b_{i_g}\right)\right)\right)\left(1+\exp\left(a\left(\theta - b_{i_{g+1}}\right)\right)\right)}\\ \frac{d}{d\theta}\log\left(P_{i_m}\right) &= a\left(1 + \exp\left(a\left(\theta - b_{i_m}\right)\right)\right)^{-1}.\end{aligned} \tag{H9}$$

For the graded response model, the score function can then be written

$$\begin{aligned}s(\theta|\mathbf{u}) &= \frac{\partial\ell}{\partial\theta}\\ &= \sum_{i=1}^{I}\left(\sum_{k=1}^{m+1}\delta_k \frac{d}{d\theta}\log\left(P_{i_g}(U_i = u_k)\right)\right),\end{aligned} \tag{H10}$$

where derivatives for different boundaries are related to the three cases.

References

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage.