ABERRANT RESPONSE PATTERNS AS A MULTIDIMENSIONAL PHENOMENON: USING FACTOR-ANALYTIC MODEL COMPARISON TO DETECT CHEATING

BY

John Michael Clark III

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

Todd D. Little
Co-Chairperson

Kristopher J. Preacher
Co-Chairperson

Pascal R. Deboeck

Wei Wu

William P. Skorupski

Date defended: June 2, 2010

The Dissertation Committee for John Michael Clark III certifies
that this is the approved version of the following dissertation:

ABERRANT RESPONSE PATTERNS AS A MULTIDIMENSIONAL PHENOMENON:
USING FACTOR-ANALYTIC MODEL COMPARISON TO DETECT CHEATING

Todd D. Little
Co-Chairperson

Kristopher J. Preacher
Co-Chairperson

Date approved: June 2, 2010

Abstract

This dissertation proposes a new factor-analytic technique for detecting cheating on exams. Person-fit statistics have been developed to assess the extent to which examinees' response patterns are consistent with expectation, with expectation defined in the context of some model. Response patterns that are inconsistent with expectation are said to be aberrant. Many person-fit statistics have been developed, mostly in the context of classical test theory or item response theory. However, in the person-fit literature, most of these techniques rely on assessing person-fit for unidimensional measurement models. This dissertation proposes that cheating can be conceptualized as a multidimensional phenomenon. A new person-fit technique that involves comparing changes in person-fit across one-factor- and two-factor exploratory factor analysis models is investigated. A statistically-significant improvement in person-fit when adding a second factor to the model is taken as evidence of cheating. Results indicate that this new technique may be useful for detecting cheating when a small-to-moderate proportion of examinees are cheaters. Suggestions are offered for future research on this new technique.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Aberrant Response Patterns as a Multidimensional Phenomenon: Using Factor-Analytic Model Comparison to Detect Cheating

Many industries are regulated by governing bodies. Such guardians of practice may be autonomous, self-governing bodies within the industry itself, or regulation may be imposed by an outside entity, such as the state or federal government. In either case, a common tool used to regulate practice is an examination. Examinations that are used to regulate entry into the field are referred to as licensure examinations, and these often are overseen directly or indirectly by a government body. Examinations that are used to demonstrate a certain level of achievement within a field, or perhaps mastery of a set of skills particular to a specialty within a field, are referred to as certification exams. The distinction between these two examination types is that examinees must pass a licensure exam to gain entry into the field in which they wish to practice, whereas completion of a certification exam is voluntary. However, both types of examinations are sometimes referred to as high stakes exams. The stakes for examinees who take licensure exams are high for obvious reasons—if an examinee fails a licensure exam, he or she will not be permitted to practice. Although certification exams are voluntary, failing such an exam may carry other undesirable consequences. Certification test takers' promotions, levels of compensation, and job duties may depend on whether or not they earn the credential they are seeking.

The necessity for these exams lies in public protection and preservation of the integrity of an industry or practice. Often, these kinds of examinations are used to protect the public from incompetent practitioners. Sometimes, these exams may be used as a tool to recognize an individual as having mastered a certain specialty, or perhaps simply being a highly skilled and knowledgeable practitioner. Regardless of the ultimate purpose or goal of the examination, the

credentialing process is dependent on a properly-functioning examination. Such exams are designed to divide test takers into two mutually exclusive groups: those who pass and those who fail. Individuals who pass the exam are said to have reached the credentialing body's criterion of minimal competence and earn the credential associated with the exam. Those who fail the exam fall below this criterion of minimal competence and are denied the credential they seek. Boards or governing bodies, in conjunction with experts from the testing industry, work to create examinations that are valid and reliable, with a minimal amount of measurement error. By following proper procedures and best practices, the Board or governing body can be confident that the pass/fail decisions rendered by the exam are accurate and appropriate. However, sometimes even when test development best practices have been followed strictly, threats to the validity of the exam cannot be avoided. One especially salient threat to the validity of an exam outcome is cheating.

Cheating will be defined in this paper as an activity carried out by a test taker in a conscious and willful effort to gain an unfair advantage on the exam. Cheating in the context of licensure and certification exams is problematic because such behavior has a negative impact on the criterion-related validity of the exam, it reduces the value of the credential, and it may potentially result in increased risk to the public.

Cheating can manifest itself in several ways. Cheating behaviors include, but are not limited to, copying answers from another test-taker, bringing a "cheat sheet" with answers into an exam, and participating in an unfair or dishonest test preparation course that uses items stolen from an examination. A variety of statistical techniques have been developed to identify cheating behavior, but methods largely vary depending on the measurement framework that is employed.

The following section contains brief overviews of three measurement frameworks that are employed in testing: classical test theory, item response theory, and factor analysis.

Common Measurement Frameworks Used in Testing

Classical test theory.

In classical test theory (CTT), a given examinee's overall test score, *Y*, is a sum of two parts: the examinee's unobserved true score (*T*) and a random error component (*E*). This relationship is represented in the equation (McDonald, 1999),

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

At the item level, this relationship is represented for a given examinee as

$$X_j = T + E_j, \tag{2}$$

where items are indexed $j = 1, 2, \ldots J$. According to classical test theory, the examinee's true score (*T*) is the only common component across items for the examinee. *E* is assumed to have a mean of 0, it is assumed to be uncorrelated with *T*, and error terms associated with all items on the exam are assumed to be independent of each other (McDonald, 1999). To describe *T* in terminology used in common factor theory: classical test theory assumes that a unidimensional factor structure underlies item responses on an exam, with individual item responses being fully independent from one another after accounting for the lone systematic component, *T*.

Much of the focus of classical test theory is directed at the overall test. Reliability and measurement error, for example, are assessed at the test level. However, techniques for assessing performance of individual items have been developed in the classical test theory measurement framework. Two important indicators of item performance include the item's difficulty level and its discrimination. In classical test theory, the difficulty parameter for a given item, π_j, is estimated by the sample statistic, p_j, which is equal to the proportion of examinees providing a

correct answer to a given item (McDonald, 1999). The discrimination index provides information regarding how well an item differentiates examinees of varying ability levels. In general, discrimination indices relate performance on a particular item to performance on the overall test. One of the common methods for estimating item discrimination in classical test theory is the item-total correlation, which is calculated as the correlation between responses to a particular item and total scores on the exam.

Variations on how to compute item-total correlations exist (McDonald, 1999). Some calculations include the item that is being assessed in calculating the overall test score, and some calculations remove it from the total score. Regardless of computational variations, the concept of discrimination fundamentally involves relating test-taker ability, as represented by test score, to performance on a particular item. If high-ability test takers tend to answer a given item correctly and low-ability test-takers tend to answer the item incorrectly, the item is displaying positive discrimination—a desirable characteristic. If test-taker ability has no appreciable relationship to performance on an item, the item does not discriminate at all. If low-ability test-takers tend to answer an item correctly while high-ability test-takers answer the item incorrectly, the item is said to have negative discrimination.

Item response theory.

Parametric IRT for dichotomous data.

Whereas the primary concern and focus of attention in classical test theory is on the overall test score, item response theory (IRT) is a measurement technique that—as its name implies—is largely focused at the item level. Item response theory is a label applied to an entire family of statistical models of varying complexity and focus. Hambleton, Swaminathan, and

Rogers (1991) provide a basic overview of IRT for dichotomous data, which will be summarized in this section.

When tests are comprised entirely of dichotomous data, three IRT models are especially common. The 3-parameter logistic (3-PL) IRT model models the probability of a correct response on item *j* with the function,

$$P(X_j = 1|\theta) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}, \quad (3)$$

where θ is the conditional value of the examinee's ability parameter, b_j is the difficulty parameter, a_j is the discrimination parameter, and c_j is the lower asymptote for item *j*. Plotting the function across a range of conditional values of θ results in an S-shaped curve (or ogive), which is referred to as the item response function (IRF), or item characteristic curve (ICC). This function represents the expected value of X_j at conditional levels of θ.

In some contexts, the *b* parameter may be referred to generically as an item's location parameter. Similarly, the θ parameter is sometimes referred to generically as the individual's trait parameter. These labels are most commonly applied in contexts in which it does not make sense to refer to the "difficulty" of an item or the "ability" level required of an individual to provide a response. For example, the *b* and θ parameters might be labeled as the location and trait parameter, respectively, in a case in which an IRT model is fit to personality data. However, for a testing situation in which item responses are coded correct or incorrect (or variations between, as in partial credit scoring), labeling *b* as the item difficulty parameter and θ as the examinee ability parameter is appropriate, so those labels will be used in this paper.

The item's *b* parameter and the examinee's θ parameter are measured on the same scale. This common scale allows for the difficulty of a particular item to be interpreted in terms of how much ability is required in order to have a particular probability of providing a correct response.

The value of the *b* parameter is set to the point where the conditional probability of a correct response equals $(1 + c) / 2$.

The *a* parameter is proportional to the slope of the item response function at its inflection point. The *a* parameter is assumed to be positive—a negative *a* parameter would represent an negative relationship between ability level and probability of success—and larger values of *a* correspond to steeper item response function slopes. In IRT, the *a* parameter functions as the item's discrimination parameter.

The *c* parameter is sometimes labeled as the item's pseudo-guessing parameter. For certain item types, particularly multiple-choice items with no penalty imposed for guessing, it may be unreasonable to assume that even an examinee with an extremely low ability level would have a probability of success on a given item that is near zero. For example, a low-ability test-taker may have a probability of success that is closer to $1 / 4 = 0.25$ than 0 for a 4-option multiple choice item. By estimating a *c* parameter, the IRF is allowed to have a lower asymptote that approaches a value greater than zero to account for behaviors such as guessing.

Historically, early applications of IRT used the normal cumulative density function (CDF) to draw the item response function. Although the logistic function is now standard, sometimes the *a* parameter is multiplied by a constant, 1.701, which causes the logistic item response function to take on a shape that closely approximates an item response function created using the normal CDF. This scaling constant is denoted *D*. An example item response function appears in Figure 1. For this example item, $a = 1.25$, $b = 0.00$, and $c = 0.25$; the scaling constant was omitted from this item response function.

The item response function shown in Figure 1 represents the probability of a correct response as a function of examinee ability level. The probability of a correct response for item *j*

is represented by the function, $P(X_j = 1|\theta)$, or more compactly as $P_j(\theta)$. Another function that will become important for future discussion is the probability of an incorrect response. The probability of an incorrect response for item *j* is equal to $P(X_j = 0|\theta) = 1 - P_j(\theta)$, which is often written simply as $Q_j(\theta)$. An example item response function with both $P_j(\theta)$ and $Q_j(\theta)$ functions plotted is shown in Figure 2.

In addition to the 3-PL model, there are other IRT models for dichotomous response data that have varying levels of constraints imposed on the number of estimated parameters for each item. The 2-parameter logistic model (2-PL) imposes the constraint on the 3-PL model that all items have a lower asymptote of 0, or $c = 0$ for all items. The probability of a correct response to item *j* is related to ability in the 2-PL model by the following function,

$$P\left(X_j = 1\middle|\theta\right) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}. \quad (4)$$

Imposing a restriction that $c = 0$ for all items allows all items to have lower asymptotes that approach 0.

Another popular IRT model for dichotomous response data is the 1-parameter logistic (1-PL) model, or sometimes referred to as the Rasch model. The 1-PL model places an additional limitation on estimated item parameters, constraining all items to have the same estimated discrimination parameter, or $a = 1$ for all items. The probability of a correct response to item *j* is represented in a 1-PL model by the item response function,

$$P\left(X_j = 1\middle|\theta\right) = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}}. \quad (5)$$

Nonparametric IRT for dichotomous data.

As described in Mokken (1997), nonparametric IRT models are similar to traditional parametric IRT models in that they relate the probability of success on individual items to

examinee ability using an item response function. However, in nonparametric item response theory, neither the item response function, nor the examinee's ability parameter are specified using parametric distributions (Mokken, 1997). Nonparametric IRT models are based on three assumptions: (1) items are unidimensional, (2) item responses are locally independent after accounting for θ, and (3) item response functions are monotonically nondecreasing. The nonparametric IRT model that meets these assumptions is known as the model of monotone homogeneity. The model of double monotonicity goes one step further and imposes the constraint that item response functions cannot intersect. Both the robustness of nonparametric item response theory models and the ability to restrict nonparametric item response functions from intersecting have made nonparametric IRT models useful for computing certain person-fit statistics and for creating person response function plots (e.g., Meijer, 2003; Emons, Sijtsma, & Meijer, 2004, 2005), however, researchers have found evidence that person-fit statistics are less accurate when nonparametric IRT models are used (St-Onge, Valois, Abdous, & Germain, 2009).

Parametric IRT for polytomous data.

Item response theory models have applicability beyond dichotomous data. Numerous IRT models have been developed for polytomous data as well. As explained by Samejima (1997), the graded response model (GRM) can be applied to data obtained from Likert scales or tests in which partial credit is awarded. While the previously-discussed IRT models for dichotomous data model the probability of success on a given item, the GRM can be used to model the probability of scoring in a particular ordered category (x) or higher at a given level of θ. The graded response model is somewhat similar to the dichotomous 2-parameter logistic model for dichotomous data, in that only the a and b parameters are modeled for items, but for item j with a

minimum possible score of 0 and a maximum possible score *m*, one *a* parameter and *m b* parameters are estimated. As shown in Figure 3, each estimated *b* parameter represents the boundary between different scores on item *j*. Each curve in this figure represents the conditional probability of obtaining a score of *x* or higher, given θ, on this item. As illustrated in Figure 4, response probabilities for individual scores on items can also be modeled using the GRM. Each curve in the plot represents the probability of a particular score, *x*, on item *j*, conditional on θ.

Factor analysis.

Factor analysis (FA) is a measurement technique that is most often used to explain observed correlations among a set of items by means of a smaller set of latent constructs, or factors (Brown, 2006; McDonald, 1999). Although factor analysis is used less often than CTT or IRT in traditional testing applications, this measurement framework has some applicability for testing. For a unidimensional test, the model for item *j* is given by

$$X_j = \mu_j + \lambda_j\theta + \varepsilon_j, \tag{6}$$

where μ_j is the item's intercept , λ_j is the item's factor loading, θ is the examinee's latent factor score, and ε_j is a unique factor, consisting of specific and error components. For an FA model applied to test data, μ_j is an indicator of item *j*'s difficulty, and λ_j is an indicator of the item's discrimination. The value $\mu_j + \lambda_j\theta$ can be conceptualized as the expected value of X_j, given the examinee's factor score. Equation (6) can easily be generalized for multidimensional models by adding additional factor scores and loadings. Although multidimensional IRT models have been developed, at the present time it is far easier to test multidimensional models using commercially-available software in the context of factor analysis.

Similarities Between the FA and IRT Measurement Frameworks

Common ground between factor analysis and item response theory has been well documented (e.g., Brown, 2006; Kamata & Bauer, 2008; McDonald, 1999; Wirth & Edwards, 2007). Although FA and IRT are often used in different applications for different purposes, the two techniques share certain similarities that make a short comparison useful for the present discussion of person-fit. Every parameter in a 1-PL or 2-PL IRT model can be represented in an approximately equivalent manner in a factor analytic model. The transformations that are discussed next have been demonstrated in Brown (2006). It is quite easy to transform parameter estimates from one measurement framework to another. To convert FA parameters into IRT parameters, the IRT *a* parameter is computed,

$$a = \frac{\lambda}{\sqrt{1 - \lambda^2}}, \tag{7}$$

and the *b* parameter is computed,

$$b = \frac{\tau}{\lambda}, \tag{8}$$

where τ is an estimated threshold parameter.

The IRT and factor-analytic measurement frameworks begin to diverge at the IRT *c* parameter. As previously discussed, the *c* parameter is used in IRT to represent the extent to which guessing affects success probabilities for examinees with very low ability levels, which affects the lower asymptote of the item response function. No parameter in factor analysis directly corresponds to IRT's *c* parameter, and there is no method to transform estimates from factor analysis into an IRT *c* parameter or vice-versa.

In the preceding paragraphs, simple item-level transformations were demonstrated to illustrate how item parameters relate to one another across the two measurement frameworks. At

the model level, making comparisons across measurement frameworks is similarly straightforward. A 2-parameter logistic IRT model is equivalently represented in factor analysis by a model with all items loading onto a single latent construct. A 1-parameter logistic IRT model is equivalently represented in factor analysis with the additional requirement that loadings be constrained to equality.

Aberrant Response Patterns

Upon inspection, certain patterns of item responses may seem strange or otherwise unexpected. For example, a low ability examinee who answers several very difficult items correctly would represent an unexpected occurrence. In broad, general terms, aberrant response patterns can be defined as response patterns that defy some expectation. Of course, the primary challenge in this endeavor is defining a level of expectation so a judgment can be rendered regarding the degree of aberrance associated with an examinee's response pattern.

A large number of person-fit indicators have been developed for the purposes of identifying aberrant response patterns (Karabatsos, 2003). Many of these techniques compare observed response patterns to expected outcomes defined by a particular model. These values are compared, and person misfit occurs where observed item response patterns are incongruous with what is implied by the model (Meijer, 1996; Meijer, Muijtjens, & van der Vleuten, 1996; Meijer & Sijtsma, 2001). However, depending on the context of measurement (CTT, IRT, or FA), methods for defining expectation in a response pattern and measuring deviations from expectation differ.

The Guttman model.

In classical test theory, one simple and intuitive technique for establishing what characteristics are to be expected in a response string is based on the concept of the deterministic

Guttman model (Guttman, 1944). Let X_j represent the score on item j, with a correct answer coded $X_j = 1$ and an incorrect answer coded $X_j = 0$. The number-correct (NC) score, r, is equal to the sum of item responses. According to the Guttman model, for item j,

$$\theta < \delta_j \Leftrightarrow P_j(\theta) = 0, \tag{9}$$

and

$$\theta \geq \delta_j \Leftrightarrow P_j(\theta) = 1, \tag{10}$$

where θ is the examinee's latent ability level, δ_j is the difficulty parameter for item j, which is measured on the same scale as θ, and $P_j(\theta)$ is the conditional probability of a given examinee of ability level θ correctly answering item j. According to this model, if a given examinee has an ability level that is greater than or equal to the difficulty level of a particular item, the probability of a correct response is 1, and if the examinee's ability level is less than the item's difficulty level, the probability of a correct response is 0 (Guttman 1994; Meijer & Sijtsma, 1995; Meijer & Sijtsma, 2001). The Guttman model therefore implies that for a given examinee's number-correct score of r on an exam, it is assumed that the examinee answered only the r easiest items on the exam correctly and that all of the r easiest items have difficulty parameters that are less than or equal to the examinee's ability level. As described by Meijer and Sijtsma (2001), a response vector in which the r easiest items are answered correctly and the $J - r$ items are answered incorrectly is known as a Guttman pattern or a conformable pattern. The opposite circumstance, in which the r most difficult items are answered correctly and the $J - r$ easiest items are answered incorrectly is known as a reverse Guttman pattern. At the item level, items that are scored contrary to expectation as outlined in the Guttman model are known as errors or inversions. For example, two items are indexed 1 and 2, and item 1 is less difficult than item 2. Three possible item response patterns would be consistent with expectation as defined by the

Guttman model: [0, 0], [1, 1], and [1, 0]. A response pattern of [0, 1] for this pair of items would represent an error according to the Guttman model. Several CTT person-fit methods that will be discussed later in this paper use the Guttman model (e.g., Karabatsos, 2003; Meijer & Sijtsma, 2001).

Causes and manifestations of aberrant response patterns.

Although cheating is one example of an examinee behavior that can lead to aberrant response patterns, it is not the only examinee behavior that can lead to aberrant response patterns. In his introductory article on person-fit, Meijer (1996) discussed seven distinct examinee behaviors that may cause aberrant response patterns and how these behaviors may manifest themselves in examinees' response vectors. The examinee behaviors that Meijer identified included sleeping, guessing, cheating, plodding, alignment errors, extreme creativity, and deficiency of subabilities.

Meijer (1996) describes sleeping behavior as an examinee with a poor test-taking strategy who does not take time to carefully check answers to some of the easier items on the test, which may result in a higher-than-expected number of mistakes on some of the easier items on the test. When guessing occurs, an examinee with a low ability level guesses blindly on medium-to-difficult items on the test. Examinees who engage in guessing may get a high proportion of easy items correct, whereas the proportion of correct answers on more difficult items may be close to the inverse of the number of response options (e.g., 0.25 for 4-option multiple choice items). Meijer describes a possible cheating scenario in which a low-ability examinee gets a high proportion of easy items correct because the difficulty of each of those items is less than or equal to the examinee's ability level. However, the examinee struggles with the medium-difficulty items and answers most of them incorrectly. After making a futile attempt to answer the

medium-difficulty items correctly, the examinee resorts to looking off of a high-ability neighbor's answer sheet to respond to the most difficult items on the test, resulting in a response pattern where many easy items are answered correctly, many medium-difficulty items are answered incorrectly, and many difficult items are answered correctly. Plodding behavior occurs when an examinee works very slowly and methodically, refusing to move on to the next item until the examinee is satisfied that the current item is answered correctly. A plodding examinee may have a high ability level, but this may not be reflected in the examinee's NC score, which may be low because the examinee runs out of time before having an opportunity to answer a significant portion of items on the exam. Meijer states that plodding behavior may result in perfect Guttman patterns, or patterns that are "too good to be true," but the overall NC score is not representative of the examinee's ability level. Alignment errors occur when an examinee uses a separate answer sheet to record responses. The examinee elects to skip a particular item but fails to adjust his or her responses on the answer sheet accordingly, thus recording a response for item j on the line for item $j - 1$, for example. Meijer defines extremely creative examinees as high-ability individuals who may over-think easy items on the test and consider them "too simple to be true." After choosing an answer to an easy item, the extremely creative examinee may rethink his or her answer and determine that the originally-selected response is too simple to be correct, and choose another, incorrect answer. These high-ability examinees may answer some of the easiest items incorrectly while getting a higher proportion of the more difficult items correct. Finally, Meijer states that if item difficulty levels happen to coincide with particular content areas, examinees with deficiencies in subabilities (i.e., less knowledge of a particular content area on the exam), may have aberrant response patterns (see also Harnish, 1983; Harnish & Linn, 1981; Meijer & Sijtsma, 1995). Meijer uses an example where the 10 easiest items on a 30-item

test are linked to content area A and the 20 most difficult items are linked to content area B. An examinee with deficient knowledge of content area A, but strong knowledge of content area B, may answer a high proportion of the 20 most difficult items correctly but answer a low proportion of the 10 easiest items correctly. Meijer provides examples of response patterns that may result from these various examinee behaviors. These different types of aberrant response patterns, along with the items' population-level proportion-correct values (π_j) are illustrated in Table 1.

It is important to note that the example response patterns provided by Meijer (1996) assume that the exam's items are presented to examinees in ascending difficulty order, with the easiest items appearing at the beginning of the exam and the most difficult items appearing at the end. Except in circumstances of (1) speeded exams that are designed such that most examinees will not be able to provide answers on all items in the allotted time limit, or (2) very poorly constructed exams, such an exam design is unlikely to be used in real-world professional testing applications. This unorthodox assumption of items being presented in order of ascending difficulty somewhat limits how Meijer's discussion of how these behaviors generalizes to a more common exam design where items are not presented in order of difficulty. However, some of Meijer's conclusion can be generalized to other testing contexts.

Regardless of the how items are presented on an exam relative to their difficulty, Meijer's description of how cheating behavior manifests itself in a response vector remains plausible. When describing the characteristics of a cheater, a relatively low ability level is a safe assumption—an examinee with a high ability level is unlikely to need to resort to cheating in order to achieve a high score on the exam. It is also reasonable to assume that a cheater may get a high proportion of the easiest items on the exam correct, even without the aid of cheating,

because some of the easiest items on the exam will have difficulty levels at or below the examinee's ability level. Finally, if the cheating examinee has access to correct answers on some of the most difficult items on the examination, this examinee likely will get more of these difficult items correct than would be expected given the examinee's ability level. In Meijer's example scenario, the cheating examinee answers the initial (easy) items correctly, answers some of the moderately-difficult items near the middle of the exam incorrectly, and resorts to looking at another examinee's answer sheet at the end of the exam in an effort to achieve a higher test score. Cheating behavior that involves one examinee looking off of another examinee's answer sheet—with or without the other examinee's consent—is typically referred to as collusion, and much of the research on detecting collusion on exams has focused on assessing the similarity of particular aspects of response patterns (e.g., choosing a higher-than-expected percentage of the same distractors across items) among pairs of respondents (e.g., Angoff, 1974). In a more common test design where items are not presented in difficulty order, collusion may be less likely to cause a response pattern such as this to emerge, unless the cheater resorts to looking at a neighbor's answer sheet on only the most difficult items. However, there are other forms of cheating that may result in aberrant response patterns.

Exposed items.

Computer-based testing has made collusion much more difficult for examinees who hope to cheat. Security measures such as administering multiple forms at a single site and scrambling the presentation order of items and response options are effective collusion deterrents, and they are easy to accomplish in computer-based testing with capable software. However, test takers looking to get an unfair advantage on an exam still have other cheating methods at their disposal. Some examinees may bring a cheat sheet or other resource when they sit for the exam. Other test

takers may gain access to compromised test questions by participating in unfair or illegal "test preparation" courses or by purchasing stolen test questions.

Of course, not all test preparation services operate in an unethical manner. Many reputable test preparation services perform services that are both helpful to their customers and within the limits of the law. Legitimate test preparation courses teach to the subject matter covered by the examination. Using resources such as the exam's published test specifications, the test preparation organization creates an educational curriculum designed to address topics and areas that appear on the exam. Although the educational curriculum may place most of its emphasis on particular topics that are addressed on the exam's specifications, the focus of the education is on the underlying concepts being tested by items on the exam—not the exam's items themselves.

However, other purported test preparation services operate illegally—using illicit means to gain inside information about the exam. Illegitimate test preparation courses teach directly to the test. Rather than focusing on underlying topics that are covered by the test specifications, they sell illegally-obtained live test items to test takers. Stolen items that are used for these purposes are referred to as exposed (or compromised) items. The methods used to steal the items vary. Some organizations collaborate with test takers and ask them to relay whatever information they can recall about the exam and its content. Sometimes, test preparation organizations send their own employees to the testing site for the sole purpose of memorizing as many items as possible. Regardless of how the stolen items were obtained, both of these forms of cheating—using cheat sheets while taking the exam or taking an illegal test preparation course prior to sitting for the exam—can result in certain suspicious outcomes for examinees who use these

methods to gain an unfair advantage. When examinees have access to exposed items prior to taking the exam, their response patterns may be aberrant.

Effects of aberrant responses on ability estimation in IRT.

In the following hypothetical scenario, two different examinees take a five-item test. The items on the test have item parameters listed in Table 2 and item response functions shown in Figure 5.

Based on the parameters for these five example items, items 1-5 can be described as generally being sorted in ascending difficulty (although some of the item response functions do intersect, it occurs at very low levels of θ). In parametric item response models, the likelihood function for an examinee's θ parameter can be computed using the formula,

$$L = \prod_{j=1}^{J} P_j^{X_j} Q_j^{1-X_j}. \tag{11}$$

where P_j represents the probability of a correct response to item j and Q_j represents the probability of an incorrect response for item j, or $1 - P_j$. As discussed before, the dichotomous item response, denoted X_j, is coded 1 for correct responses and 0 for incorrect responses. It can be seen that when item j is coded as correct, the outcome is $P_j^1 Q_j^{1-1} = P_j^1 Q_j^0 = P_j$, and when item j is coded as incorrect, the outcome is $P_j^0 Q_j^{1-0} = Q_j^1 = Q_j$. Therefore, the likelihood function is the product of P_j for all items answered correctly and Q_j for all items answered incorrectly. The maximum likelihood estimate (MLE) of θ, $\hat{\theta}$, occurs at the maximum of this function—or the point at which the first derivative of the function equals 0.

Because possible values of P and Q range between 0 and 1, multiplying P and Q across many items will result in extremely small products, which a computer may have difficulty representing accurately, so a log-likelihood function can be used instead to estimate $\hat{\theta}$ and avoid

computational issues with handling extremely small numbers. The log-likelihood function is defined as

$$\ln L = \sum_{j=1}^{J} \ln \left[P_j^{X_j} Q_j^{1-X_j} \right]. \quad (12)$$

Using the five-item test from Table 2, it can be easily illustrated how different combinations of item responses affect properties of the MLE of an examinee's θ parameter. For example, two different examinees take this five-item test. Both examinees achieve number-correct scores of 3 on the test, but examinee 1 answers the three easiest items correctly (items 1-3), while examinee 2 answers the three most difficult items correctly (items 3-5). As will be shown shortly, the response pattern for examinee 1 is consistent with expectation under the IRT model, while the response pattern for examinee 2 is contrary to expectation, or aberrant. A log-likelihood function for each examinee can be created using formula (12).

The log-likelihood function for the non-aberrant response string belonging to examinee 1 is presented in Figure 6. As illustrated in this figure, a non-aberrant response vector results in a log-likelihood function that is peaked at the MLE for θ, with fairly sharp drop-offs in the function at other conditional values of θ.

The log-likelihood function for the aberrant response string belonging to examinee 2 is shown in Figure 7. Both examinees in this example answered three items correctly, but they differed in which items were answered correctly. As illustrated in these figures, examinees with aberrant response patterns have log-likelihood functions that are flatter than log-likelihood functions computed from non-aberrant response vectors. This example also shows that it can sometimes be unclear where the best estimate of the examinee's θ level lies when a response

string is aberrant. In Figure 7, a local maximum occurs near the vicinity of $\theta = 0.5$, but the log-likelihood function is higher at large negative values of θ.

This simple example illustrates basic maximum likelihood estimation of θ in IRT, but there are also other parameter estimation techniques that make use of Bayesian methods and mix the log-likelihood function with a prior distribution to form a posterior distribution of θ (e.g., expected *a posteriori* [EAP], modal *a posteriori* [MAP]; Hambleton et. al 1991). Regardless of which particular estimation method is used, the presence of aberrant response patterns causes the function that is used to estimate ability to flatten when compared to a function computed from a data set with no aberrance present.

Methods for Assessing Person-Fit

Model-fit is an integral and well-known component to statistical modeling. In factor analysis, for example, model-fit indicates how well the covariance matrix implied by the factor analysis model reproduces the observed covariance matrix obtained from the sample (Brown, 2006; McDonald, 1999). When model fit is good, the implied covariance matrix is highly similar to the observed covariance matrix, and when model fit is poor, the implied covariance matrix is not very similar to the observed covariance matrix. In a broader sense, model-fit indicates the extent to which expectations as outlined in a particular statistical model align with the reality of what was observed in a sample. It is noteworthy that model fit provides a single indicator of the adequacy of the model for the entire sample.

The concept of person-fit, as its name implies, is similar to model-fit in that it provides an indication of the extent to which characteristics of observed data conform to expectations defined by a model; however, model-fit and person-fit differ greatly in scope. A model-fit statistic indicates how well a model fits the aggregated data. A person-fit statistic indicates how

reasonable one individual's response pattern is, given some expectation—to be defined shortly. As mentioned before, a very large number of person-fit statistics have been developed, and person-fit statistics have been developed to be applied in a number of measurement frameworks. Because of the vast number of person-fit statistics in existence, not every person-fit statistic that has been developed and published will be discussed in this paper. Rather, each of the major groups of person-fit statistics will be discussed, and several of the most common and popular techniques within each group will be discussed. The two most commonly researched and applied classes of person-fit statistics are group-based techniques and IRT-based techniques. Also discussed in this paper are less commonly researched and applied—but relevant—techniques based on factor analysis and Bayesian estimation.

Group-based person-fit.

As previously mentioned, all person-fit statistics more or less can be summarized as establishing the reasonableness of an individual's response pattern by comparing the individual's response pattern to some expected outcome. How the basis of expectation is formed is what separates the different types of person-fit statistics. Group-based person-fit statistics define expectation based on aggregate-level item characteristics estimated from the overall sample (Karabatsos, 2003). In general, group-based person-fit statistics classify response patterns as aberrant when items with proportion-correct (p) values near 0 tend to be answered correctly and items with p near 1 tend to be answered incorrectly (Meijer & Sijtsma, 2001). Of course, there is considerable variability in how each of these group-based person-fit statistics are calculated, but all involve comparing individual response patterns to response patterns belonging to other examinees to some extent—although some researchers have also recommended repeated-

measures comparisons as well (Tatsuoka & Tatsuoka, 1983). Several of the most common group-based statistics are discussed in the following sections.

The personal point-biserial correlation.

In the overview of classical test theory item analysis, the item-total correlation was discussed as an indicator of an item's discrimination. A positive item-total correlation for a particular item indicates a positive relationship between item performance and overall exam performance—an outcome that is desirable in testing. Conceptually, the personal point-biserial correlation coefficient (r^*_{pb}) can be thought of as the transpose of the traditional item-total correlation coefficient. With examinees represented as rows and items represented as columns in the data matrix, the item-total correlation represents the correlation between a particular column (i.e., responses for a single item) and a column comprised of row sums (i.e., total exam scores). Conversely, the personal point-biserial correlation coefficient represents the correlation between a row (i.e., an examinee's complete response string) and a row comprised of items' proportion-correct values (Karabatsos, 2003; Meijer & Sijtsma, 2001). The term point-biserial correlation is used with this particular statistic because it was originally developed for use with tests comprised of dichotomous item responses.

The personal point-biserial correlation coefficient's interpretation is relatively simple and straightforward. Negative values of r^*_{pb} indicate aberrant response patterns and positive values indicate response patterns that are consistent with expectation. Being a correlation coefficient, possible values for the coefficient range from +1 to -1. However, the actual range of observed values for this statistic will be attenuated to some extent when dichotomous item responses are used to calculate the correlation coefficient.

The caution index.

The personal point-biserial correlation coefficient holds a certain intuitive appeal: even psychometricians who are new to person-fit are familiar with the item-total correlation, and calculating a correlation between a string of item responses and a string of proportion-correct values makes sense. However, the issue of attenuation is a limitation. The caution index (*C*; Sato, 1975) is similar to the personal point-biserial correlation coefficient in that the covariance between item responses and item difficulty indicators is estimated, but the caution index also incorporates Guttman pattern information into the formula. The caution index can be calculated as one minus the ratio of two covariances,

$$C = 1 - \frac{\text{cov}(\mathbf{X}, \mathbf{p})}{\text{cov}(\mathbf{X}^*, \mathbf{p})}, \tag{13}$$

where $\mathbf{X}$ is the observed response vector for a given examinee, $\mathbf{X}^*$ is a response vector containing correct responses only for the easiest *r* items, and $\mathbf{p}$ represents the items' proportion-correct vector. When $\mathbf{X}$ is a perfect Guttman vector, $\mathbf{X} = \mathbf{X}^*$, which results in an estimated value of 0 for *C*. As $\mathbf{X}$ becomes less like a perfect Guttman vector, *C* becomes larger (Meijer & Sijtsma, 1995). Although the caution index has a lower bound of 0, it has no upper bound. This limitation has led to the development a variation on this person-fit statistic: the modified caution index.

The modified caution index.

The modified caution index (C^*) alters the previously-discussed caution index to be calculated as

$$C^* = \frac{\text{cov}(\mathbf{X}^*, \mathbf{p}) - \text{cov}(\mathbf{X}, \mathbf{p})}{\text{cov}(\mathbf{X}^*, \mathbf{p}) - \text{cov}(\mathbf{X}', \mathbf{p})}, \tag{14}$$

where $\mathbf{X}'$ is a reverse Guttman vector containing correct responses only for the most difficult r items. This modification to the formula has the effect of limiting both the upper and lower limits of the statistic. Once again, when $\mathbf{X}$ is a perfect Guttman vector, $\mathbf{X} = \mathbf{X}^*$, which causes the numerator (and therefore the value of C^*) to equal 0. When $\mathbf{X}$ is a perfect reverse Guttman pattern, $\mathbf{X} = \mathbf{X}'$, which results in the numerator and denominator being equal, thus resulting in C^* equaling 1.

The H^T statistic.

The H^T statistic (Sijtsma, 1986; Sijtsma & Meijer, 1992) is another group-based person-fit statistic, but unlike C and C^*, H^T is not normed against the Guttman pattern (Meijer & Sijtsma, 2001). The H^T statistic for examinee a is calculated as

$$H^T = \frac{\sum_{a \neq b} \sigma_{ab}}{\sum_{a \neq b} \sigma_{ab}^{\max}}. \tag{15}$$

As shown in equation (15), the observed covariances between all pairs of examinees (indexed a and b here) are calculated and summed, and that value is divided by the sum of the maximum possible covariances between all pairs of examinees. For an exam comprised entirely of dichotomously-scored items, let β_a represent the proportion of items answered correctly by examinee a. Let β_{ab} represent the proportion of items answered correctly by both examinee a and examinee b. The covariance of the two examinees' response strings can then be computed as

$$\sigma_{ab} = \beta_{ab} - \beta_a \beta_b, \tag{16}$$

and assuming that examinee indices $a < b$ imply that $\beta_a < \beta_b$, the maximum covariance between the two response vectors is

$$\sigma_{ab}^{\max} = \beta_a (1 - \beta_b). \tag{17}$$

Like the personal point-biserial correlation coefficient, a positive value of H^T is indicative of a response vector that is consistent with other examinees' response vectors and a negative value is indicative of an aberrant response vector.

In addition to person-fit assessment, H^T has also been applied as a statistical test to determine whether or not nonparametric item response functions intersect (e.g., Sijtsma & Junker, 1996; Sijtsma & Meijer 1992). Such applications demonstrate the utility of this statistic, but a detailed discussion of H^T applied for that purpose is beyond the scope of the current paper.

IRT-based person-fit statistics.

As previously discussed in the overview of item response theory, the item response function relates the probability of a correct response at conditional ability levels. If person-fit is conceptualized as the extent to which an examinee's response vector conforms with expectation, then the item response function provides a very useful tool to use in evaluating person-fit. Recall that the item response function represents the probability of a correct response at conditional values of θ. As an example, assume that a particular item on an exam has the following item parameters: $a = 1.70$, $b = 0.00$, $c = 0.00$, and a particular examinee has an ability level of $\theta = 2.00$. In this example, the examinee answers this item correctly. The item response function, with the examinee's response also plotted, is illustrated in Figure 8.

Treating the line plotted by the item response function as the expected outcome and the point representing the response as the observed outcome, it is easy to see that for this particular item, a correct response is quite reasonable for an examinee with an ability level of $\theta = 2.00$. As shown in Figure 9, it would be far less likely for an examinee with a lower ability level of $\theta = -2.00$ to answer this same item correctly, as indicated by the distance between the plotted item response and the probability of success.

Although not presented in graphical form here in this paper, it would also be easy to demonstrate that an incorrect response to this item would fall in line with expectation for an examinee with an ability level of θ = -2.00, and an incorrect response to the same item would be highly unexpected for an examinee with an ability level of θ = 2.00. Of course, actual person-fit assessment in IRT involves additional steps than what has been presented so far, but these simple examples illustrate how the item response function makes IRT a useful measurement framework to use when conducting person-fit assessment. The small selection of group-based person-fit statistics that were reviewed in the previous section used several different methods to establish the expected performance criterion: item difficulty, Guttman patterns, and consistency with other examinees' response vectors. Item response theory has the expectation criterion component built in already.

IRT has proven to be a popular measurement framework in the person-fit research literature (e.g., Karabatsos, 2003), with a number of person-fit statistics having been developed for use with IRT. The log-likelihood (l_0) person-fit statistic and the standardized log-likelihood statistic (l_z), especially, have proven to hold enduring interest in the person-fit research literature. Those person-fit statistics will be discussed in next section.

The l_0 and l_z statistics.

The l_0 person-fit statistic (Levine & Rubin, 1979) is calculated using the formula,

$$l_0 = \sum_{j=1}^{J} \{X_j \ln [P_j(\theta)] + \left(1 - X_j\right) \ln\left[1 - P_j(\theta)\right]\}. \quad (18)$$

This formula is a slightly altered version of formula (12), which was used to create the log-likelihood function of the examinee's θ parameter for the purpose of parameter estimation. As with formula (12), $\ln[P_j(\theta)]$ is calculated for correctly-answered items, and $\ln[1 - P_j(\theta)]$ is

calculated for incorrectly-answered items. Recall that $1 - P_j(\theta)$ in formula (18) corresponds to Q_j in formula (12). These values are then summed across all items to calculate l_0. Aside from minor differences in notation, the only major difference between formula (12) and formula (18) is that formula (12) is calculated across a range of conditional θ values in order to create a function (see Figure 6), while in formula (18), $P_j(\theta)$ and $1 - P_j(\theta)$ are computed at a single fixed value of θ*:* the examinee's MLE, or $\hat{\theta}$, thus resulting in a scalar value of l_0, as opposed to a function. Therefore, the l_0 person-fit statistic can be described as the logarithm of the likelihood function, computed at the examinee's MLE of θ (Meijer & Sijtsma, 2001). In their 1996 article, Drasgow, Levine, and Zickar assert that person-fit statistics based on information obtained from likelihood ratio tests are optimal from both the perspectives of Type I error and power. Some authors have classified IRT-based person-fit statistics into residual-based statistics and likelihood-based statistics: the l_0 person-fit statistic is one that fits into both of these categories (Kogut, 1986).

The l_0 statistic provides an intuitive basis for assessing person-fit. Because both $P_j(\theta)$ and $Q_j(\theta)$ range between 0 and 1, taking the natural logarithm of any value between the upper and lower limits of the range will result in a negative value. Referring back to the example shown in Figure 8, when observed item responses are congruent with expected outcomes, $P_j(\theta)$ or $Q_j(\theta)$ is a value close to 1. Taking the natural logarithm of a number close to 1 will result in a value that is close to 0. In an ideal scenario where $P_j(\theta)$ and $Q_j(\theta)$ are equal to values extremely close to 1 for all items that are answered correctly or incorrectly, respectively, l_0 approaches its maximum possible value of 0. Conversely, when observed item responses are incongruent with expectation, $P_j(\theta)$ or $Q_j(\theta)$ is a value closer to 0, which is illustrated in Figure 9. Taking the natural logarithm of a value that is near 0 results in a larger, negative value.

Drawing together what has been previously discussed, it becomes clear why the l_0 statistic provides a sensible method for assessing person-fit. The more an examinee's responses tend to be congruent with expectation, the closer l_0 will be to 0. An examinee with responses that are incongruent with expectation will have an l_0 statistic that is larger, in terms of absolute value, and negative. Although the l_0 statistic is a scalar value and not a function, this discussion is consistent with what was previously noted about how log-likelihood functions differ overall for aberrant and non-aberrant response patterns. As illustrated in Figure 6, a log-likelihood function created from a non-aberrant response pattern will tend to be peaked, while a log-likelihood function created from an aberrant response pattern (as illustrated in Figure 7) will tend to be flatter. The maxima of the two functions will also vary, with a function associated with a non-aberrant response pattern reaching a maximum closer to 0 than the maximum of a function created from an aberrant response pattern. By measuring the height of the log-likelihood function at $\hat{\theta}$, the l_0 statistic provides an intuitive method of assessing the level of aberrance associated with a vector of responses. However, the l_0 statistic has serious limitations that reduce its applicability.

The l_0 person-fit statistic has two major limitations that have been identified in the research literature (Kogut, 1988; Meijer & Sijtsma, 2001; Molenaar & Hoijtink, 1990). First, it is not standardized. As a result of this limitation, classification of a particular response pattern as model-fitting or misfitting depends on θ. Second, in order to classify response patterns as misfitting, a distribution of l_0 under the null hypothesis is required, but the null distribution of l_0 is unknown.

In response to these limitations, Drasgow, Levine, and Williams (1985) developed a standardized version of the l_0 statistic: l_z. This statistic is computed

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{\text{var}(l_0)}}, \tag{19}$$

where $E(l_0)$ and var(l_0) are the expectation and variance of l_0, respectively:

$$E(l_0) = \sum_{j=1}^{J} \{P_j(\theta) \ln[P_j(\theta)] + [1 - P_j(\theta)] \ln[1 - P_j(\theta)]\}, \tag{20}$$

and

$$var(l_0) = \sum_{j=1}^{J} P_j(\theta)[1 - P_j(\theta)] \left[\ln \frac{P_j(\theta)}{1 - P_j(\theta)}\right]^2. \tag{21}$$

The l_z statistic was initially purported to be asymptotically standard normally distributed (Drasgow et al., 1985). Because l_z was believed to be standard normal, it was originally thought that it could be interpreted like a *z* score, with decisions regarding the relative aberrance of response patterns based on values from a *z* score table. However, researchers have since disputed this conclusion. Van Krimpen-Stoop and Meijer (1999) found evidence that l_z is not normally distributed when used with computer adaptive tests. Several other studies have shown that l_z may not be normally distributed when $\hat{\theta}$ is substituted for θ in computing the statistic (Nering, 1995; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999) and its variance may be underestimated as well (Snijders, 2001). Nering reports that the l_z statistic is closer to being normally distributed when θ is more accurately estimated. However, the skewness and kurtosis of the statistic's sampling distribution were found to be generally problematic, as was the Type I error rate associated with cut-off values obtained from a *z* table. For these reasons, Nering recommends using empirically-derived cutoff values when assessing person-fit using l_z, rather than uniformly applying a *z* table cut in all circumstances.

In their 2007, study, Armstrong, Stoumbous, Kunk, and Shi found that detection power may also be greatly reduced when l_z is calculated using the MLE of θ. This finding is not altogether surprising, because the examinee's item responses are used in estimating $\hat{\theta}$. Therefore, cheating on the exam is likely to bias the examinee's MLE of θ, which will also affect the estimate of l_z. Of course, this problem is not limited to the l_z person-fit statistic. Indeed, any person-fit statistic that makes use of some estimate of examinee ability from the empirical data is going to be affected by characteristics of the observed data (e.g., Emons, Meijer, & Sijtsma, 2002). Ironically, Brown and Villarreal (2007) advocate using l_z as a weighting function to correct biased estimates of ability that may arise when aberrant response patterns are present. Other research on this statistic has suggested employing corrections to the MLE of θ for unreliability prior to computing the l_z statistic (de la Torre & Deng, 2008).

IRT-based person response functions.

The item response function was previously introduced as an IRT plot relating the probability of success on a given item to examinee ability. At the conceptual level, the person response function (PRF) can be thought of as the transpose of the item response function. As shown in Figure 1, the IRF illustrates the probability of a correct answer as a function of θ and fixed item parameters, while the PRF illustrates the proportion of correctly-answered items as a function of item difficulty and a fixed person parameter (Emons, Sijtsma, & Meijer, 2004; Ferrando & Lorenzo, 2000; Sijtsma & Meijer, 2001). The PRF can also be applied to polytomous data obtained from personality inventories, for example (Ferrando, 2004).

A PRF is plotted using observed item responses. The exam's J items are sorted in ascending difficulty order, $(1 - \pi_1) < (1 - \pi_2) < \ldots < (1 - \pi_J)$, and then grouped into G strata, with an individual stratum denoted A_g (indexed $g = 1, 2, \ldots, G$). Each stratum consists of $m = J / G$

items (or some variant if J / G yields a remainder). The proportion of correctly-answered items within each stratum is plotted in a line chart. Emons, Sijtsma, and Meijer (2004) have shown that when an IRT model meets four important assumptions: (1) unidimensionality, (2) local item independence, (3) monotonicity, and (4) non-intersecting item response functions, the plotted person response function is expected to be a nonincreasing function. The first three assumptions are common assumptions for parametric IRT models, but the fourth assumption (non-intersecting item response functions) is met in parametric IRT only in the case of the Rasch (1-PL) IRT model, which constrains all IRF slopes to equality and all IRF lower asymptotes to equal 0. Because the IRFs in the Rasch model do not intersect, when comparing two items, $b_1 > b_2$ implies that $P_1(\theta) > P_2(\theta)$ for all θ. In addition to the Rasch model, nonparametric IRT models that meet the double monotonicity assumption can be used for assessing person-fit using PRFs as well (Emons Sijtsma, & Meijer, 2004). Person response functions that significantly depart from a nonincreasing appearance (i.e., PRFs that go back up for the strata that contain more difficult items) provide visual evidence of aberrance in the response vector.

Two examples of person response functions, adapted from Emons, Sijtsma, and Meijer (2005), are shown in Figure 10 and Figure 11. Both figures are examples of discrete person response functions. The term discrete is used here because some researchers have advocated taking an additional step and applying kernel smoothing to the discrete person response function as well (e.g., Emons, Sijtsma, & Meijer, 2004, 2005). No smoothing is applied to the person response functions illustrated in this paper.

Both Figure 10 and Figure 11 illustrate person response functions for examinees who took a test with items that were grouped into 9 strata. Figure 10 is an example of a person response function generated from a non-aberrant response vector. The person response function

is monotonically decreasing, indicating that the examinee's performance declines as the difficulty of the items increases. This outcome is consistent with expectation, which will be explained in greater detail shortly.

Figure 11 is an example of a person response function that was generated from an aberrant response vector. In this example, the examinee's performance steadily declines across the first 7 strata, but performance improves for strata 8 and 9—the strata comprising the most difficult items on the exam—which is contrary to expectation.

Although a visual inspection of an examinee's observed performance across strata may provide useful diagnostic information regarding person-fit, a comparison of observed performance to some representation of expected performance would likely prove helpful in assessing the level of a response vector's aberrance. A clearly aberrant person response function, such as the example illustrated in Figure 11, is quite easy to identify without any comparison to an expected level of performance necessary. However, assessing a PRF for a response vector that is less obviously aberrant becomes problematic when relying on a visual interpretation of observed performance alone. For example, in Figure 11, the examinee answered 60% of items in stratum 8 correctly and 70% of items in stratum 9—an obvious departure from the percentage of correct answers observed in stratum 7 (40%) and the observed downward trend across strata 1-7. However, if the examinee had instead answered 45% of items in stratum 8 and 50% of items in stratum 9 correctly, the final determination regarding the aberrance of the response vector becomes less clear. Although the person response function would indeed no longer be monotonically decreasing in that situation, judging whether or not the PRF significantly diverges from expectation enough to indicate a problematic level of response vector aberrance is difficult

without additional information. Therefore, it becomes necessary to define a level of expected performance for each stratum to serve as a basis for comparison for observed performance.

According to Emons, Sijtsma, and Meijer (2004), the expected proportion of correct answers for a given examinee to items in stratum A_g is equal to

$$\tau_g = m^{-1} \sum_{j \in A_g} P_j(\theta). \tag{22}$$

Assuming an IRT model with no intersecting item response functions (e.g., the Rasch model) is being used, ordering items in ascending difficulty order implies that for each examinee,

$$m^{-1} \sum_{j \in A_g} P_j(\theta) \geq m^{-1} \sum_{j \in A_{g+1}} P_j(\theta), \tag{23}$$

for all θ. For the *G* strata, it also follows that

$$\tau_1 \geq \tau_2 \geq \cdots \geq \tau_G, \tag{24}$$

for all θ. Therefore, a nonincreasing person response function is expected, with stratum-level expected values computed using equation (22).

Comparing observed and expected proportion-correct scores within each stratum allows for a detailed visual person-fit assessment to be conducted. If an observed PRF is found to increase over one or more strata, for example, the observed proportion correct values for the strata can be compared to their respective expected values. Expected values across strata are not necessarily expected to fall into a neat, 45-degree downward slope; the only constraint imposed on expected values is that they are nonincreasing, but the rate at which expected values decrease may change across some strata, depending on the characteristics of the items and the examinee's ability level. Comparing observed proportion-correct values to expected proportion-correct values across strata provides additional valuable information when conducting a visual person-fit assessment of an examinee's response vector.

A related technique that has been used to assess person-fit in conjunction with person response functions incorporates multilevel logistic regression (e.g., LaHuis & Copeland, 2007; Reise, 2000; Woods, 2008). As described by Woods (2008), with this method item- and person parameters are estimated from a set of item responses using an IRT model. Next, item responses are regressed onto item difficulty,

$$\begin{aligned} \log\left(\frac{p_{ji}}{1-p_{ji}}\right) &= b_{0i} + b_{1i}\beta_j \\ b_{0i} &= \gamma_{00} + u_{0i} \\ b_{1i} &= \gamma_{10} + u_{1i}, \end{aligned} \tag{25}$$

where the first row in (25) represents Level 1 of the two-level logistic regression model and the second and third rows represent Level 2. Examinees are indexed *i* and items are once again indexed *j*. In equation (25), p_{ji} represents the probability that examinee *i* correctly answers item *j*, β_j is the difficulty parameter for item *j*, γ_{00} and γ_{10} are Level 2 intercepts, and u_{0i} and u_{1i} are error terms.

Woods (2008) explains that significant heterogeneity in the person response function slope (b_{1i}) is taken as an indication that person-fit varies significantly over individuals and may be poor for some. If this outcome is found, Woods recommends employing empirical Bayes methods (Snijders & Bosker, 1999, pp. 58-63) to estimate individual person response function slopes.

Each of the previously-discussed person-fit statistics provides a single scalar value for use in assessing the aberrance of an examinee's response vector. The added benefit of using a person response function is that it provides several pieces of person-fit information across the entire range of examinee's response string. Nering and Meijer (1998) recommend using a traditional scalar person-fit statistic as an initial screening tool to flag aberrant response patterns,

and then creating person response functions for flagged examinees for diagnostic purposes. Considering the item response pattern characteristics thought to coincide with various test-taker behaviors (see Table 1; Meijer, 1996), the researcher can use a person response function to gain insight into possible explanations as to why a particular examinee's response pattern was classified as aberrant, thus saving time and resources by targeting follow-up only for examinees that have response patterns that (1) are flagged as aberrant and (2) match a particular pattern of interest.

A related concept that makes use of person response functions is differential person functioning (DPF). Much like a person response function can be conceptually thought of as the transpose of an item response function, differential person functioning can be conceptually thought of as the transpose of differential item functioning (Johanson & Alsmadi, 2002). Differential item functioning (DIF) occurs when two or more groups of examinees who are matched on ability are found to have differential probabilities of success on a particular item. For example, groups may be formed based on some demographic variable that is of interest, and then examinees are matched across groups based on their ability levels. If the item appears to be significantly easier for one of the matched groups, the item is exhibiting DIF. Differential person functioning assessment is similar to DIF assessment, with the main difference being that assessment is conducted for one examinee at a time, and performance is assessed across different groups of items.

As described in Johanson and Alsmadi (2002), much of the methodology used to assess DPF is identical to what was previously described regarding the construction of person response functions. Once again, items are ordered by difficulty and then grouped into *G* strata. However, DPF assessment includes an additional step where items are broken down into two groups: the

focal group and the reference group. The groups may be comprised of particular content domains or perhaps item types (e.g., recognizing or recalling concepts vs. applying them). Items from each group are matched on difficulty and placed into *G* strata. A person response function may be plotted, with a different line plotted for each group of items. A technique such as this is especially useful for performing a visual assessment of aberrance due to subabilities (Meijer, 1996).

Factor-analytic techniques for assessing person-fit.

Although classical test theory and item response theory are by far the most popular measurement frameworks in the person-fit research literature, a small number of studies have explored aberrant response pattern detection using factor analysis. Reise and Widaman (1999) proposed an FA-based person-fit statistic that is based on partitioning overall model fit down to the individual's level. At the model level, the fit function can be related to the log-likelihood of the model parameters given the data,

$$LL = -(N \times F_{ML}) \,/\, 2. \tag{26}$$

The authors go on to state that the log-likelihood value can be partitioned to determine a given examinee's contribution. Equation (27) gives the log-likelihood of the model at the level of the examinee:

$$P_{LL} = -\frac{1}{2}[J \ln(2\pi) + \ln|\mathbf{\Sigma}^*| + (\mathbf{X} - \mathbf{M})\mathbf{\Sigma}^{*-1}(\mathbf{X} - \mathbf{M})], \tag{27}$$

where $\mathbf{\Sigma}^*$ is the reproduced covariance matrix, and $\mathbf{M}$ is a vector of sample means. The left side of the equation is constant for all individuals, and the right side is the Mahalanobis squared distance formula and varies across examinees (Comrey, 1985; Reise & Widaman, 1999).

In discussing this statistic, Reise and Widaman (1999) report that small negative values indicate good person-fit, while large negative values indicate poor person-fit, but a direct

conditional standardization of this statistic is not offered. However, the authors offer a method that is more objective and standardized than simply comparing values of P_{LL} among examinees. Using equation (27), values of P_{LL} are computed for both a saturated and substantive model, and then -2 is then multiplied by the difference in log likelihoods between the two models. The authors interpret this value as the individual examinee's contribution to the overall model chi-square, which the authors denote IND_{CHI}. Large, positive values of IND_{CHI} are indicators of individuals whose response vectors are making larger contributions to overall model misfit.

Ferrando (2007) proposed person-fit techniques that are described in his article as factor-analytic counterparts to the familiar IRT-based l_0 and l_z statistics. The context of the study was assessing individual response vectors comprised of J personality items. In the article, Ferrando states that a single latent trait is assumed to underlie an instrument made up of graded response format items. As previously shown, an individual's response to item j can be represented in the factor analytic framework as a sum of item difficulty (μ_j), the product of the item's factor loading, (λ_j) and the examinee's factor score (θ), and the unique component (ε_j). According to Ferrando, the conditional distribution of responses for a fixed θ is assumed to be normal, with a mean $\mu_j + \lambda_j\theta$ and variance $\sigma^2_{\varepsilon_j}$. The *lco* scalability index, which Ferrando described as the factor-analytic counterpart to IRT's l_0 person-fit index, is calculated,

$$lco = \sum_{j=1}^{J} \left[\frac{X_j - \mu_j - \lambda_j\theta}{\sigma_{\varepsilon_j}} \right]^2, \tag{28}$$

where the maximum-likelihood estimate of the examinee's factor score is estimated by Bartlett's weighted least-squares formula (Ferrando, 2007),

$$\hat{\theta}(ML) = \frac{\sum_{j=1}^{J} \frac{\lambda_j (X_j - \mu_j)}{\sigma_{\varepsilon_j}^2}}{\sum_{j=1}^{J} \frac{\lambda_j^2}{\sigma_{\varepsilon_j}^2}}. \quad (29)$$

The lco person-fit statistic could be characterized as a sum of squared naïve standardized residuals (Bollen & Arminger, 1991). Standardized residuals like this have been used for other testing issues as well, such as applications involving response time (Ferrando & Lorenzo-Seva, 2007). When $\hat{\theta}$ is substituted for θ in equation (28), the distribution of individual *lco* values across respondents is expected to be χ^2 with $J - 1$ degrees of freedom (Ferrando, 2007). In comparing *lco* with l_0, Ferrando (2007) notes two important differences between the statistics. First, the *lco* statistic is referred to the χ^2 distribution, whereas the l_0 statistic is asymptotically referred to the normal distribution (due to the central limit theorem, l_0 is expected to approach a normal distribution as the number of items increases). Second, because the l_0 index is a log-likelihood, large, negative values indicate misfit, while *lco* is a minimum chi-square, so large, positive values indicate misfit.

Ferrando (2007) writes that an ideal person-fit index should satisfy three criteria: (1) it should have reference values so it can be interpreted, (2) it should be independent of test length, and (3) it should be independent of trait levels (i.e., able to detect misfitting response patterns equally well at all trait levels). The *lco* index meets conditions (1) and (3), but fails to meet condition (2), because the length of the test changes the sampling distribution of the statistic via changes in degrees of freedom. In response to this limitation, Ferrando (2007) proposed an FA counterpart to the IRT l_z person-fit statistic, *lcz*, which uses a normal approximation to the χ^2 distribution,

$$lcz = \sqrt{2lco} - \sqrt{2J - 3}. \quad (30)$$

The *lcz* statistic is referred to a standard normal distribution, so it is interpreted as a *z* score (Ferrando, 2007). As was the case with the l_0 and *lco* statistics, signs are once again reversed when comparing the l_z and *lcz* statistics. Large, negative values of l_z indicate person-level misfit, and large, positive values of *lcz* indicate misfit.

In his follow-up article, Ferrando (2009) generalized the *lco* and *lcz* person-fit statistics for models with *K* estimated latent factors. The multidimensional *lco* person-fit index, *M-lco*, is computed

$$M - lco = \sum_{k=1}^{K} \sum_{j=1}^{J} \left[\frac{X_j - \mu_j - \lambda_{jk} \hat{\theta}_{mk}}{\sigma_{\varepsilon j}} \right]^2 . \quad (31)$$

Like its unidimensional counterpart, the *M-lco* person-fit statistic follows a χ^2 distribution with *J* – *K* degrees of freedom. Ferrando also developed a standardized version of the *M-lco* statistic,

$$M - lcz = \sqrt{2M - lco} - \sqrt{2(J - K) - 1}. \quad (32)$$

Similar to the unidimensional *lcz* statistic, *M-lcz* is expected to follow a standard normal distribution.

Other methods used to assess person-fit.

In addition to the previously-described techniques that have been used to evaluate person-fit, there are several other methods that have received some attention in the research literature. Although these person-fit techniques are somewhat less common than the group-based and IRT-based person-fit statistics, they warrant a brief mention in a discussion of person-fit assessment.

Bayesian techniques.

As previously mentioned in the discussion of IRT parameter estimation, some IRT estimation routines make use of Bayesian methods by mixing the log-likelihood function with a prior distribution (Hambleton et. al 1991). However, methods such as these rely predominately

on maximum likelihood methods for parameter estimation. The Bayesian component incorporating the prior distribution is included as an additional step that is added to the maximum likelihood estimation procedure by mixing the prior distribution with the log-likelihood function. IRT models can also be fit to data using a fully Bayesian estimation procedure, with numerous draws repeatedly taken from a posterior distribution using a Markov Chain Monte Carlo (MCMC) procedure (e.g., Glas & Meijer, 2003; Hendrawan, Glas, & Meijer, 2005; Kim & Bolt, 2007). Person-fit assessment performed as part of MCMC estimation is incorporated by way of posterior predictive checks.

According to Hendrawan et. al (2005), the posterior distribution of parameters for an IRT model is simulated using the MCMC method. This step estimates the probability of the item parameters given the data, $p(\xi \mid y)$, where ξ represents the person and item parameters and y represents the observed data. Next, person-fit is assessed using a posterior predictive check based on the index $T(y, \xi)$, where T is an IRT-based person-fit statistic. Once the chain has converged, draws from the posterior distribution are used to generate model-conforming data, y^{rep}, and to compute the Bayes p value,

$$\text{Bayes } p \text{ value} = \Pr\left(T(y^{\text{rep}}, \xi) \geq T(y, \xi) | y\right). \tag{33}$$

For every saved iteration, the $T(y, \xi)$ person-fit statistic is computed, a new model-conforming response pattern is generated, and a value $T(y^{rep}, \xi)$ is computed. The Bayes p value is computed as the proportion of iterations where $T(y^{rep}, \xi) \geq T(y, \xi)$.

Cumulative sum statistics.

Cumulative sum (CUSUM) statistics take a slightly different approach than the other common person-fit statistics. Rather than compute a scalar value that is meant to be interpreted as an indicator of the degree of aberrance for an overall response pattern, CUSUM statistics are

designed to detect aberrance that occurs over a segment of an exam. In general, CUSUM statistics work as follows: the statistic is initialized at 0 and accumulates as aberrant responses occur over the course of the exam. When a non-aberrant response is provided, the CUSUM statistic resets back to 0 and begins to re-accumulate again when aberrant responses are provided. Meijer (2002) used a CUSUM procedure that has an upper statistic, C_j^U, and a lower statistic, C_j^L. These statistics are initialized, $C_0^U = C_0^L = 0$, and accumulate across items,

$$C_j^U = \max\{0, C_{j-1}^U + X_j - p_j(\theta)\}, \tag{34}$$

and

$$C_j^L = \min\{0, C_{j-1}^L + X_j - p_j(\theta)\}. \tag{35}$$

Armstrong and Shi (2009) explain that if expectation is defined by an item response model, then non-aberrant responses are represented as $p_j(\theta)$ and the measure of aberrance is $X_j - p_j(\theta)$. Therefore, positive and negative values will accumulate over the course of the exam for non-aberrant examinees, but they will average around 0. Aberrant response behavior would be characterized by a high number of positive or negative deviations over a segment of the test.

Variations to Meijer's (2002) method for computing CUSUM statistics have been developed (e.g., Armstrong & Shi, 2009), but all of the techniques in this family of person-fit indices utilize the same basic approach of accumulating the person-fit statistic over responses. Armstrong and Shi (2009) claim that one potential drawback of traditional person-fit indices is that they do not make use of item sequencing information in their computation. The authors claim that a run of positive (or negative) deviations on the exam can be counter-acted by a similar run of negative (or positive) deviations later on the exam, which will result in an examinee being characterized as non-aberrant if a traditional person-fit statistic is used, whereas a CUSUM method may be more likely to flag the examinee as aberrant on the basis of the string

of concurrent aberrant responses. Of course, in more advanced testing settings, such as computer-based testing with scrambled presentation order, the CUSUM approach may have less applicability. For example, if two examinees of equal ability levels have access to the same subset of exposed items and then they each take scrambled versions of the same form, a CUSUM approach is likely to yield different results for each examinee, depending on how and where the exposed items appear on their respective forms. For this reason, a CUSUM approach is best used on a fixed test form, where order is consistent for all examinees.

Research Methodologies Used by Investigators in Person-Fit Research

Simulation studies.

As expected, much of person-fit research uses simulated data. Using simulated data affords the researcher an opportunity to exercise full control over the characteristics of the data and the response patterns under investigation. Although it is generally true in all circumstances in which a statistic is being investigated or developed that retaining direct control over the data that are used in the investigation is beneficial, incorporating simulated data is especially important in the context of person-fit research.

Meijer (1996) illustrated how different examinee behaviors result in various manifestations in the individual's response vector. However, as Meijer explained, sometimes two different examinee behaviors may result in similar response pattern manifestations. For example, if most of the difficult items on the exam come from one particular content domain and most of the easiest items come from a different content domain, the response pattern obtained from an examinee who has strong knowledge of the difficult content domain but deficient knowledge of the easier content domain may appear very similar to a response string obtained from an individual who cheated on the exam and answered a higher-than-expected proportion of difficult

items correctly. Although both scenarios may ultimately result in aberrant response vectors, which may furthermore appear similar to one another, the underlying cause for the aberrance differs greatly. If understanding and identifying aberrant response patterns that result from various kinds of examinee behavior is of primary interest to the researcher, data simulation provides the best method for ensuring that aberrance is not attributed to an incorrect underlying cause.

If the underlying cause of aberrant response patterns is of less importance to the researcher than simply making an accurate assessment of issues such as hit- and error rates for a person-fit statistic, then data simulation remains the method of choice when designing a research study. A researcher who simulates all data for an investigative person-fit study not only maintains control over the general characteristics of simulated items and participants, but the researcher can also control factors such as the number, type, and degree of aberrant response patterns within the simulated data sets. Because the researcher has simulated all data in the study and has knowledge of which patterns were simulated to be aberrant, it is easy to draw conclusions about issues such as power, and Type I and Type II error rates when investigating person-fit using simulated data.

Many person-fit studies that utilize simulated data have used item response theory for data simulation (e.g., Armstrong & Shi, 2009; Karabatsos, 2003; Woods, 2008). The use of item response theory in data simulation is not surprising, given the prevalence of IRT methodologies in person-fit research. However, regardless of whether or not IRT methodologies are used in the actual person-fit data analysis, IRT models provide a useful and intuitive framework for simulating item response data.

The steps for simulating data using IRT are quite simple and can be summarized rather succinctly as a series of steps. In the first step, a particular IRT model is chosen for data simulation. In the second step, the researcher chooses population-level values for both person- and item parameters. Once the person- and item population parameters have been chosen, they are treated as known, and expected success probabilities (conditional on each simulated examinee's true θ value) are generated for all items. In the case of simulating dichotomous item response, a random draw is pulled from a uniform distribution [0, 1] and compared to the expected probability for each cell in the data set. If the random draw is less than or equal to the expected probability of success given the population-level person- and item parameters, the item response is coded 1; if the random draw is greater than the expected probability of success, the item response is coded 0.

This procedure for simulating dichotomous item responses can easily be generalized to be appropriate for polytomous data as well. Using an IRT model for polytomous data (e.g., the graded response model), person- and item parameters are simulated and treated as known. For each item response, a random draw is pulled from the uniform distribution. The random draw is compared to boundaries for the possible score categories, and the value of the random draw determines which score will be simulated on the item.

The IRT model is especially useful for simulating item responses because it provides an easy and intuitive method for generating data that conform to specifications at both the person- and item level. Over repeated sampling using the same person- and item parameters, the proportion of correct answers to each item will approach the expected proportion, as defined by the person- and item parameters, but including random draws from the uniform distribution ensures that data simulation remains a random process. The final step in the simulation process

involves identifying a subset of simulated examinees and recoding their responses to make them aberrant.

When simulating cheating behavior, researchers often begin by simulating model-fitting data from an IRT model and then identify a subset of simulated examinees to be identified as cheaters and a subset of items to be identified as exposed. Most often, the researcher selects a subset of very difficult items to be simulated as exposed (e.g., Hendrawan, Glas, & Meijer, 2005; Karabatsos, 2003; Meijer, 2003). The researcher alters item responses where cheaters encounter exposed items. Karabatsos (2003) simply imputed correct answers wherever a cheater encountered an exposed item, while Meijer (2003) imputed correct responses with a fixed success probability of 0.90 for cheaters on exposed items while Hendrawan, Glas, and Meijer (2005) used a fixed success probability of 0.80.

Studies have varied in manipulating the proportion of exposed items and/or cheaters in data sets. For example, Hendrawan, Glas, and Meijer (2005) varied the proportion of exposed items in their conditions (1/6, 1/3, or 1/2 of the exam) but all of their conditions had a fixed percentage of cheaters set at 10% throughout. Conversely, Karabatsos (2003) varied the percentage of cheaters across conditions (5%, 10%, 25%, or 50% of examinees) but the percentage of exposed items was fixed at 18%. In all of these cases, the cheaters and exposed items are fixed across replications within each condition.

Empirical studies.

In developing and investigating person-fit statistics, some researchers have elected to base their studies—in whole or in part—on empirical data obtained either from study participants or database records (e.g., Armstrong & Shi, 2009; Brown & Villarreal, 2007; Comrey, 1985; Emons, Sijtsma, & Meijer, 2005, Reise & Widaman, 1999). With established person-fit statistics,

such studies can be helpful in assisting researchers to better understand how these various person-fit assessment techniques behave with real—often messy—data obtained from actual people. Empirical studies are beneficial to furthering research on person-fit because they provide an opportunity to assess characteristics of person-fit statistics when applied to more realistic data sets. However, in most situations with empirical data, the identity of cheaters and exposed items is not known by the researcher, so conclusions regarding power and error rates can be difficult to reach.

Method

The overarching goal of this study was to investigate the utility of a new application of existing person-fit statistics for detecting cheating. Specifically, this study was undertaken to explore the applicability of Ferrando's factor-analytic *lco* and *M-lco* (2007, 2009) statistics in a novel model comparison approach for cheating detection. With rare exceptions, such as Woods (2008), cheating is not often discussed or conceptualized in the literature as a multidimensional phenomenon. In a unidimensional test design where raw exam scores are assumed to be directly influenced by the examinee's underlying θ level, once the common underlying trait has been accounted for, item responses should be fully independent for all examinees. However, if some examinees take the exam with prior knowledge of a subset of exposed items, then a unidimensional factor structure would not be adequate for these examinees on these exposed items. Rather, I posit that the factor structure for these individuals on the exposed items becomes multidimensional, where the second factor that emerges on these exposed items will account for covariance among these exposed items due to prior exposure.

To date, no researchers have studied cheating in the multidimensional context that is explored in this paper. As previously mentioned, Woods (2008) discussed cheating as a

multidimensional phenomenon, but that study used multilevel logistic regression as opposed to the factor analytic technique that I use in the present study. Drasgow, Levine, and McLaughlin (1991) discussed person-fit in the context of a multidimensional test, but cheating was not one of the underlying dimensions in that study. Rather, their study was focused on tests that were designed to be multidimensional from a content perspective—or a multidimensional test consisting of independent, unidimensional subtests. Similarly, Ferrando's (2009) *M-lco* statistic has so far been researched using only data that are expected to be multidimensional: personality inventories comprised of items designed to assess multiple, distinct constructs, for example. I investigated a new application of Ferrando's person-fit statistics for data that come from an exam that was designed and intended to be unidimensional.

If cheating violates the unidimensionality of exposed items for cheaters, then this violation of unidimensionality should be systematic and detectable at the examinee level by comparing person-fit computed by a one-factor model with person-fit computed by a two-factor model, with expectation being that the two-factor model should fit significantly better at the person level for cheaters. I investigated a new method for assessing person-fit, where one- and two-factor models are fit to item responses.

The *lco* and *M-lco* person fit statistics follow a χ^2 distribution with $J - K$ degrees of freedom (Ferrando, 2007, 2009). For each examinee, *lco* from the one-factor model and *M-lco* from the two-factor model were computed and person-fit was assessed using χ^2 difference tests. A significant χ^2 difference test was taken as indication that an individual's response pattern is aberrant and was not adequately represented by the unidimensional model. Two simulation studies were conducted to investigate this new technique.

Simulation Study 1

The first simulation study was conducted for the purposes of assessing Type I error rates for the statistics used in this study. Polytomous item responses with score categories ranging between 0 – 4 were generated from the graded response model using the program WinGen (Han, 2007). Population parameters were drawn from the following distributions: $\theta \sim N(0, 1)$, $a \sim U(0.5, 2.0)$, and $b \sim N(0, 1)$. A total of 2,000 replicated data sets were generated in this simulation study, with each generated data set consisting of 25 items and 1,000 simulated examinees.

The test length and number of score categories used in the present study were chosen based on Ferrando's 2009 article, which was the first study to use the *M-lco* statistic. Ferrando (2009) used test lengths with 10, 18, and 24 items, and all items had score categories ranging from 0 – 4. One item was added to Ferrando's longest test length so the test length would be an odd number, which would eventually facilitate some aspects of item exposure simulation to be discussed in the next section. No systematic misfit was introduced at any point in the data generation process during this simulation study.

Following data generation, the 2,000 replicated data sets were analyzed using R, version 2.10.1. (R Development Core Team, 2009). For each replicated data set, a one-factor exploratory factor analysis model was fit to the data using R's built-in ML exploratory factor analysis procedure: factanal. As recommended by Ferrando (2007, 2009), factor scores were estimated using Bartlett's method. The *lco* and *lcz* person-fit statistics were computed for each simulated examinee using the factor scores and parameter estimates obtained from the one-factor model. Next, a two-factor exploratory factor analysis model was fit to the data. Oblique target rotation was used in the two-factor model, with the target matrix consisting entirely of 1s in the first column and 0s in the second column. Once again, factor scores in the two-factor model were

estimated using Bartlett's method and *M-lco* was computed for each simulated examinee. Changes in person-fit between the one-factor- and two-factor models were assessed by computing the *lco* difference: *lco* – *M-lco*. Because the *lco* difference should be distributed χ^2 with $df = 1$, any response pattern resulting in an *lco* difference greater than 3.841 was flagged as aberrant. The *lcz* statistic that was calculated for the one-factor model is expected to follow a standard normal distribution, and poor person-fit is indicated by large, positive values of *lcz*, so any simulated examinee with a response pattern resulting in an *lcz* value greater than 1.645 was flagged. Factor loadings from the one-factor model and rotated factor loadings from the two-factor model were also aggregated across replications.

This process was repeated for all 2,000 replicated data sets in the first simulation study. The number of simulated examinees flagged as aberrant based on the criteria described in the previous paragraph were used to calculate Type I error rates for the *lcz* statistic and the *lco* difference. For each person-fit method, approximately 5% of all simulated examinees were expected to be flagged as aberrant based on these criteria. Because data were simulated to be unidimensional and no items were simulated to be exposed in this study, no consistently strong loadings on the second factor were expected to be found.

Simulation Study 2

The second simulation study took the methodology employed in the first simulation study and extended it by adding several manipulations to the procedure. As in the first study, polytomous item responses ranging between 0 – 4 were simulated from the graded response model using WinGen (Han, 2007). The same person- and item parameters were employed in the second simulation study, and each simulated data set once again consisted of 1,000 simulated examinees and 25 items. However, in the second study, additional manipulations were performed

on the item responses prior to fitting factor analysis models to the data. The second simulation study took the form of a 3 × 4 design, where the manipulated conditions were number of simulated cheaters (10, 50, 100, or 250 out of 1,000) and number of exposed items (3, 7, or 13 out of 25).

Cheaters were selected as follows: all 1,000 simulated examinees were sorted in descending order to true ability level such that the simulated examinee with the highest true ability level was located in row 1 and the simulated examinee with the lowest true ability level was located in row 1,000. In each condition, cheaters were selected from three ranges within the bottom half of the ability distribution: approximately one-third of cheaters were selected from the range of ability just below the distribution's mid-point, approximately one-third were selected from the bottom of the ability distribution, and the remaining approximate one-third were selected from the mid-point between those two ranges. For example, in the conditions in which 10 cheaters were present, the following simulated examinees (designated by row number) were selected to serve as cheaters: 501-503, 749-752, and 998-1,000.

Exposed items were selected from three ranges of difficulty. Items were sorted in descending order of difficulty, with difficulty judged by the true *b* parameter associated with the highest score category of $X_j = 4$. Approximately 1/3 of the exposed items were the easiest items on the exam, approximately 1/3 of the exposed items were near the mid-point of the difficulty range, and approximately 1/3 of the exposed items were the most difficult items on the exam. For example, in the condition in which 7 items were exposed, the following items (designated by column number) were selected to serve as exposed items: 1-2, 12-14, and 24-25.

Once all cheaters and exposed items had been selected, item responses were manipulated. In each condition of the study, cheaters and exposed items were identified, and cheating was

simulated in cells in the data matrix in which cheaters encountered exposed items. In these cells, cheaters were assigned a 0.80 probability of achieving the maximum possible score of 4 on exposed items, regardless of the examinee's true ability level or the item's true difficulty level. All subsequent analyses in the second simulation study were identical to the methodology described for the first simulation study.

Results

Simulation Study 1

All factor analysis models in the first simulation study converged successfully. On average, both methods had Type I error rates that slightly deviated from the nominal 0.05 level. The *lcz* person-fit statistic's Type I error rate was somewhat conservative. The overall Type I error rate for *lcz* across all replications was equal to 0.040. Conversely, the *lco* difference method had a Type I error rate that was somewhat inflated, with an overall error rate of 0.071.

A cursory review of Type I error rates across the range of simulated examinee ability levels indicated that error rates may vary as a function of examinee ability level, so this was investigated further. Simulated examinees were sorted in ascending order of true ability and then grouped into 10 ability strata, each of which contained 100 simulated examinees. Type I error rates were aggregated across the 100 simulated examinees grouped within each stratum. Type I error rates for each ability stratum are reported in Table 3, and Figure 12 graphically illustrates the relationship between Type I error rate and examinee ability level. Both methods have conservative Type I error rates for examinees with very low ability levels (i.e., stratum 1) and very high ability levels (i.e., stratum 10), and both methods have inflated Type I error rates for examinees who are have near-average ability levels (i.e., strata 5 and 6). However, on average the *lcz* person-fit statistic's Type I error rate was lower than expected, with Type I rates for 6 out

of 10 examinee ability strata less than the 0.05 level, and the *lco* difference method's Type I error rate was higher than expected, with Type I rates for 8 out of 10 strata greater than the 0.05 level.

A summary of factor loading is provided in Table 4. Loadings from the one-factor model appear under the FA 1 heading, and loadings from the two-factor model appear under the FA 2 heading. Loadings onto the second factor were somewhat higher than anticipated, given that data in this study were simulated to be unidimensional and model-fitting. However, no overwhelmingly strong evidence of multidimensionality was found, with loadings onto the second factor ranging between 0.135 and 0.194. The factors in the two-factor model had a correlation of -0.644. The higher-than-expected loadings onto the second factor and the factor correlation will be addressed in greater detail in the Discussion section.

Simulation Study 2

All factor analysis models in the second simulation study converged successfully. A summary of the proportions of correctly-identified cheaters and the proportions of incorrectly-identified non-cheaters based on the flagging criteria outlined in the Method section is provided in Table 5. As reported in this table, the *lcz* person-fit statistic was most successful when 13 out of 25 items (or 52% of the exam) were exposed and 10 out of 1,000 examinees (or 1%) were cheaters, achieving a detection rate of 0.668 in this condition. Detection rates for *lcz* in other conditions in this study generally ranged from moderate to poor. The *lco* difference method was most successful in detecting cheating when 13 out of 25 items were exposed and 50 out of 1,000 examinees (or 5%) were cheaters, achieving a detection rate of 0.878 in this condition. Detection rates for the *lco* difference method were generally better than what was observed for the *lcz* statistic, but there were some conditions—most notably those in which (1) very few items were

exposed, and (2) there were either very few or very many cheaters present—in which detection rates for the *lco* difference method were also quite poor.

As illustrated in Figure 13 for *lcz* and Figure 14 for the *lco* difference method, detection rates generally improved with increased proportions of exposed items for both techniques. However, the impact of the number of cheaters differed for each method. As illustrated in Figure 15, the *lcz* person-fit statistic was most powerful in the condition with the fewest number of cheaters, where only 10 out of 1,000 (or 1%) of examinees were cheaters. Detection rates for *lcz* declined as more cheaters were added to the data set. The relationship between detection rate and number of cheaters was somewhat more complex for the *lco* difference method. As illustrated in Figure 16, detection of simulated cheaters with this method was most successful when moderate amounts of cheaters (i.e., 5% or 10% of examinees) were present in the data set, and detection rates were poorer when very few or very many cheaters were present.

Detection rates for the two methods will be compared in two different contexts: number of exposed items and number of cheaters. The line graphs shown in Figure 17 through Figure 20 compare detection rate performance across levels of item exposure, with each individual graph corresponding to a particular cheating condition. As illustrated in Figure 17, when only 1% of examinees are cheaters, the *lcz* method performed as well or better than the *lco* difference method at all investigated levels of item exposure. However, Figure 18 through Figure 20 show that the *lco* difference method consistently detected more cheaters than the *lcz* statistics across all levels of item exposure when 5% or more of examinees are cheaters. Next, the line graphs shown in Figure 21 through Figure 23 compare detection rate performance across cheating conditions, with each graph corresponding to a particular level of item exposure. These figures reinforce that when very few cheaters (i.e., 1% of examinees) are present in a data set, the *lcz* method performs

as well or better than the *lco* difference method. However, the *lco* difference method outperforms *lcz* when more cheaters are present, regardless of the number of exposed items.

Factor loadings for the second simulation study are shown in Table 6 through Table 8. In conditions where only 1% of examinees were cheaters, the exposed items' loadings onto the second factor were not particularly noteworthy when compared to the non-exposed items' loadings on this factor. However, in the conditions in which at least 5% of examinees were cheaters, exposed items' loadings onto the second factor began to differ from the non-exposed items' loadings, with exposed items loading more strongly than the non-exposed items onto the second factor in the two-factor model. As would be expected, in the conditions in which 3 or 7 items are exposed, the exposed items' loadings onto the second factor generally increase when more cheaters are present in the data set. In the final four conditions of this study, in which 52% of the exam is exposed, loadings onto the second factor change dramatically when compared to the other eight conditions of the study. When more than half of the exam has become exposed and 5% or more examinees are cheaters, the role of the second factor changes, with the exposed items now loading weakly onto the second factor and the non-exposed items loading more strongly onto this factor. Across all conditions, the factor correlations from the two-factor models ranged between -0.456 and -0.745, with 9 out of the 12 conditions in this study resulting in a factor correlation for the two-factor model between -0.6 and -0.7.

Discussion

The results of these two studies highlight some of the promise and several important limitations of the *lco* difference method for assessing person-fit. One of the most notable apparent strengths of the *lco* difference method, when compared to a traditional person-fit statistic like *lcz*, is that the *lco* difference method remains powerful when more than just a few

cheaters are present in a data set. In fact, the *lco* difference method appears to reach its highest levels of power for detecting cheating when a moderate number of cheaters are present. This outcome is not altogether surprising, given that success in identifying cheaters using the *lco* difference method is dependent upon the factor analysis model's ability to extract a second factor when cheating has occurred. It seems reasonable to conclude that a second (cheating) factor is more likely to emerge when at least a moderate proportion of examinees are cheaters. Conversely, the *lcz* statistic's power appears only to diminish as more cheaters are added to the data set. However, in conditions in which very few cheaters were present, the simulation study results indicate that the *lco* difference method is no more effective than the *lcz* statistic. In conditions where only 1% of examinees were cheaters, the *lcz* person fit statistic performed as well or slightly better than the *lco* difference method. The difference in how the proportion of cheaters relates to the effectiveness of each method is reflective of the different approaches the two methods take to evaluating person-fit.

As previously discussed, all person-fit statistics essentially amount to comparing a vector of observed item responses to their expected values, given some model, and estimating the extent to which observed performance differs from expected performance. However, when many cheaters are present in a data set, their better-than-expected performance on the exposed items affects the exposed items' parameter estimates. This makes the exposed items appear to be easier than they really are, which in turn causes the cheaters' correct answers to the exposed items to appear less aberrant, thus reducing the gap between observed and expected performance for these cheaters on the exposed items and reducing a person-fit statistic's power to detect cheating.

For this reason, a person-fit statistic such as *lcz* is most effective when very few of the examinees are cheaters. As more cheaters are added to the data set, their influence on the item

statistics has the net effect of making the cheaters more difficult to detect. The results of the second simulation study support my assertion that a traditional person-fit statistic like *lcz* loses power as more cheaters are added to a data set. Performance for the *lcz* person-fit statistic declined as more cheaters were added to the data set. Because the *lco* difference method also involves comparing observed performance to expected performance, it also loses power when a very large proportion of examinees are cheaters, but this method remains relatively effective when small-to-moderate proportions of cheaters are present while the *lcz* method's effectiveness for detecting cheating drops off considerably when more than a few examinees are cheaters.

One of the factors that drew me to investigate the *lco* difference method for investigating person-fit was the potential for using loadings as a diagnostic indicator of item exposure. The person-fit research literature is largely focused on identifying individuals who have aberrant response patterns, but no effort is made to identify the exposed items that contribute to the response pattern's aberrance. This may be due—wholly or in part—to the aforementioned common practice in much of the person-fit simulation research to limit simulation of exposed items to the most difficult items on the exam. However, in real-world testing situations, exposure may not be limited to the most difficult items on an exam, so a tool for identifying exposed items would be very useful for a test administrator conducting an investigation. A factor-analytic technique such as the *lco* difference method holds promise because it may be useful for identifying both cheaters and exposed items.

The factor loading results were somewhat mixed. Loadings may be useful tools for identifying exposed items, but this occurs only when two conditions have been met: first, at least a moderate proportion of examinees (i.e., 5% or more) are cheaters; and second, less than half of the items are exposed. If a small proportion of examinees are cheaters, or if more than half of the

items on the exam have been exposed, factor loadings may be less useful or possibly even misleading for this use. When very few cheaters are present, loadings for the exposed items are indistinguishable from loadings for the non-exposed items on the second factor. In situations in which more than half of the exam's items are exposed, an investigation based solely on factor loadings may lead to the incorrect conclusion that the non-exposed items have been exposed, because their factor loadings onto the second factor become much stronger than the loadings for the exposed items.

To investigate the use of rotated factor loadings further, I performed a small-scale follow-up investigation and fit the two-factor model to data sets across several conditions using oblique quartimin rotation. In terms of utility for identifying exposed items, quartimin-rotated factor loadings fared no better or worse than the oblique target rotation that was reported in this paper. Two apparent differences between the two rotation methods were observed, however.

The first major difference between rotation methods was the effect of the rotation method on the pattern of the loadings. When target rotation was applied to factor loadings in conditions where at least a moderate proportion of examinees were cheaters and less than half of the items were exposed, exposed items generally manifested themselves by having strong loadings onto both the first and second factor in the two-factor model, while non-exposed items only loaded strongly onto the first factor. Given the target matrix that was used for the two-factor models in this study, this outcome would be expected. In the two-factor model, the target matrix specified strong loadings for all items onto the first factor. When quartimin rotation was used in these conditions, exposed items loaded strongly onto the second factor only, while non-exposed items loaded strongly onto the first factor only. In the simulation study conditions where at least moderate proportions of examinees were cheaters and less than half of the items were exposed,

exposed items were easy to identify regardless of which rotation was used, but they manifested themselves differently depending on the rotation method. In the study conditions with very few cheaters, all items loaded strongly onto the first factor and had mostly weak loadings onto the second factor regardless of rotation method, but loadings onto the second factor tended to be closer to zero when quartimin rotation was used. However, neither rotation method appeared to be noticeably better-suited for detecting exposed items in this condition. In study conditions with more than half of the items exposed, target rotation resulted in strong loadings onto both factors for non-exposed items and strong loadings onto the first factor only for exposed items, while quartimin rotation resulted in strong loadings onto the second factor only for non-exposed items and strong loadings onto the first factor only for exposed items. When more than half of the items on an exam are exposed, an investigator using either rotation method likely would arrive at the wrong conclusions when flagging exposed items based on factor loadings.

The other major difference between the two oblique rotation methods is the effect of the rotation method on the estimated factor correlation. The strong factor correlations that were found in the first simulation study were unexpected. Data in the first study were simulated to be unidimensional, so the second factor in the two-factor model was not expected to have a strong correlation with the first factor. A short follow-up investigation revealed that the stronger-than-expected factor correlations that were observed in the first study likely were byproducts of the target rotation method that was used. Factor correlations in the first simulation study tended to be much closer to zero when quartimin rotation was used. When quartimin rotation was applied to factor loading matrices from a sample of data sets across conditions in the second simulation study, factor correlations were generally weaker in conditions with 10 or 250 cheaters, with absolute values generally ranging between 0.00 – 0.30. In conditions with 50 or 100 cheaters,

factor correlations obtained from quartimin rotation were much stronger, with absolute values generally ranging between 0.50 – 0.80.

Limitations

The results of this study show that this new method for assessing person-fit holds promise and warrants further investigation. However, both the new person-fit technique proposed in this paper and the research methodology that was used to investigate it have limitations that bear mentioning. These limitations will be addressed in the following sections.

Limitations of the *lco* difference method.

In developing the *lco* and *M-lco* person-fit statistics, Ferrando (2007, 2009) asserts that these statistics follow a χ^2 distribution with degrees of freedom equal to $J - K$. If this is true, then their difference should be distributed χ^2 with $df = 1$ when comparing the difference of *lco* from a one-factor model with *M-lco* from a two-factor model. Furthermore, if *M-lco* and *lco* are indicators of model fit evaluated at the level of the individual, fit should never decrease when adding an additional factor to the model. Recall that with these statistics, large values are indicative of poor fit, so if a one-factor model and a two-factor model are both fit to the same data set, then subtracting the two-factor model's *M-lco* statistic from the one-factor model's *lco* statistic for a given examinee should always yield a positive value. However, this was not always the case in the simulation studies presented in this paper. The distribution of the *lco* difference from Study 1, in which no cheaters or other sources of systematic model misfit were included, is shown in Figure 24. As illustrated in this figure, a small amount of these *lco – M-lco* difference values are negative. Taken at face value, this outcome would seem to imply that the two-factor model fits worse for some individuals than the one-factor model, but this is not likely to be the case with these examinees.

Most of the negative *lco* difference values are large enough that they cannot be dismissed as mere cases of rounding error; some observed *lco* difference values approach -2. Prior to investigating further, the computational methods employed in this study were reviewed and verified. All statistics from this study were calculated exactly as described by Ferrando (2007, 2009), thus ruling out the possible explanation that these negative *lco* difference methods are due to computational errors. These occasionally-negative *lco* difference values, taken with the slightly inflated observed Type-I error rate for the *lco* difference method, indicate that the sampling distribution for the *lco* difference method may not follow a χ^2 distribution as well as might be expected. Similarly, the *lcz* statistic may deviate from the normal distribution somewhat. The distribution of *lcz*, shown in Figure 25, appears to have a slight negative skew (recall that negative values of *lcz* indicate good person-fit and positive values indicate poor person-fit), and the observed Type I error rate for *lcz* was slightly below the expected 0.05 level.

One possible explanation for these observations is that using estimated person- and item parameters may alter the sampling distributions for these statistics somewhat. As previously discussed in the Introduction, several researchers (e.g., Nering, 1995; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999) have found that the IRT-based l_z person-fit statistic's sampling distribution deviates from $N(0, 1)$ when $\hat{\theta}$ is substituted for θ in calculating the statistic. If *lco* and *lcz* are the factor-analytic counterparts to IRT's l_0 and l_z person-fit statistics, then it is possible that the sampling distributions for these factor-analytic person-fit statistics also may deviate from their expected sampling distributions when estimates are used in place of true parameters to calculate these statistics.

Limitations of the research methodology.

When designing this study, I purposefully limited the amount and type of misfit that was introduced into the data. Because the purpose of this study was to test a new technique, I chose to generate model-fitting data, and the only source of systematic misfit in the data was created by the simulated cheating behavior. Introducing population model misfit—such as other sources of multidimensionality beyond cheating, for example—into the methodology of this study likely would have resulted in data sets that more closely resembled data observed in real-world testing contexts. Data as "clean" as those used in this study are unlikely to be obtained when administering a real-world test to real-world examinees. Other systematic sources of misfit—such as multidimensionality and non-independent errors, for example—may show up in such data sets, but those other sources of misfit were not simulated in the present study.

Using data that have unrealistically good model fit (aside from misfit introduced by cheating) somewhat limits the generalizability of the results of this study. The reader is encouraged to consider the rather ideal qualities of the data used in this study and interpret the reported detection rates with an appropriate degree of caution. In a real-world application of this method, detection rates may be somewhat lower than the rates reported in this study due to the influence of potential sources of model misfit other than cheating. Because this study was the first to use this new method, I elected to utilize fairly ideal simulated data mainly for the purposes of establishing best-case scenario performance levels for the *lco* difference method. Had more realistic (i.e., messier) data sets been used in this study and performance of the *lco* difference method been found to be poor, it would have been difficult to determine if the poor performance was due to deficiencies in the *lco* difference method or due to the level of misfit in

the data. In this early stage of development for this technique, the present study has established that this method may have utility for detecting cheating in certain conditions.

Another limitation of this study that bears discussion is the method that was used to simulate cheating behavior. One of the methodological goals of this study was to investigate the performance of person-fit statistics under more realistic conditions of item exposure, namely, by allowing more than just the most difficult items on the exam to be exposed; however, I make no claims that the cheating simulation methodologies that were employed in the present study were without fault. Each simulated cheater had a 0.80 probability of success on each exposed item. A success probability that was less than 1 was chosen to incorporate an element of random error into the simulation. This ensured that all cheaters would have a high success probability on each exposed item, but these simulated cheaters occasionally would answer some exposed items incorrectly as well. Although a 0.80 probability of success for cheaters on exposed items sometimes has been used in past person-fit research (e.g., Hendrawan, Glas, & Meijer, 2005), this value is admittedly arbitrary. Reasonable arguments could be made for employing a higher or lower success probability. Also, in certain rare instances, assigning a 0.80 success probability has the potential to disadvantage a cheater with a moderate ability level on an easy exposed item, because the cheater's model-implied probability of success may be greater than 0.80. Furthermore, by applying the same success probability to all cheaters across all exposed items, I make the somewhat unreasonable implicit assumption that it is equally easy to cheat on all of the exposed items, and all cheaters are equally capable of successfully cheating. Although I would suggest that including exposed items from a broader range of difficulty was an important first step toward more a more realistic simulation of item exposure in studies like this, additional

steps could have been taken to strengthen the fidelity of the item exposure simulation even further.

Future Research

The results of this study indicate that further investigation of the *lco* difference method is warranted. The present study has shown the *lco* difference method performs similarly well compared to the *lcz* method when few cheaters are present, and the *lco* difference method appears to be more powerful than *lcz* when larger proportions of cheaters are present. However, many issues need to be explored further in future research on this method, and several of the most salient issues will be discussed here.

Differences in the Type-I error rates should be considered when comparing the cheating detection rates for the *lco* difference method and the *lcz* person-fit statistic. The results of the present study indicate that further research on the sampling distributions for the *lco* difference method and the *lcz* person-fit statistic are necessary. Traditional cut-off values based on the χ^2 or normal distributions may not always be appropriate for these person-fit techniques. Additional research on the sampling distributions of these statistics is necessary.

Factor loadings may hold some value as diagnostic tools for identifying exposed items, but clearly this particular area needs further investigation and development. In certain circumstances, factor loadings may be useful indicators of exposed items. However, the present study demonstrated that in some situations, loadings are less useful or even misleading. Further investigation into the behavior of factor loadings in the presence of cheating would be an informative and valuable contribution to the research literature on this topic. Future research devoted to investigating other methods (e.g., OLS) and other rotation methods may provide valuable contributions to this research field.

Although using the *lco* difference method as a means to identify cheaters amongst the examinees was the main focus of the present study, some additional investigating was performed on the factor loadings. A histogram displaying the distribution of loadings onto the second factor across replications for one exposed item is shown in Figure 26. More specifically, this distribution of factor loadings comes from item 12 from the condition with 7 exposed items and 50 cheaters in the dataset. As seen in this figure, the vast majority of the loadings onto the second factor across replications were large and positive, although some replications had weaker loadings or even large, negative loadings. The distribution of loadings for the same item, but in the condition with 7 exposed items and only 10 cheaters, is displayed in Figure 27. As shown in this figure, loadings were widely dispersed across replications when fewer cheaters were present, and on average, loadings tended to be smaller than what was observed in the condition with more cheaters.

Future research on aberrant response pattern detection may benefit from more realistic simulation of cheating behavior. As previously discussed, the method used in the present study essentially assumed that all cheaters were equally able to cheat, and that all exposed items were equally easy for cheaters to answer correctly. However, this may not be a very realistic representation of real-world cheating behavior. It is more likely that some individuals are more capable cheaters than others and some exposed are easier for cheaters to answer correctly. Future studies in person-fit would be well-served by incorporating distributions for cheaters and exposed items.

Another topic for future *lco* difference research is applying a model comparison method similar to the one employed in this study to dichotomous data. Applying the statistical methodologies employed in the present study to dichotomous data adds certain complications,

due to the complexities of calculating unstandardized factor loadings and uniquenesses for dichotomous items. Another area that holds potential for this line of research is comparing changes in person-fit across IRT models of varying complexity (e.g., change in person-fit going from a unidimensional IRT model to a multidimensional IRT model). Regardless of which measurement framework may be used in future research, extending model comparison person-fit methods presented in this study to dichotomously-scored items may provide a valuable contribution to person-fit research, given the prevalence of tests comprised entirely of dichotomously-scored items.

Conclusion

This dissertation investigated a new technique for conceptualizing and measuring person-fit. Findings indicate that under the right set of circumstances, cheating can be detected by measuring changes in person-fit through factor-analytic model comparison. Questions arise about the sampling distribution of the statistic that has been proposed, and further research on this matter is warranted.

References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69,* 44-49.

Armstrong, R. D & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33,* 391-410.

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the l_z person-fit statistic. *Practical Assessment, Research & Evaluation, 12(16).* Available online: http://pareonline.net/getvn.asp?v=12&n=16

Bollen, K. A. & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 235-262). New York: Basil Blackwell.

Brown, R. S. & Villarreal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing, 7,* 1-25.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

Comrey, A. L. (1985). A method for removing outliers to improve factor analytic results. *Multivariate Behavioral Research, 20,* 273-281.

de la Torre, J. & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and it reference distribution. *Journal of Educational Measurement, 45,* 159-177.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

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,* 67-86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education. Special Issue: Person-fit research: Theory and applications, 9,* 47-64.

Emons, W. H. M., Meijer, R. R, & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement, 26,* 88-108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39,* 1-35.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10,* 101-119.

Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement, 28,* 126-140.

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42,* 481-507.

Ferrando, P. J. (2009). Multidimensional factor-analysis-based procedures for assessing scalability in personality measurement. *Structural Equation Modeling, 16,* 109-133.

Ferrando, P. J. & Lorenzo, U. (2000). WPerfit: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement, 60,* 479-487.

Ferrando, P. J. & Lorenzo-Seva, U. (2007). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research, 42,* 675-706.

Glas, C. A. W. & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27,* 217-233.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Han, K. T. (2007). *WinGen* [Computer software]. Amherst, MA: University of Massachusetts at Amherst.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20,* 191-206.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effects of person misfit on classification decisions. *Applied Psychological Measurement, 29,* 26-44.

Johanson, G. & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement, 62,* 435-443.

Kamata, A. & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136-153.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277-298.

Kim, J. & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26,* 38-51.

Kogut, J. (1986). *Review of IRT-based indices for detecting and diagnosing aberrant response patterns* (Research Report No. 86-4). Enschede, The Netherlands: University of Twente.

Kogut, J. (1988). *Asymptotic distribution of a person-fit statistic* (Research Report No. 88-13). Enschede, The Netherlands: University of Twente.

LaHuis, D. M. & Copeland, D. (2007). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods, 12,* 296-319.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Meijer, R. R. (1996). Person-Fit research: An introduction. *Applied Measurement in Education, 9,* 3-8.

Meijer, R. R. (2003). Diagnosing item score patters on a test using item response theory-based person-fit statistics. *Psychological Methods, 8,* 72-87.

Meijer, R. R., Muijtjens, A. M. M., & van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9,* 77-89.

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261-272.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. Psychometrika, 55, 75-106.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. Applied Psychological Measurement, 19, 121-129.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person fit statistic. *Applied Psychological Measurement, 22,* 53–69.

R Development Core Team (2009). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543-568.

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,* 3-21.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Sato, T. (1975). *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23,* 41-53.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7,* 131-145.

Sijtsma, K. & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49,* 79-105.

Sijtsma, K. & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16,* 149-157.

Sijtsma, K. & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,* 191-208.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66,* 331-342.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2009). A Monte Carlo study of the effect of item characteristic curve estimation on the accuracy of three person-fit statistics. *Applied Psychological Measurement, 33,* 307-324.

Tatsuoka, K. K. & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221-230.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. Applied Psychological Measurement, 23, 327-345.

Wirth, R. J. & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58-79.

Woods, C. M. (2008). Monte Carlo evaluation of a two-level logistic regression for assessing person fit. *Multivariate Behavioral Research, 43,* 50-76.

Table 1

Aberrant Item Score Patterns on a Fictitious 12-Item Test (Meijer, 1996)

	Item											
Behavior	1	2	3	4	5	6	7	8	9	10	11	12
Sleeping	0	0	0	1	1	1	1	1	1	1	0	1
Guessing	1	1	1	1	0	0	1	0	0	0	0	1
Cheating	1	1	0	1	0	1	0	0	0	1	1	1
Alignment Errors	1	1	1	1	1	0	1	1	0	0	0	0
Plodding	1	1	1	1	1	1	0	0	0	0	0	0
Extreme Creativity	0	0	0	0	1	1	1	1	0	1	1	1
Deficiency of Subabilities	0	0	1	0	1	1	1	0	1	1	1	0
π	.90	.85	.83	.82	.57	.55	.50	.49	.30	.25	.21	.15

Table 2

Item Parameters for a Hypothetical Five-Item Test (Hambleton et. al 1991)

	Item				
Parameter	1	2	3	4	5
a	0.67	1.00	1.14	1.34	1.27
b	-2.00	-0.59	0.15	0.59	1.19
c	0.01	0.20	0.15	0.15	0.10

Table 3

Type I Error Rates Across Ability Strata for Simulation Study 1

Stratum	*lcz*	*lco* Difference
1	0.004	0.031
2	0.022	0.060
3	0.043	0.078
4	0.061	0.087
5	0.071	0.093
6	0.070	0.095
7	0.061	0.090
8	0.043	0.080
9	0.022	0.064
10	0.005	0.035
Overall	0.040	0.071

Table 4
Factor Loadings for Simulation Study 1

	FA 1	FA 2	
Item	F1	F1	F2
1	0.495	0.628	0.160
2	0.438	0.572	0.161
3	0.658	0.822	0.182
4	0.553	0.702	0.178
5	0.593	0.748	0.178
6	0.269	0.380	0.135
7	0.510	0.654	0.172
8	0.480	0.618	0.171
9	0.292	0.405	0.139
10	0.310	0.427	0.146
11	0.600	0.761	0.187
12	0.567	0.727	0.187
13	0.559	0.713	0.185
14	0.519	0.663	0.172
15	0.481	0.624	0.179
16	0.325	0.446	0.152
17	0.300	0.418	0.145
18	0.501	0.645	0.173
19	0.674	0.847	0.193
20	0.612	0.768	0.178
21	0.665	0.838	0.194
22	0.546	0.696	0.178
23	0.610	0.774	0.189
24	0.466	0.600	0.164
25	0.541	0.693	0.184

Table 5

Detection Rates and Type I Error Rates for Simulation Study 2

	Exposed		*lcz* Rates		*lco* Difference Rates	
Condition	Items	Cheaters	Detection	Type I	Detection	Type I
1	3	10	0.184	0.038	0.162	0.070
2		50	0.105	0.033	0.600	0.044
3		100	0.064	0.031	0.569	0.033
4		250	0.033	0.034	0.272	0.062
5	7	10	0.615	0.037	0.497	0.067
6		50	0.447	0.028	0.854	0.036
7		100	0.201	0.023	0.776	0.031
8		250	0.027	0.032	0.453	0.093
9	13	10	0.668	0.036	0.654	0.066
10		50	0.557	0.024	0.878	0.035
11		100	0.454	0.017	0.781	0.032
12		250	0.201	0.014	0.470	0.086

Table 6

Factor Loading Comparison for Conditions With 3 Exposed Items

	10 Cheaters			50 Cheaters			100 Cheaters			250 Cheaters		
	FA 1	FA 2		FA 1	FA 2		FA 1	FA 2		FA 1	FA 2	
Item	F1	F1	F2	F1	F1	F2	F1	F1	F2	F1	F1	F2
1	***0.495***	***0.628***	***0.160***	***0.452***	***0.666***	***0.201***	***0.400***	***0.703***	***0.485***	***0.251***	***0.636***	***0.647***
2	0.438	0.572	0.161	0.435	0.517	0.085	0.435	0.491	0.065	0.436	0.504	0.093
3	0.658	0.822	0.182	0.659	0.802	0.134	0.658	0.787	0.175	0.658	0.794	0.198
4	0.553	0.702	0.178	0.554	0.680	0.125	0.554	0.669	0.160	0.554	0.680	0.188
5	0.593	0.748	0.178	0.593	0.713	0.116	0.592	0.688	0.123	0.593	0.696	0.146
6	0.269	0.380	0.135	0.270	0.338	0.077	0.270	0.319	0.066	0.270	0.326	0.082
7	0.510	0.654	0.172	0.509	0.628	0.118	0.509	0.618	0.152	0.510	0.631	0.181
8	0.480	0.618	0.171	0.480	0.586	0.105	0.481	0.572	0.124	0.481	0.584	0.153
9	0.292	0.405	0.139	0.292	0.359	0.074	0.292	0.339	0.061	0.292	0.348	0.080
10	0.310	0.427	0.146	0.309	0.388	0.085	0.309	0.368	0.082	0.307	0.375	0.101
11	0.600	0.761	0.187	0.600	0.733	0.128	0.600	0.721	0.166	0.601	0.736	0.200
12	0.567	0.727	0.187	0.568	0.712	0.141	0.568	0.713	0.209	0.568	0.735	0.258
13	***0.559***	***0.713***	***0.185***	***0.488***	***0.779***	***0.257***	***0.406***	***0.822***	***0.656***	***0.190***	***0.697***	***0.859***
14	0.519	0.663	0.172	0.519	0.639	0.119	0.520	0.631	0.154	0.518	0.640	0.183
15	0.481	0.624	0.179	0.481	0.603	0.123	0.480	0.596	0.166	0.481	0.613	0.203
16	0.325	0.446	0.152	0.325	0.408	0.087	0.324	0.390	0.091	0.325	0.403	0.118
17	0.300	0.418	0.145	0.298	0.375	0.083	0.298	0.356	0.078	0.298	0.365	0.099
18	0.501	0.645	0.173	0.502	0.619	0.118	0.502	0.609	0.150	0.502	0.625	0.186
19	0.674	0.847	0.193	0.674	0.824	0.142	0.674	0.811	0.186	0.674	0.825	0.226
20	0.612	0.768	0.178	0.611	0.740	0.125	0.610	0.720	0.146	0.612	0.737	0.182
21	0.665	0.838	0.194	0.665	0.813	0.139	0.666	0.802	0.187	0.665	0.818	0.229
22	0.546	0.696	0.178	0.544	0.653	0.108	0.544	0.630	0.111	0.545	0.648	0.148
23	0.610	0.774	0.189	0.609	0.735	0.124	0.609	0.713	0.136	0.609	0.733	0.180
24	0.466	0.600	0.164	0.466	0.567	0.102	0.466	0.553	0.116	0.466	0.565	0.146
25	***0.541***	***0.693***	***0.184***	***0.453***	***0.778***	***0.267***	***0.355***	***0.816***	***0.719***	***0.108***	***0.646***	***0.919***

Note. Exposed items are emphasized.

Table 7

Factor Loading Comparison for Conditions With 7 Exposed Items

	10 Cheaters			50 Cheaters			100 Cheaters			250 Cheaters		
	FA 1	FA 2		FA 1	FA 2		FA 1	FA 2		FA 1	FA 2	
Item	F1	F1	F2	F1	F1	F2	F1	F1	F2	F1	F1	F2
1	***0.497***	***0.638***	***0.173***	***0.469***	***0.678***	***0.373***	***0.428***	***0.648***	***0.450***	***0.273***	***0.586***	***0.573***
2	***0.431***	***0.559***	***0.153***	***0.403***	***0.584***	***0.325***	***0.363***	***0.555***	***0.397***	***0.221***	***0.498***	***0.514***
3	0.657	0.815	0.158	0.654	0.739	0.115	0.652	0.718	0.092	0.656	0.724	0.099
4	0.552	0.695	0.156	0.549	0.627	0.110	0.548	0.609	0.092	0.552	0.620	0.104
5	0.593	0.741	0.153	0.586	0.647	0.072	0.583	0.624	0.042	0.590	0.635	0.056
6	0.269	0.372	0.120	0.266	0.298	0.042	0.265	0.288	0.029	0.268	0.296	0.041
7	0.508	0.654	0.166	0.505	0.580	0.106	0.505	0.563	0.089	0.508	0.574	0.102
8	0.479	0.613	0.152	0.476	0.538	0.084	0.474	0.521	0.066	0.478	0.531	0.080
9	0.290	0.396	0.117	0.289	0.320	0.036	0.288	0.308	0.021	0.293	0.320	0.038
10	0.309	0.422	0.135	0.306	0.345	0.053	0.306	0.336	0.042	0.308	0.344	0.055
11	0.600	0.757	0.171	0.595	0.677	0.114	0.593	0.656	0.092	0.598	0.670	0.110
12	***0.554***	***0.730***	***0.226***	***0.500***	***0.832***	***0.599***	***0.427***	***0.773***	***0.710***	***0.184***	***0.645***	***0.854***
13	***0.565***	***0.731***	***0.207***	***0.515***	***0.807***	***0.525***	***0.446***	***0.753***	***0.628***	***0.220***	***0.639***	***0.773***
14	***0.510***	***0.667***	***0.204***	***0.470***	***0.746***	***0.499***	***0.411***	***0.701***	***0.598***	***0.201***	***0.600***	***0.738***
15	0.479	0.625	0.176	0.477	0.561	0.129	0.478	0.546	0.114	0.479	0.558	0.129
16	0.324	0.441	0.137	0.322	0.369	0.067	0.323	0.358	0.054	0.324	0.367	0.067
17	0.299	0.410	0.131	0.297	0.335	0.052	0.295	0.323	0.039	0.298	0.335	0.056
18	0.501	0.645	0.163	0.498	0.573	0.109	0.497	0.556	0.091	0.501	0.570	0.110
19	0.673	0.844	0.178	0.668	0.761	0.129	0.666	0.738	0.106	0.672	0.752	0.122
20	0.610	0.761	0.156	0.605	0.677	0.093	0.603	0.658	0.072	0.609	0.672	0.091
21	0.664	0.834	0.178	0.659	0.753	0.131	0.658	0.730	0.106	0.663	0.745	0.128
22	0.545	0.687	0.151	0.536	0.589	0.062	0.534	0.573	0.041	0.542	0.591	0.069
23	0.608	0.769	0.174	0.602	0.671	0.086	0.598	0.647	0.059	0.606	0.668	0.089
24	***0.457***	***0.602***	***0.188***	***0.415***	***0.678***	***0.481***	***0.355***	***0.633***	***0.579***	***0.144***	***0.527***	***0.714***
25	***0.548***	***0.717***	***0.220***	***0.485***	***0.817***	***0.603***	***0.401***	***0.748***	***0.715***	***0.141***	***0.594***	***0.845***

Note. Exposed items are emphasized.

Table 8

Factor Loading Comparison for Conditions With 13 Exposed Items

	10 Cheaters			50 Cheaters			100 Cheaters			250 Cheaters		
	FA 1	FA 2		FA 1	FA 2		FA 1	FA 2		FA 1	FA 2	
Item	F1	F1	F2	F1	F1	F2	F1	F1	F2	F1	F1	F2
1	***0.502***	***0.643***	***0.153***	***0.496***	***0.581***	***0.127***	***0.490***	***0.556***	***0.113***	***0.466***	***0.528***	***0.118***
2	***0.435***	***0.562***	***0.138***	***0.427***	***0.496***	***0.103***	***0.419***	***0.469***	***0.081***	***0.404***	***0.444***	***0.071***
3	***0.657***	***0.826***	***0.168***	***0.652***	***0.773***	***0.185***	***0.643***	***0.740***	***0.176***	***0.599***	***0.701***	***0.200***
4	***0.553***	***0.690***	***0.151***	***0.551***	***0.618***	***0.085***	***0.549***	***0.598***	***0.062***	***0.538***	***0.575***	***0.059***
5	0.585	0.779	0.176	0.545	0.829	0.505	0.490	0.762	0.598	0.286	0.640	0.730
6	0.265	0.381	0.129	0.250	0.374	0.224	0.228	0.347	0.267	0.144	0.297	0.326
7	0.503	0.670	0.173	0.477	0.700	0.397	0.439	0.656	0.478	0.279	0.572	0.608
8	0.474	0.638	0.166	0.446	0.663	0.388	0.409	0.619	0.465	0.254	0.531	0.579
9	0.288	0.411	0.130	0.270	0.409	0.254	0.243	0.374	0.295	0.151	0.319	0.357
10	0.305	0.429	0.137	0.287	0.427	0.253	0.264	0.397	0.300	0.169	0.343	0.370
11	***0.598***	***0.743***	***0.160***	***0.592***	***0.649***	***0.062***	***0.586***	***0.624***	***0.028***	***0.577***	***0.594***	***0.011***
12	***0.566***	***0.696***	***0.158***	***0.563***	***0.572***	***-0.039***	***0.564***	***0.558***	***-0.089***	***0.580***	***0.547***	***-0.112***
13	***0.574***	***0.713***	***0.157***	***0.564***	***0.609***	***0.039***	***0.558***	***0.583***	***-0.001***	***0.554***	***0.552***	***-0.033***
14	***0.519***	***0.644***	***0.147***	***0.517***	***0.551***	***0.020***	***0.517***	***0.533***	***-0.019***	***0.524***	***0.517***	***-0.043***
15	***0.480***	***0.590***	***0.136***	***0.483***	***0.478***	***-0.054***	***0.489***	***0.472***	***-0.103***	***0.515***	***0.470***	***-0.134***
16	0.321	0.445	0.139	0.303	0.446	0.257	0.281	0.419	0.309	0.182	0.365	0.387
17	0.296	0.416	0.137	0.278	0.413	0.245	0.255	0.383	0.288	0.164	0.333	0.359
18	0.496	0.658	0.169	0.469	0.686	0.386	0.434	0.647	0.469	0.279	0.565	0.594
19	0.665	0.874	0.192	0.626	0.929	0.533	0.570	0.870	0.644	0.353	0.752	0.809
20	0.602	0.796	0.181	0.564	0.848	0.503	0.512	0.787	0.601	0.311	0.673	0.744
21	0.656	0.862	0.193	0.618	0.916	0.523	0.564	0.857	0.632	0.351	0.743	0.796
22	***0.543***	***0.678***	***0.154***	***0.532***	***0.569***	***0.026***	***0.526***	***0.541***	***-0.025***	***0.530***	***0.514***	***-0.065***
23	***0.606***	***0.746***	***0.154***	***0.591***	***0.619***	***-0.002***	***0.578***	***0.584***	***-0.059***	***0.577***	***0.547***	***-0.107***
24	***0.466***	***0.578***	***0.136***	***0.462***	***0.473***	***-0.019***	***0.465***	***0.459***	***-0.067***	***0.482***	***0.443***	***-0.113***
25	***0.560***	***0.685***	***0.143***	***0.548***	***0.546***	***-0.059***	***0.542***	***0.524***	***-0.118***	***0.554***	***0.499***	***-0.165***

Note. Exposed items are emphasized.

Figure 1

Example 3-PL Item Response Function

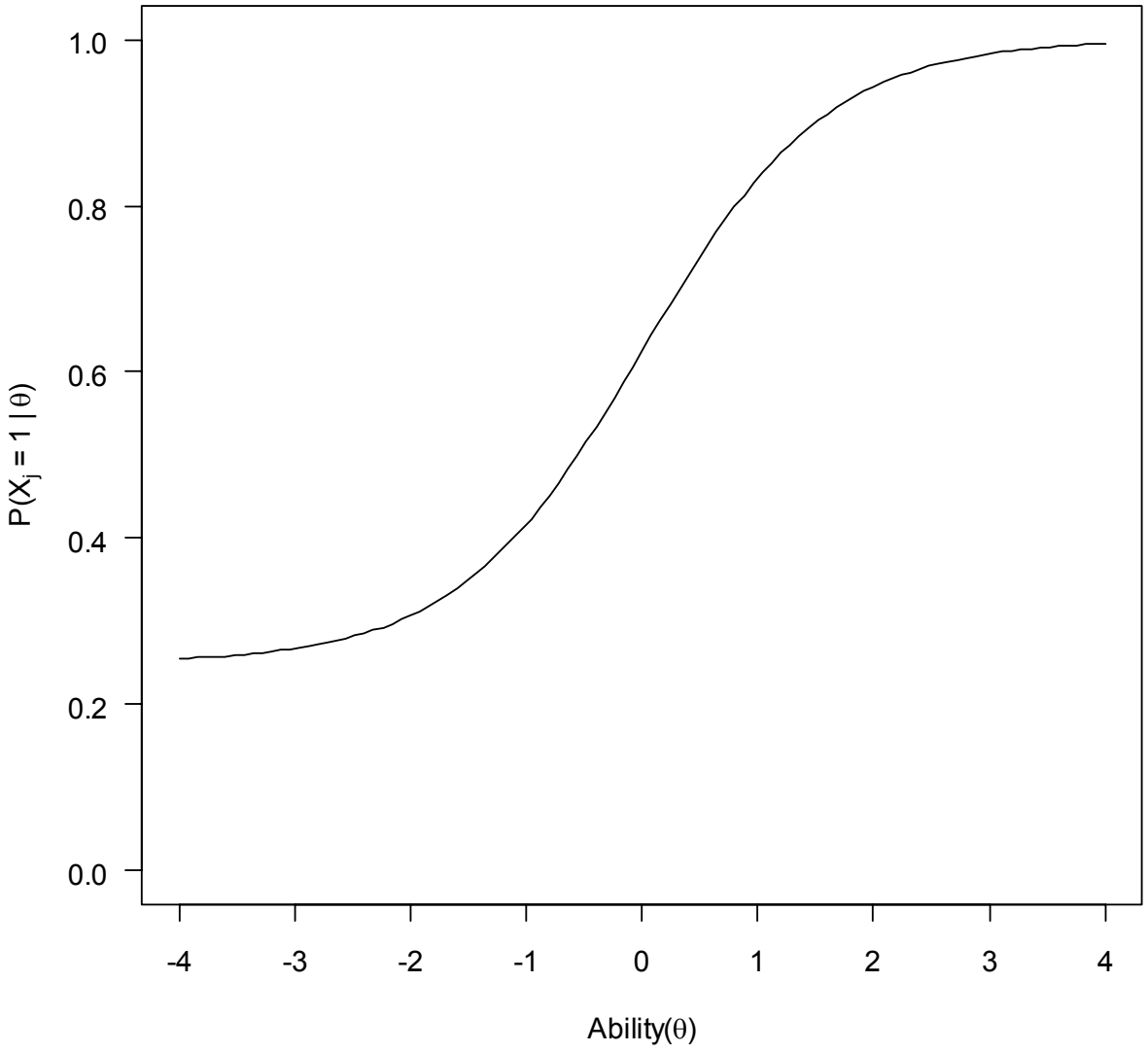

Figure 2

Example Item Response Function With Both $P_j(\theta)$ and $Q_j(\theta)$ Shown

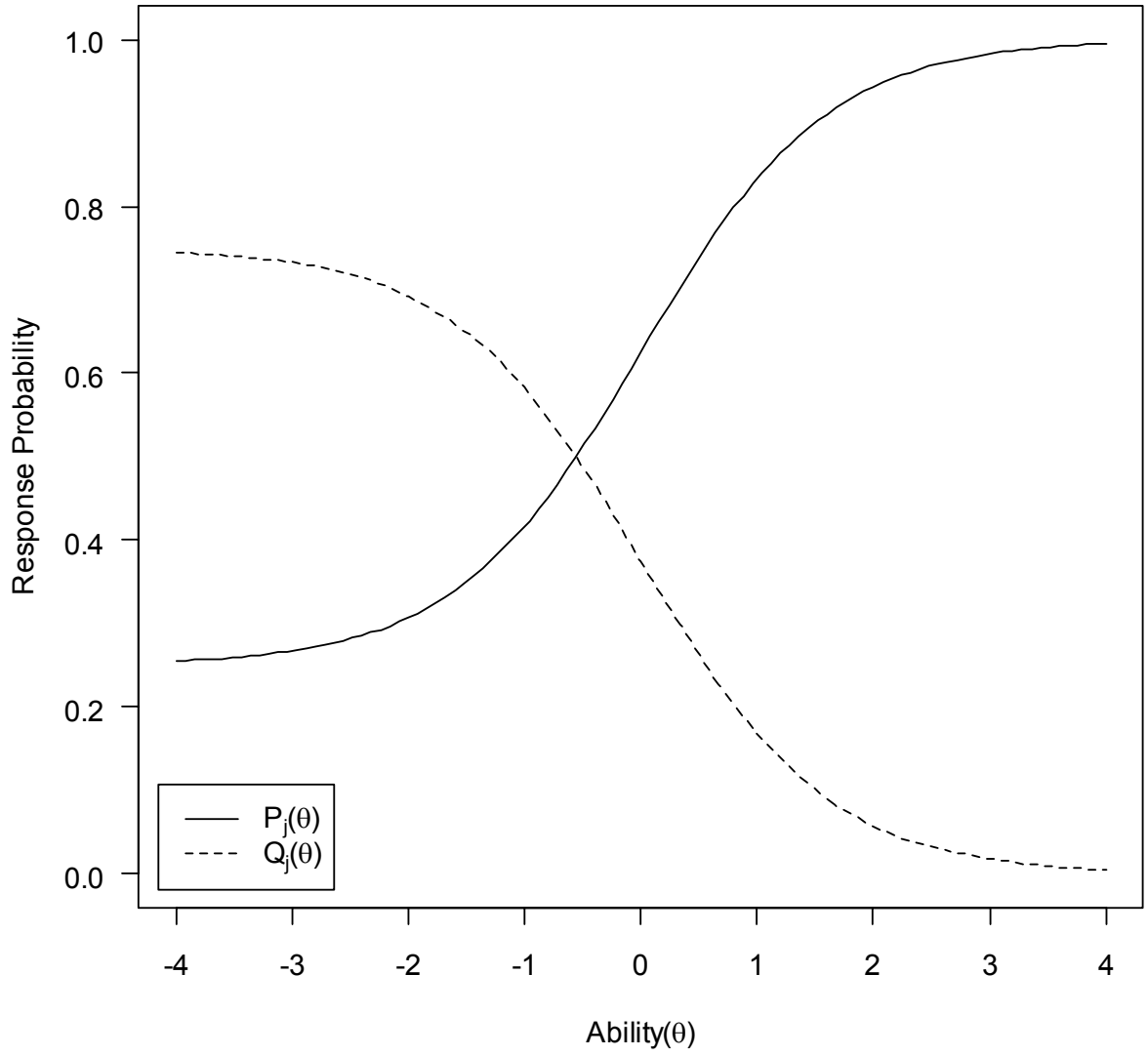

Figure 3

Cumulative Category Response Function for Graded Response Model

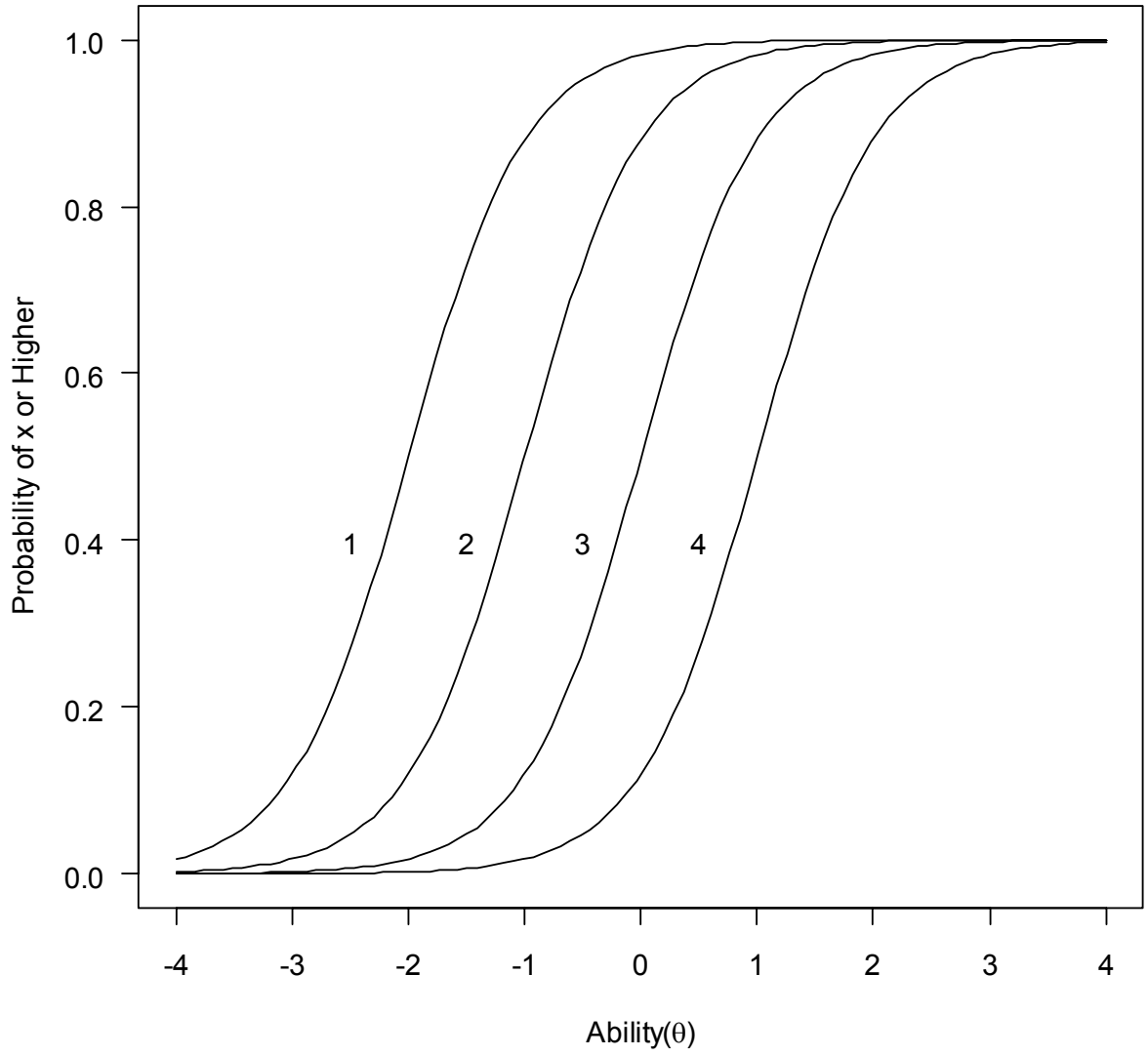

Figure 4

Score Category Response Function for Graded Response Model

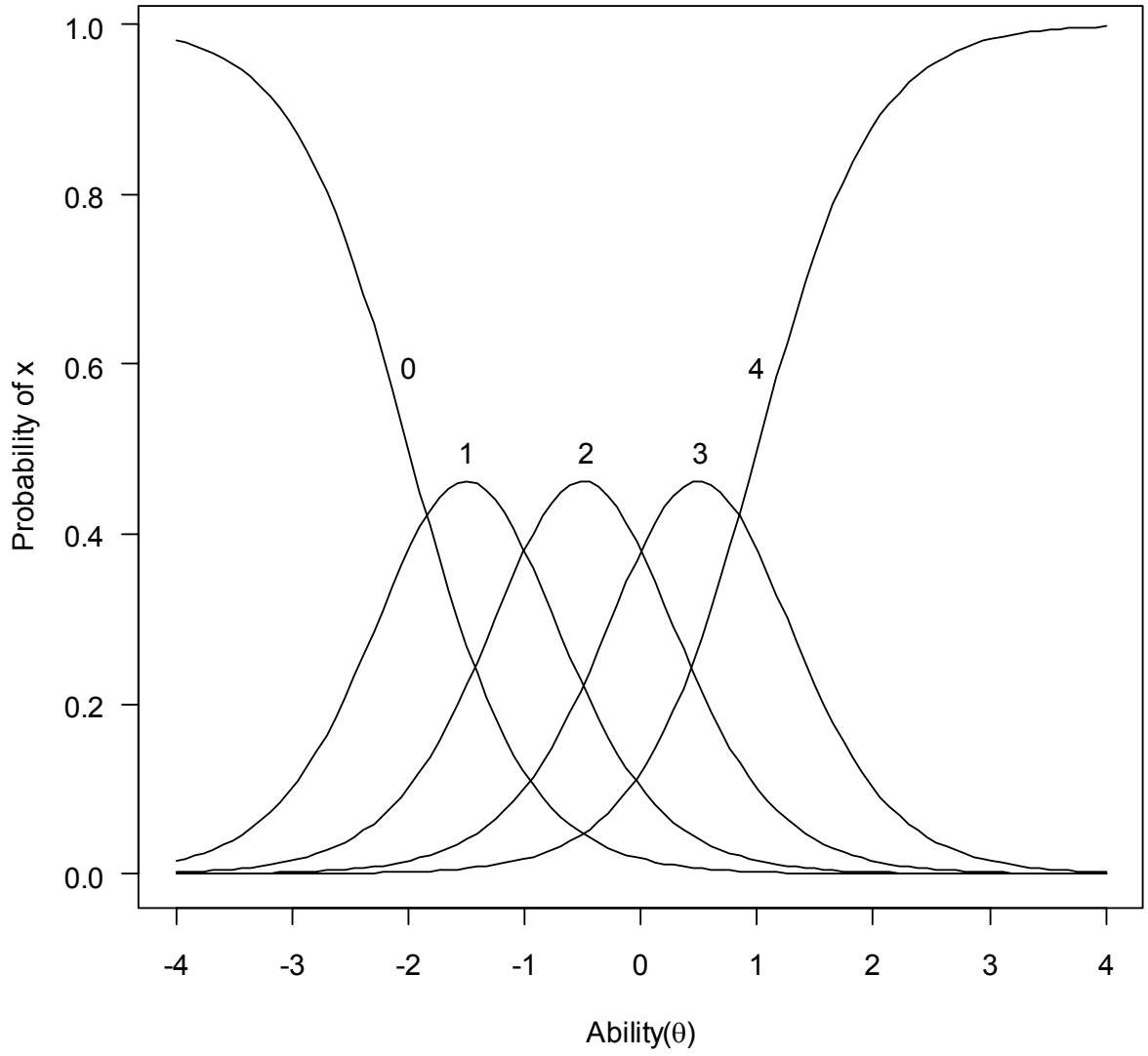

Figure 5

Item Response Functions for Five Example Items

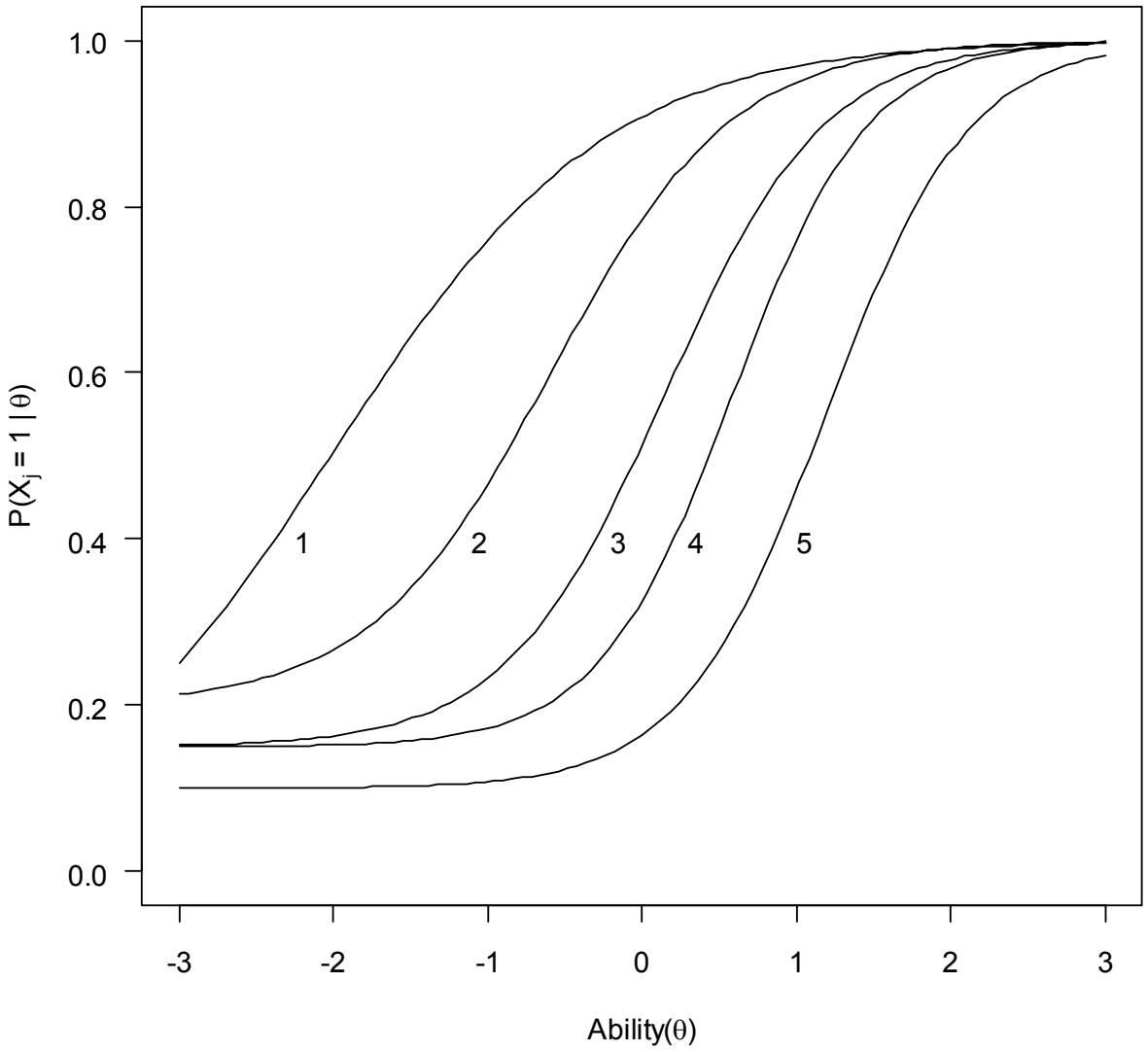

Figure 6

Log-Likelihood Function for a Non-Aberrant Response Pattern

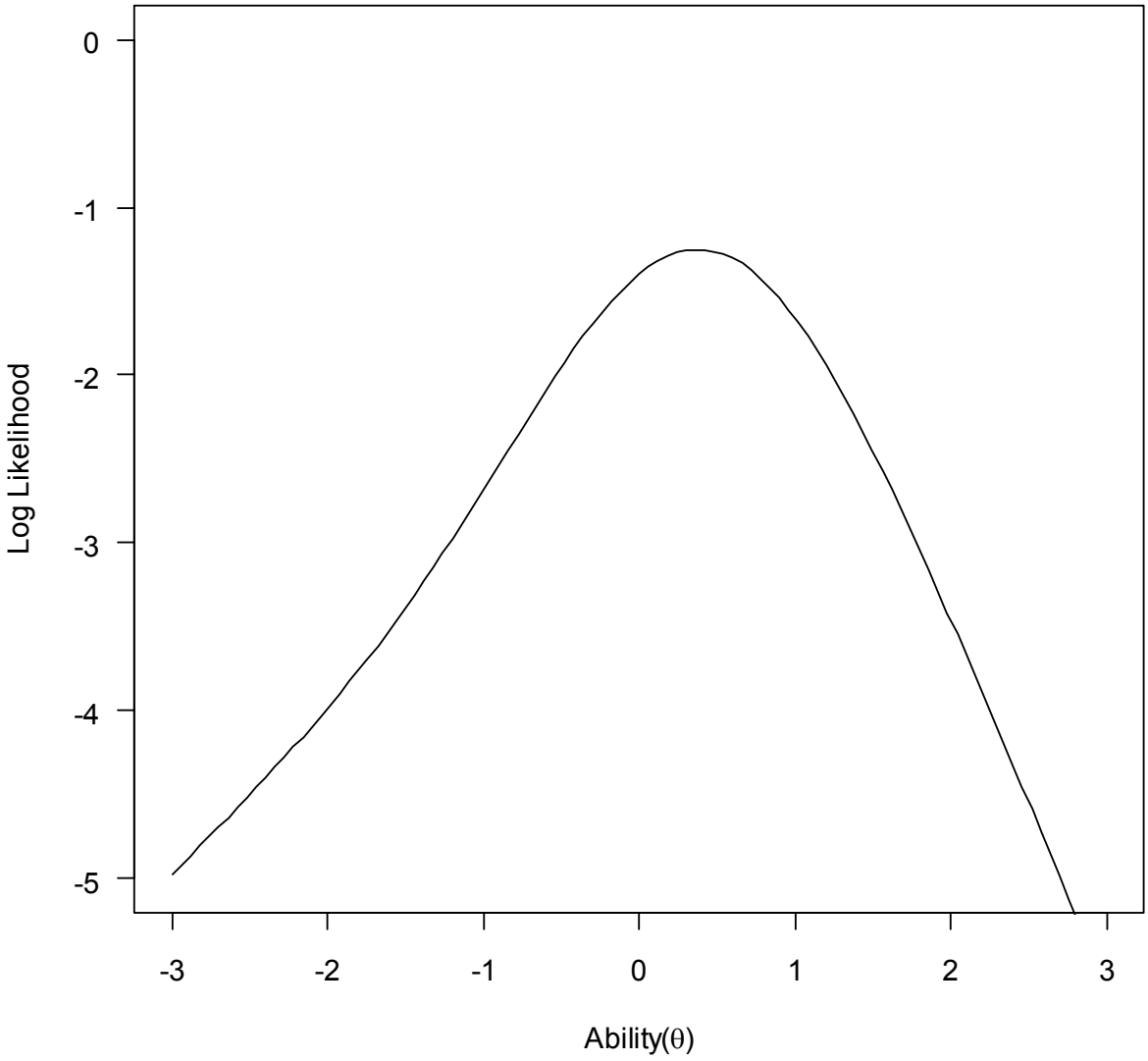

Figure 7

Log-Likelihood Function for an Aberrant Response Pattern

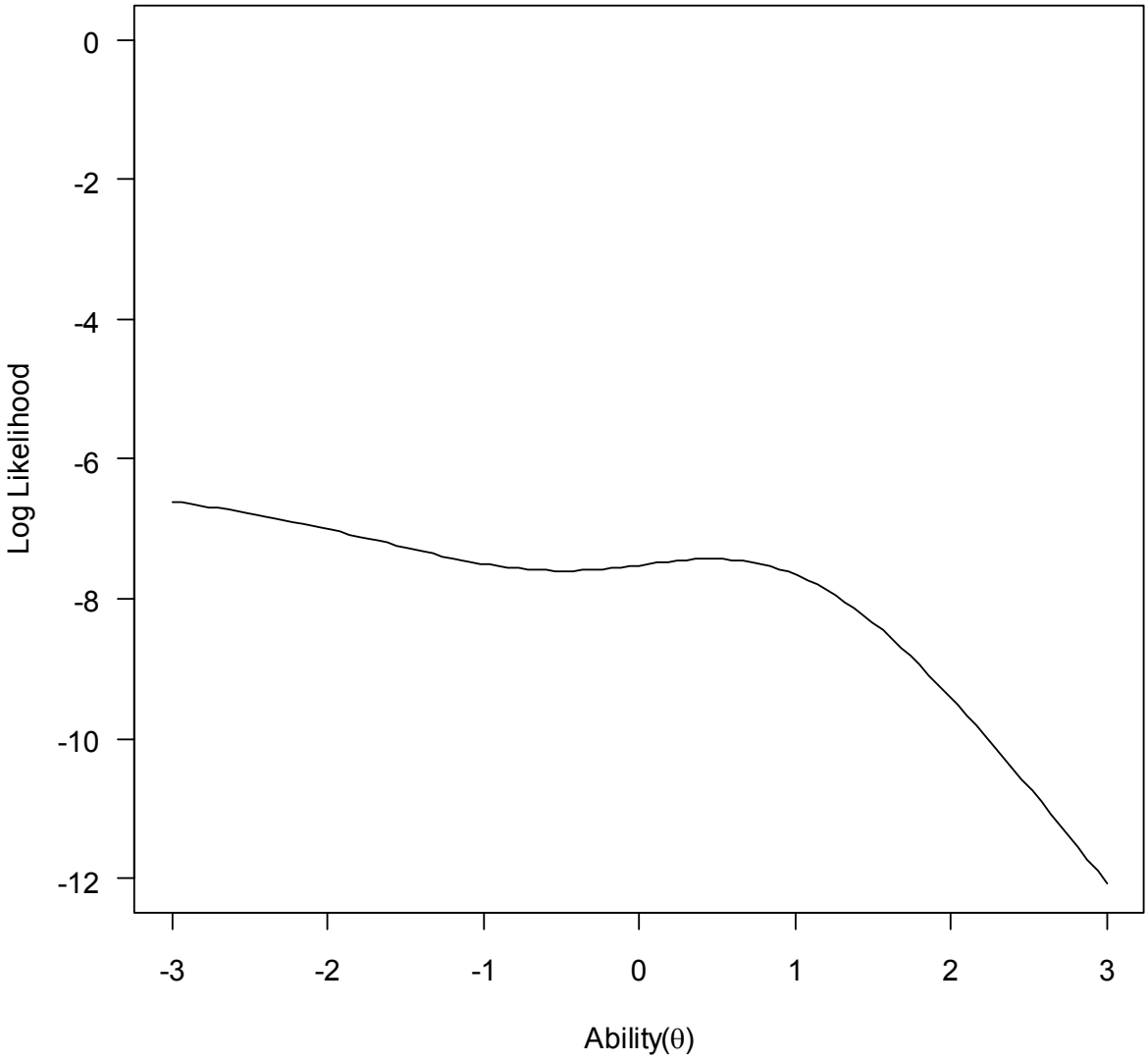

Figure 8

IRF With an Item Response That Is Consistent With Expectation

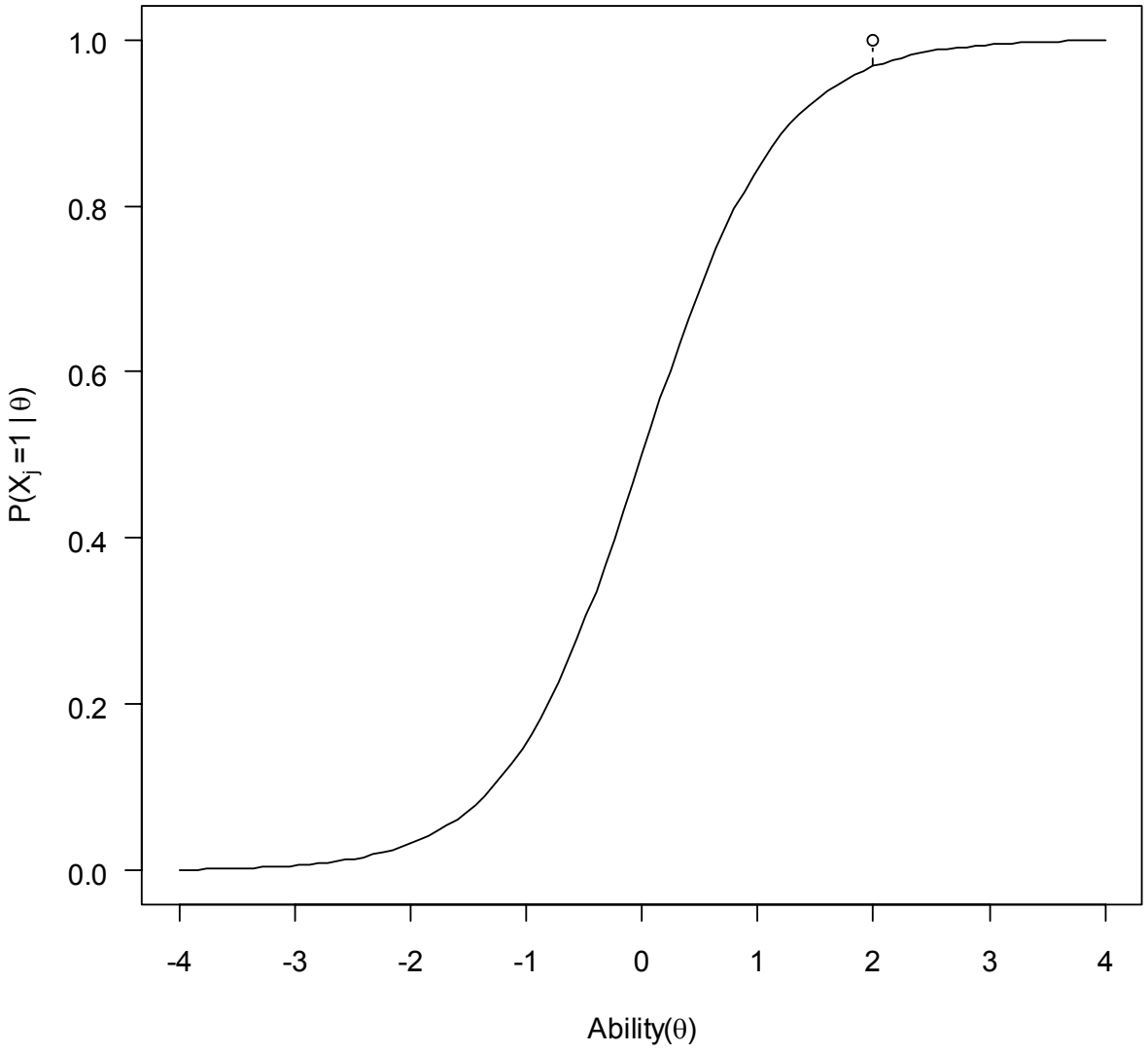

Figure 9

IRF With an Item Response That Is Less Consistent With Expectation

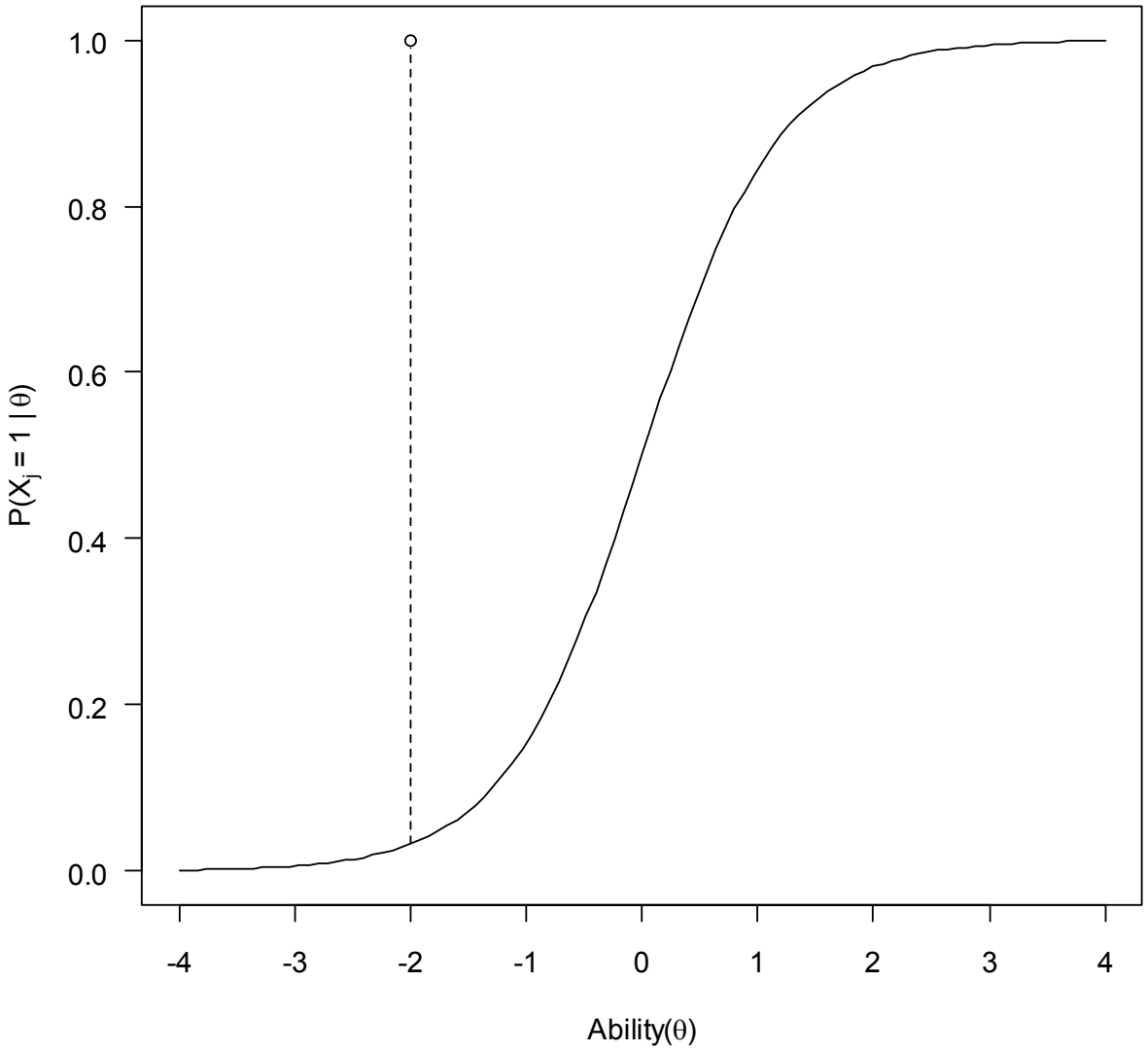

Figure 10

Person Response Function for a Non-Aberrant Response Vector

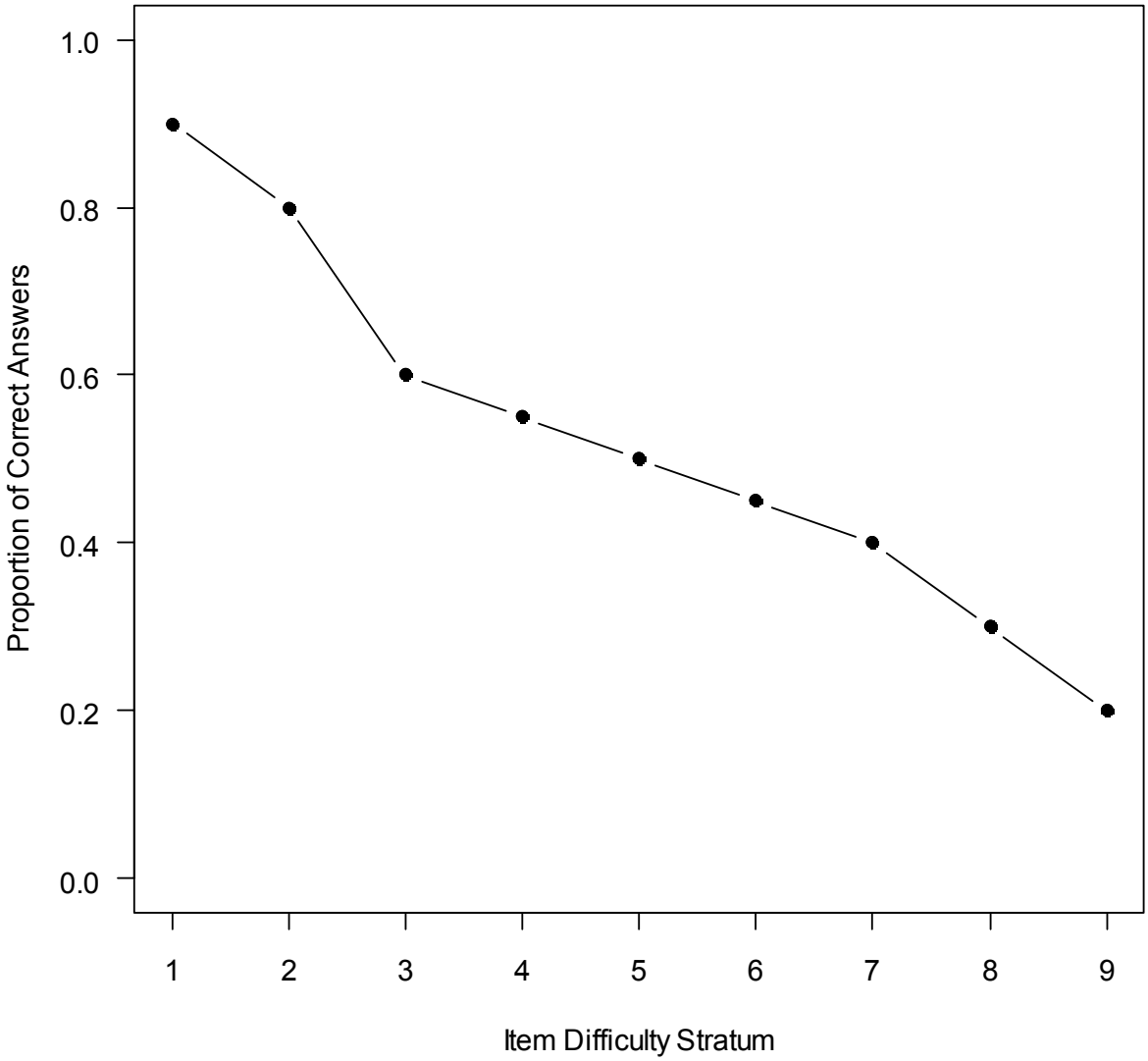

Figure 11

Person Response Function for an Aberrant Response Vector

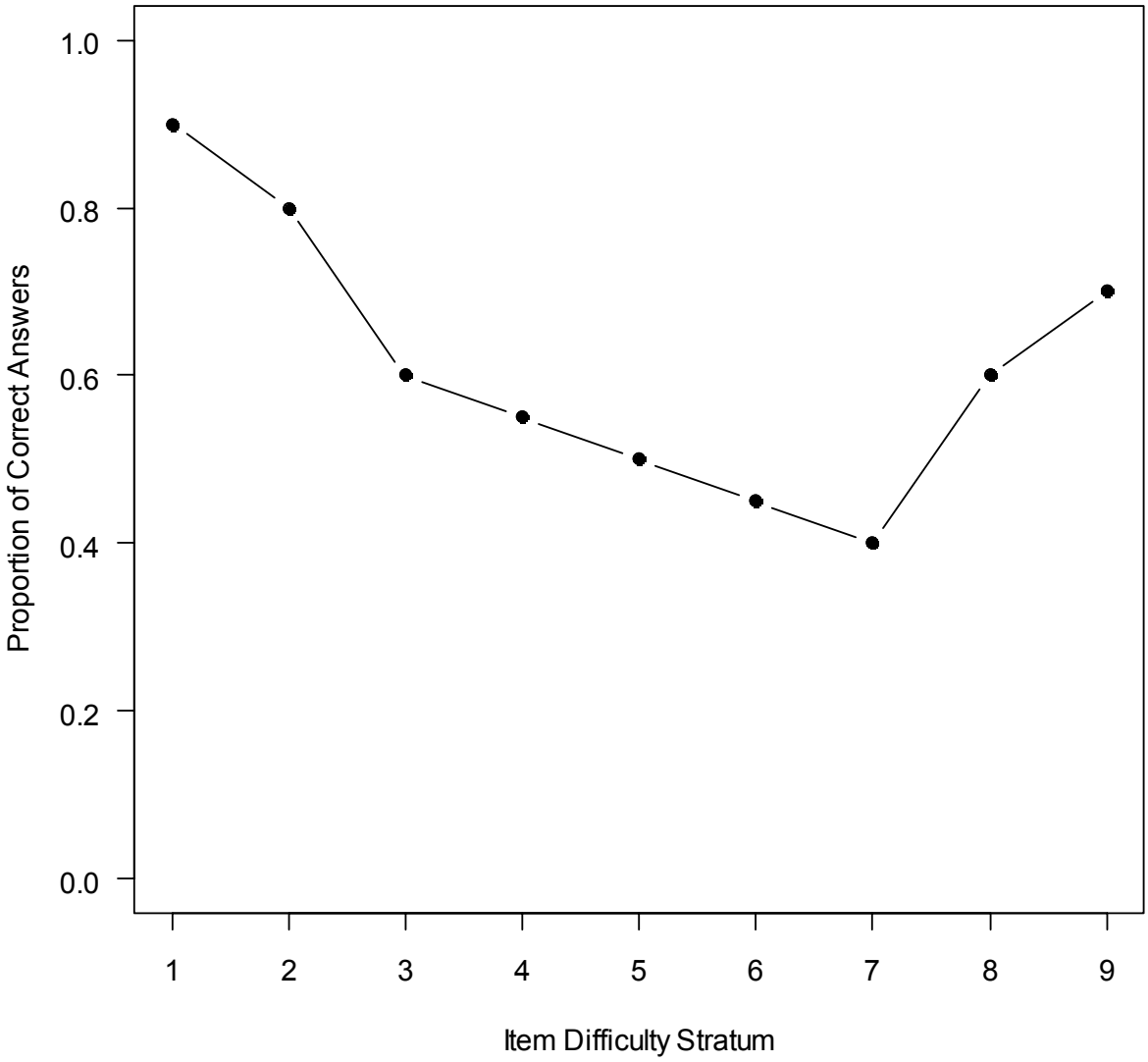

Figure 12

Type I Error Rates for *lcz* and *lco* Difference Across Examinee Ability Strata

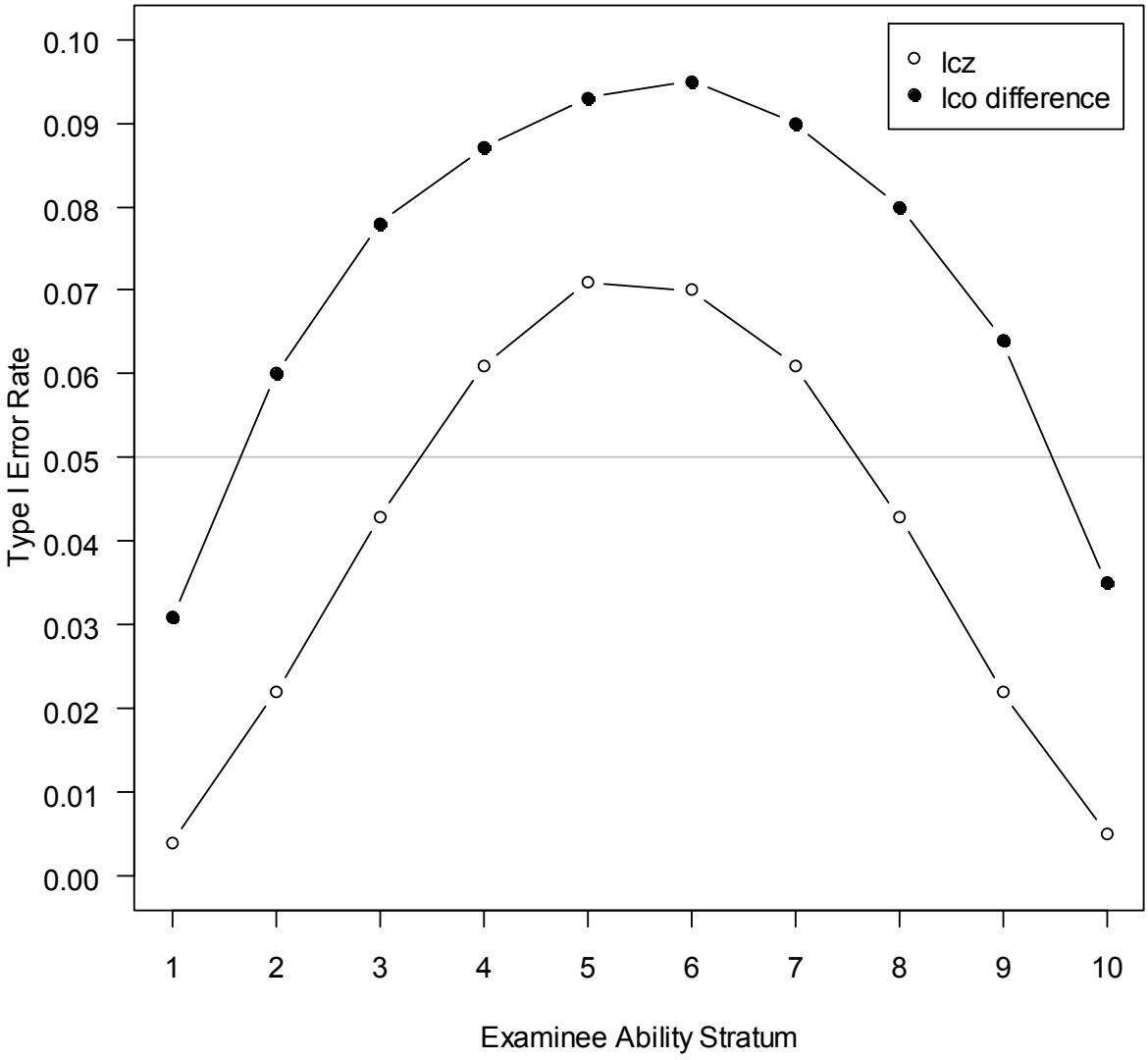

Figure 13

Detection Rates for *lcz* Across Numbers of Exposed Items

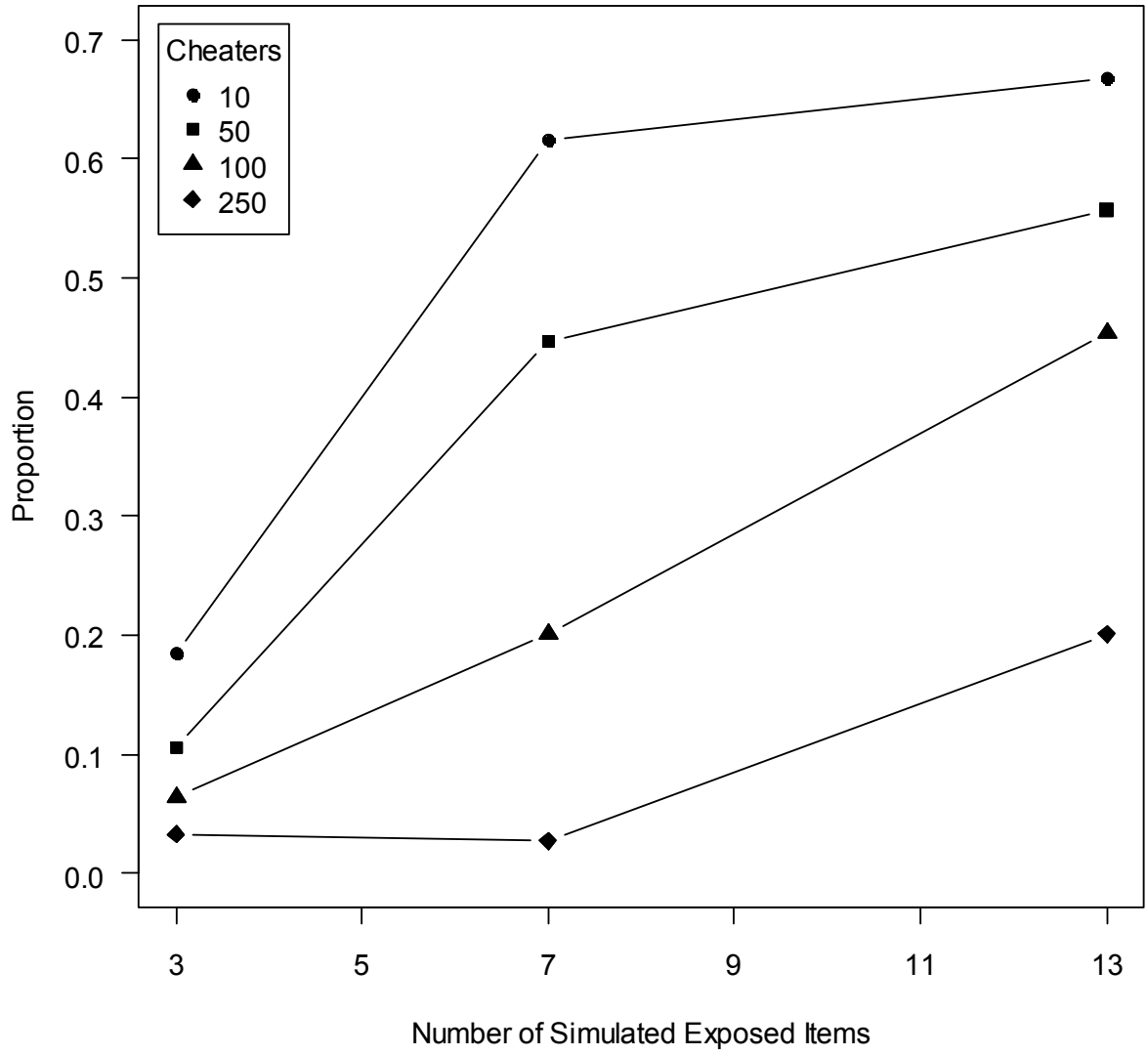

Figure 14

Detection Rates for *lco* Difference Method Across Numbers of Exposed Items

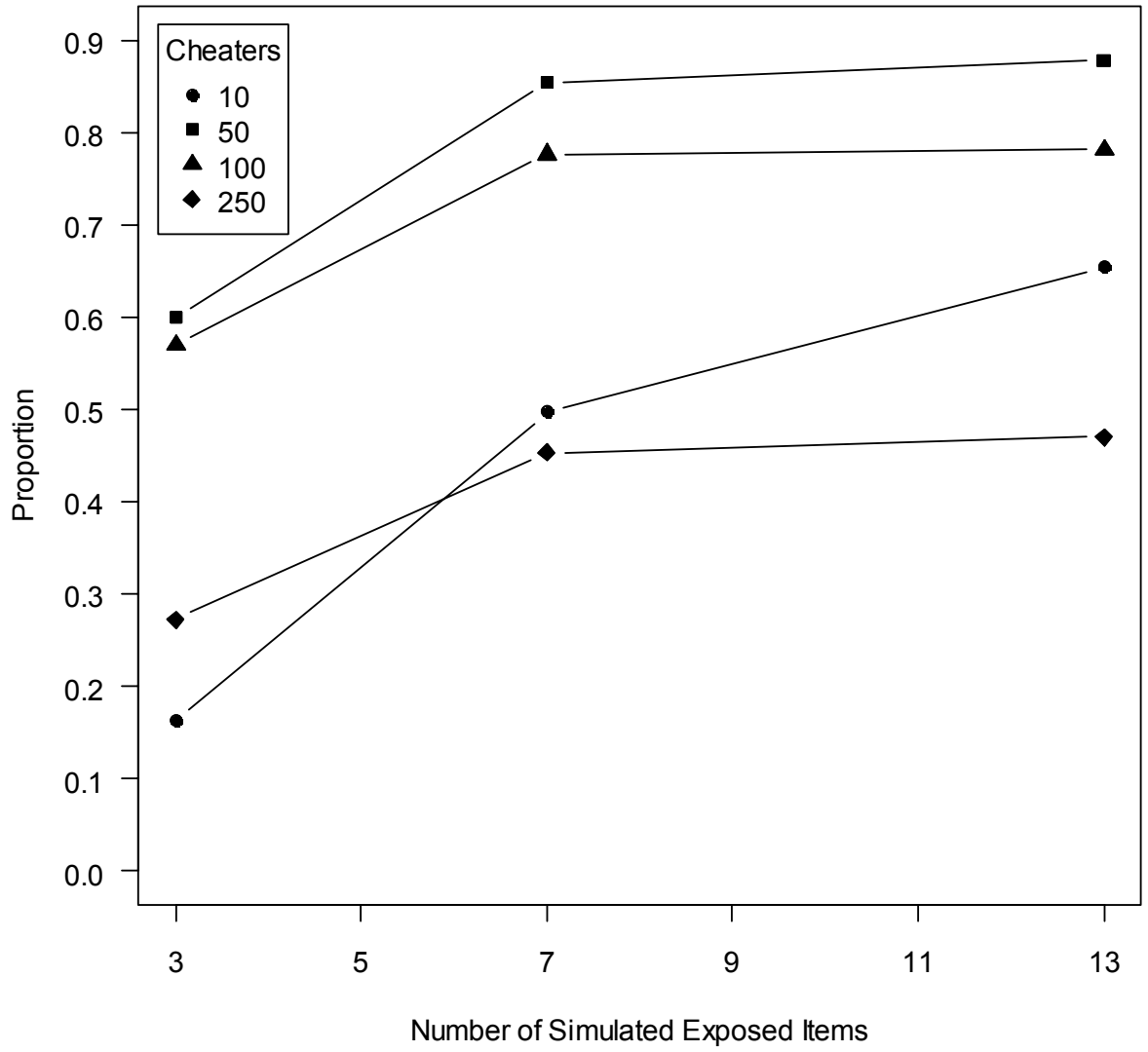

Figure 15

Detection Rates for *lcz* Across Numbers of Cheaters

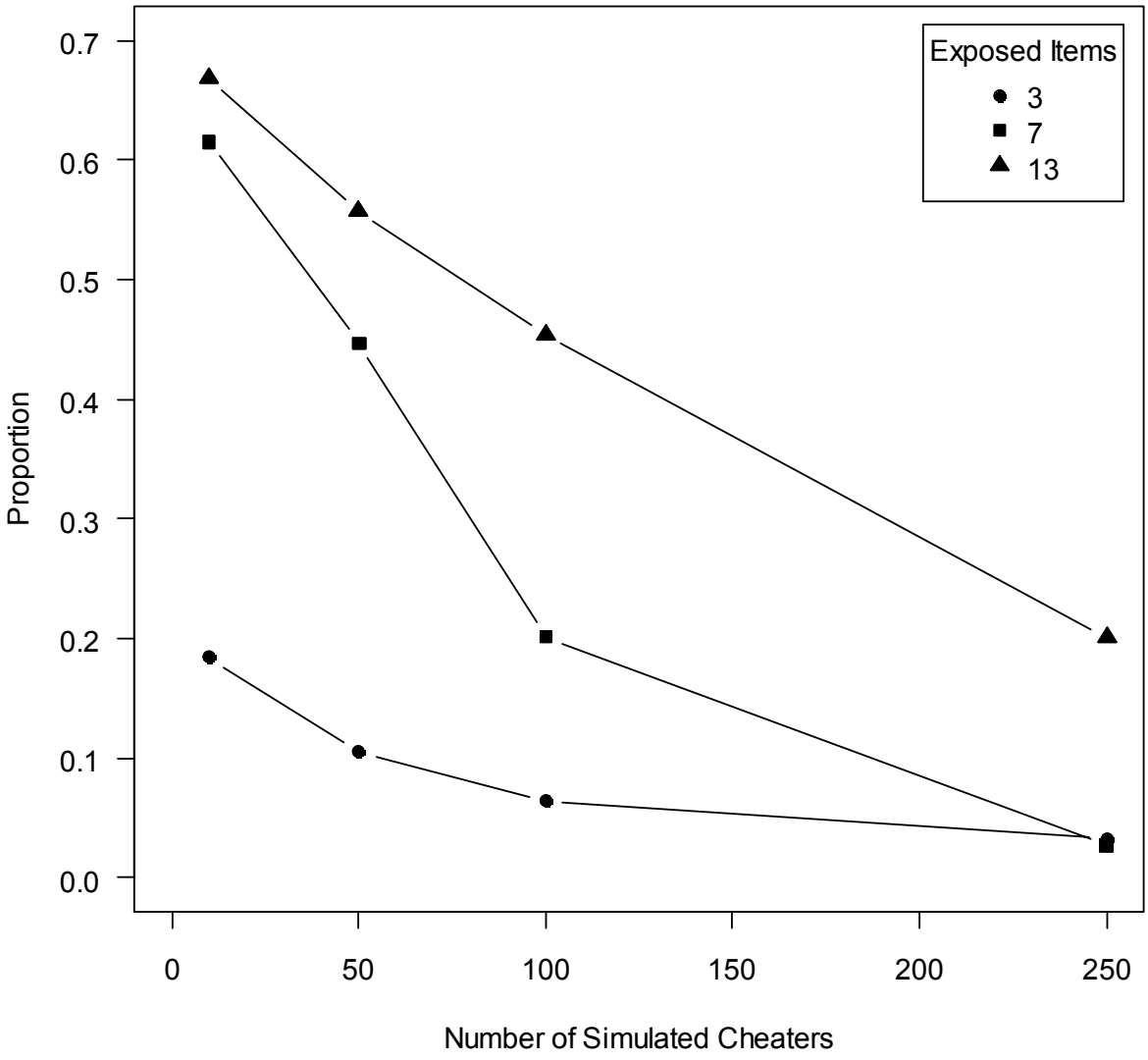

Figure 16

Detection Rates for *lco* Difference Across Numbers of Cheaters

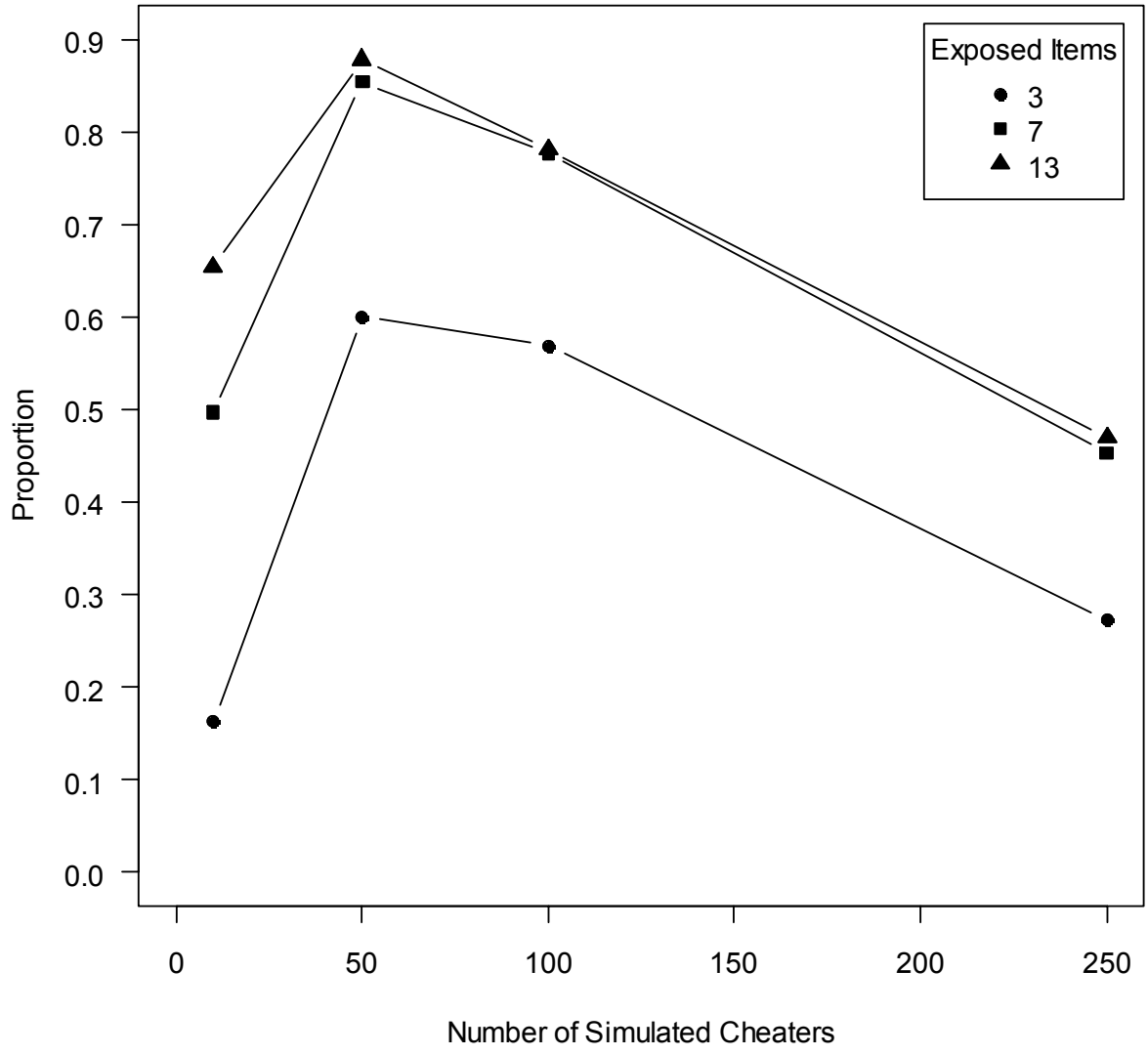

Figure 17

Comparison of Detection Rates for Conditions With 10 Cheaters

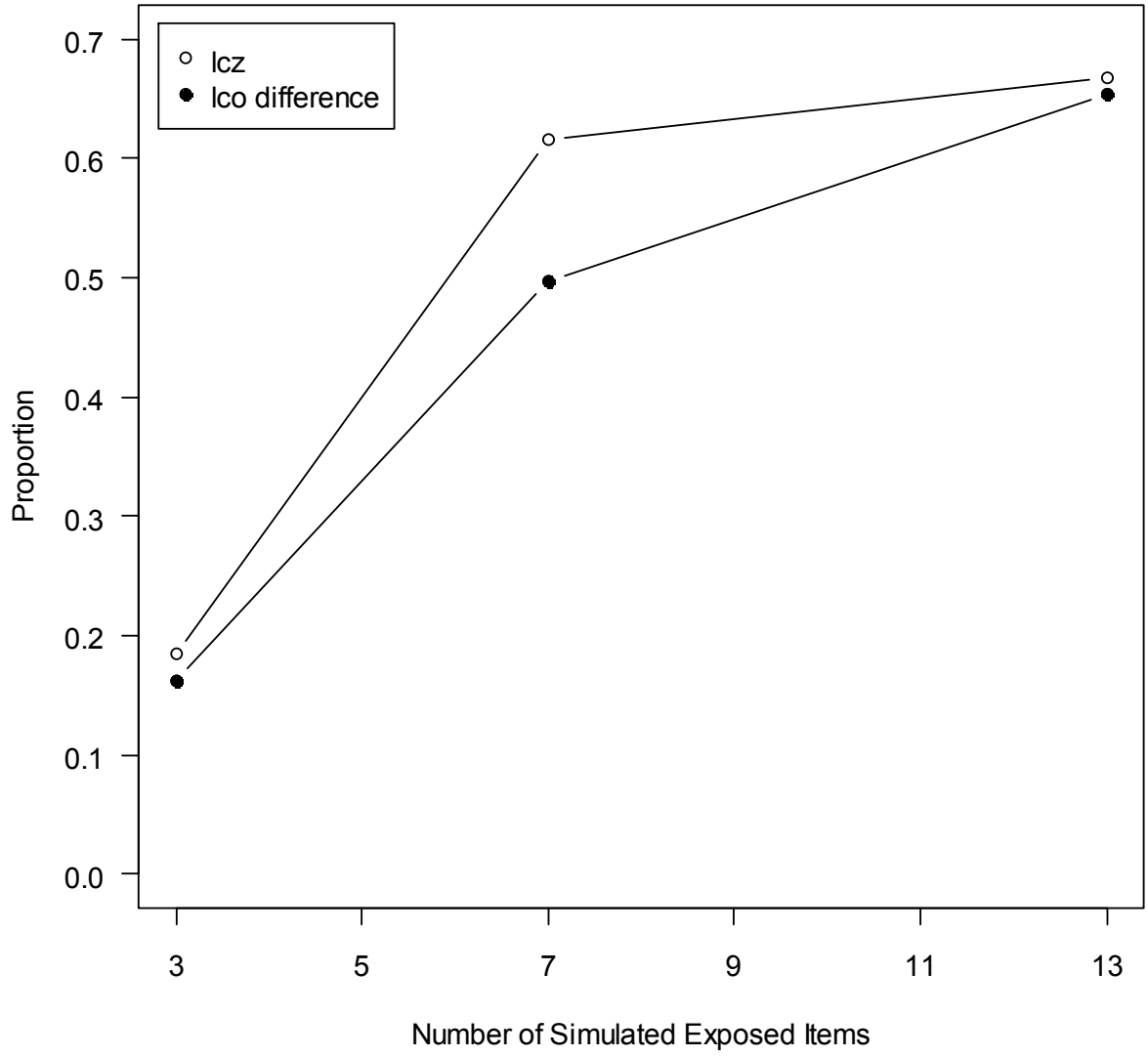

Figure 18

Comparison of Detection Rates for Conditions With 50 Cheaters

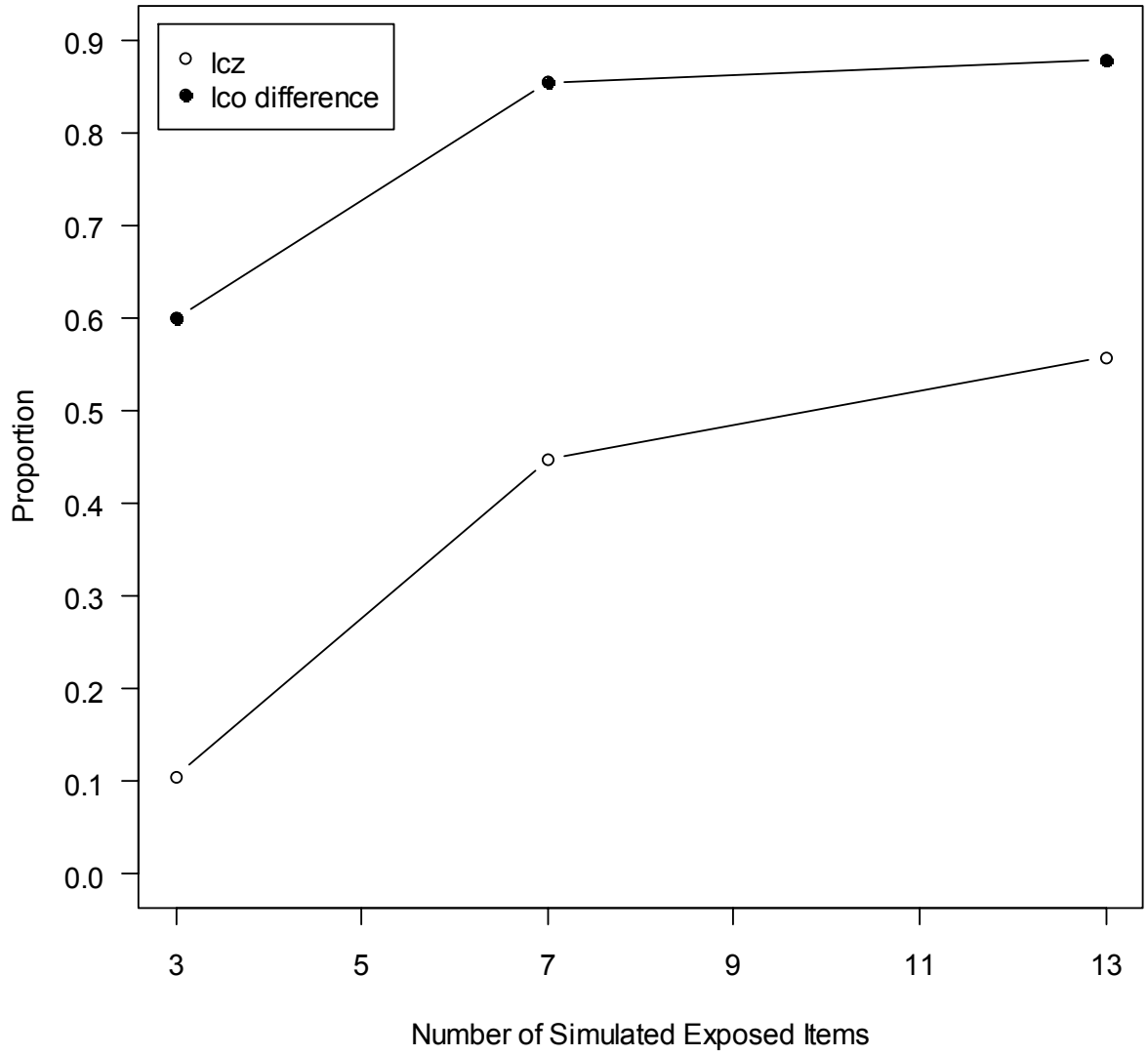

Figure 19

Comparison of Detection Rates for Conditions With 100 Cheaters

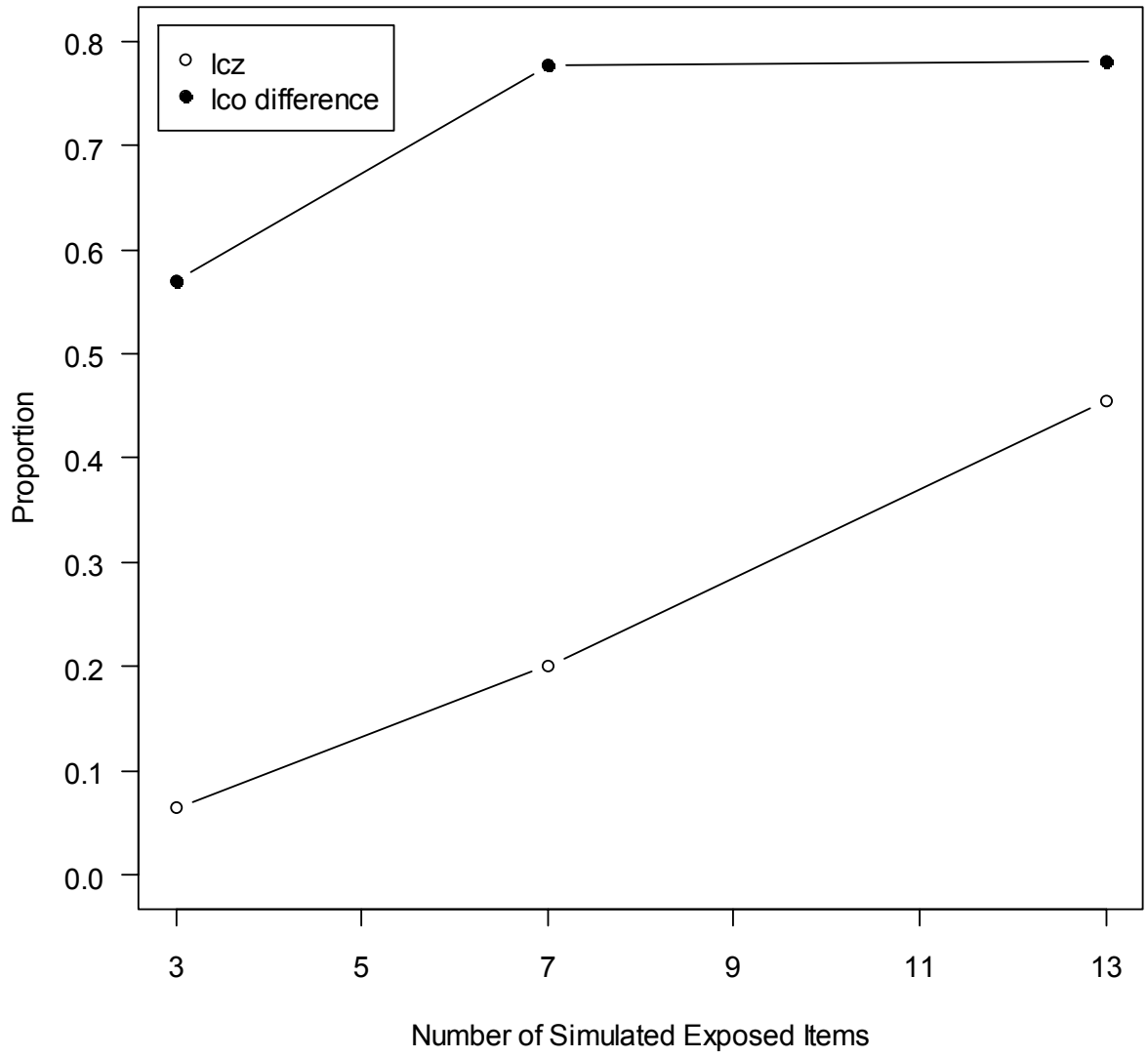

Figure 20

Comparison of Detection Rates for Conditions With 250 Cheaters

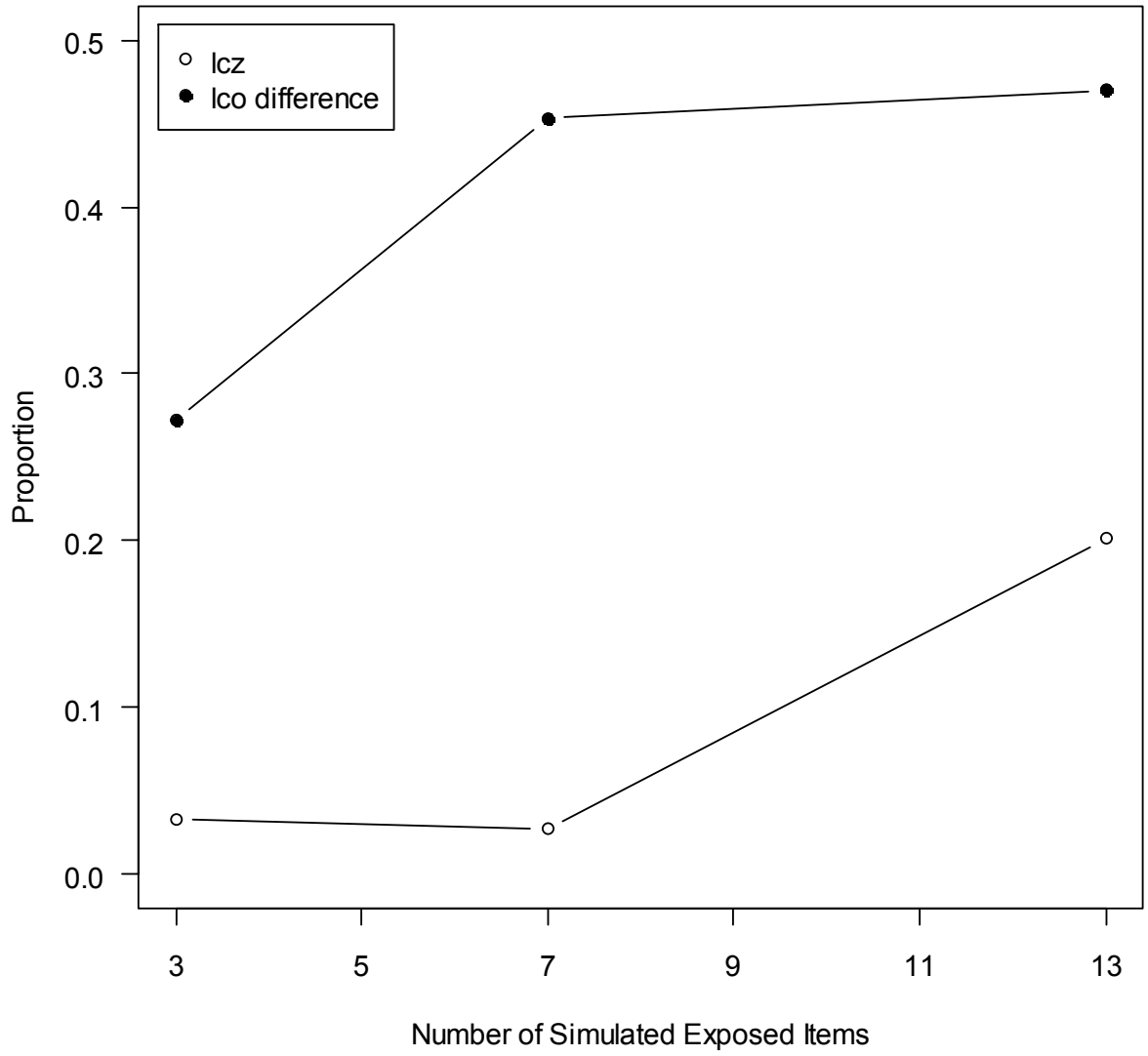

Figure 21

Comparison of Detection Rates for Conditions With 3 Exposed Items

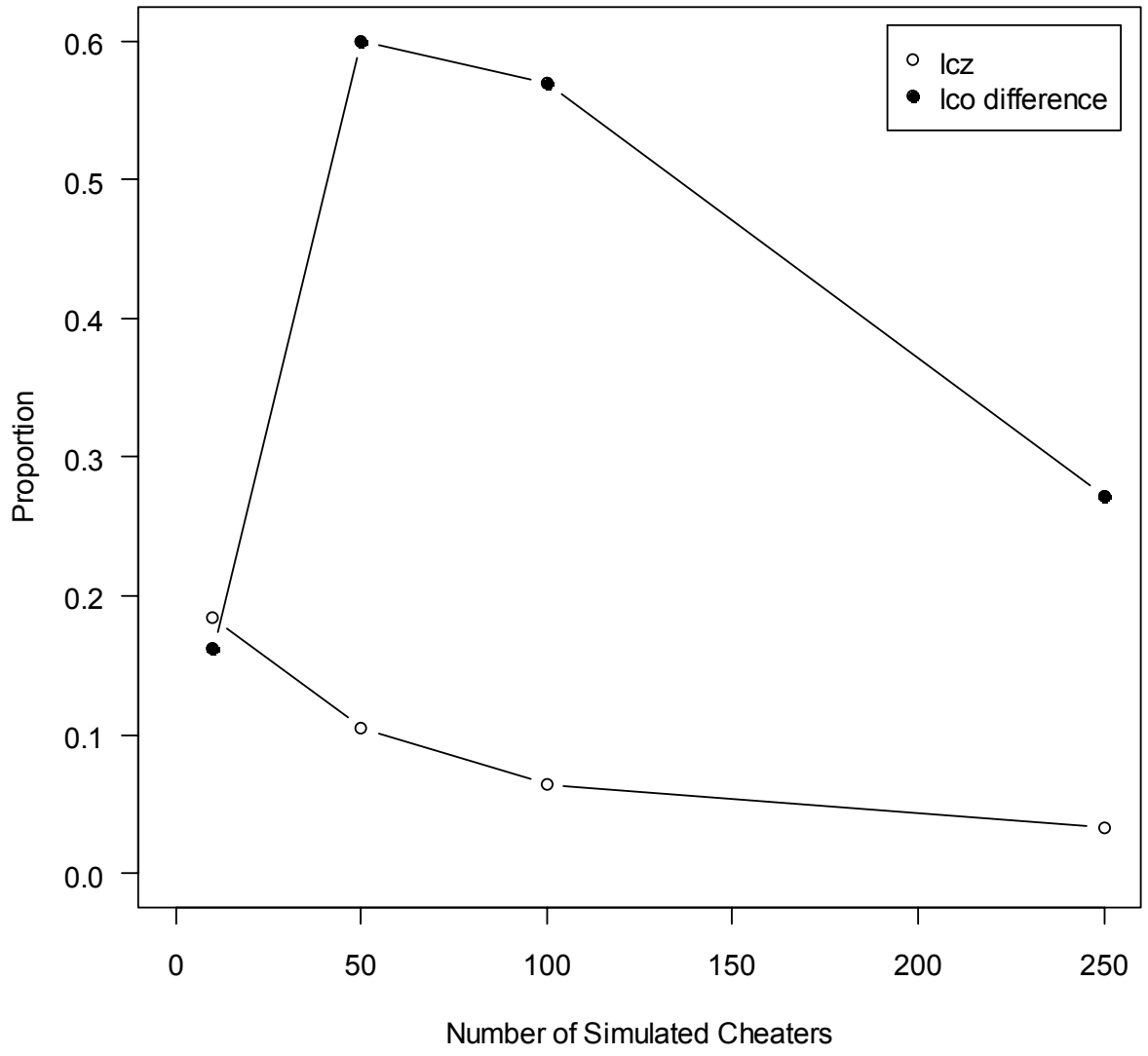

Figure 22

Comparison of Detection Rates for Conditions With 7 Exposed Items

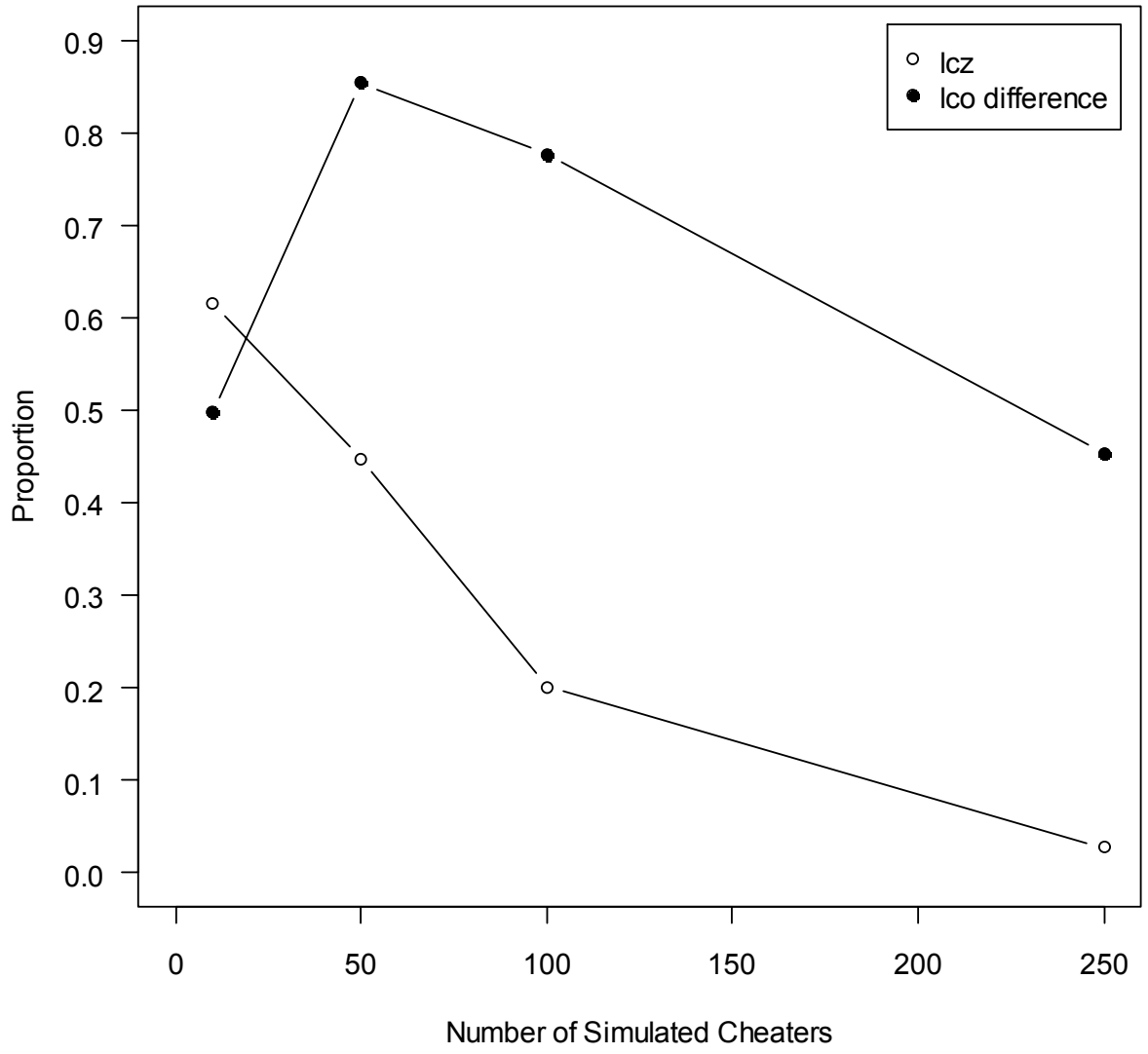

Figure 23

Comparison of Detection Rates for Conditions With 13 Exposed Items

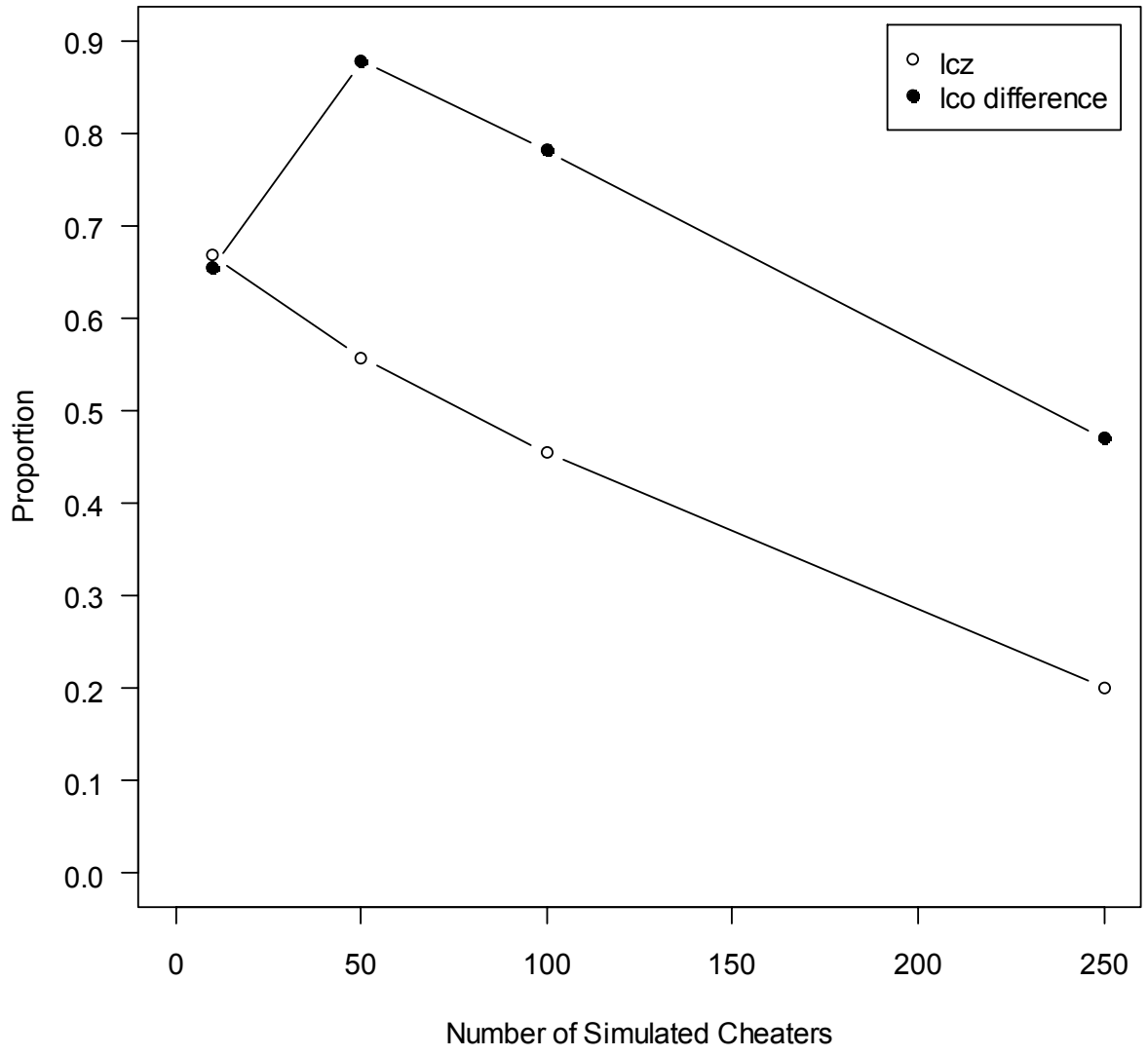

Figure 24

Distribution of *lco* Differences Across Replications When No Cheaters Are Present

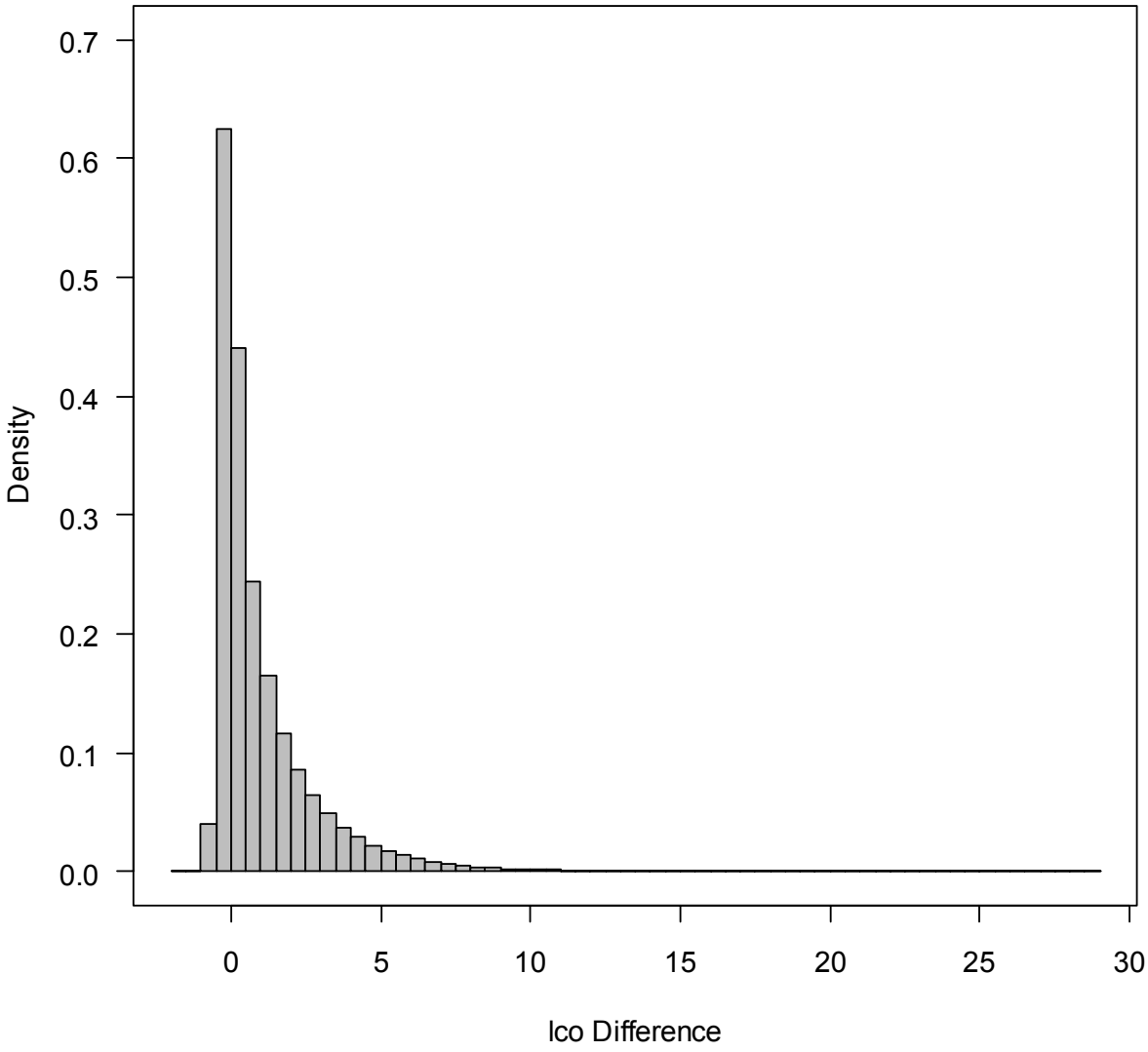

Figure 25

Distribution of *lcz* Across Replications When No Cheaters Are Present

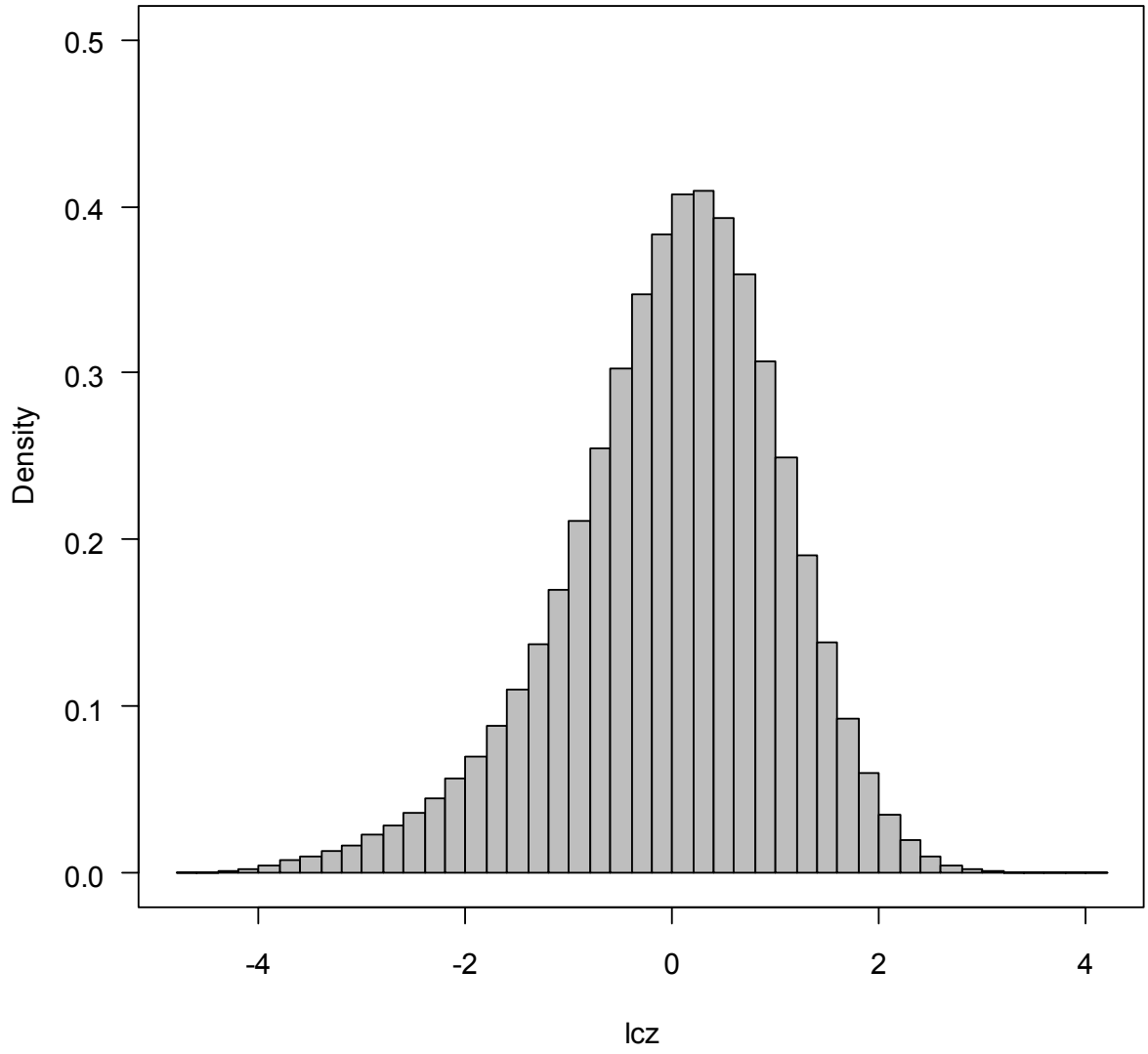

Figure 26

Distribution of Factor Loadings for Item 12: Condition With 7 Exposed Items and 50 Cheaters

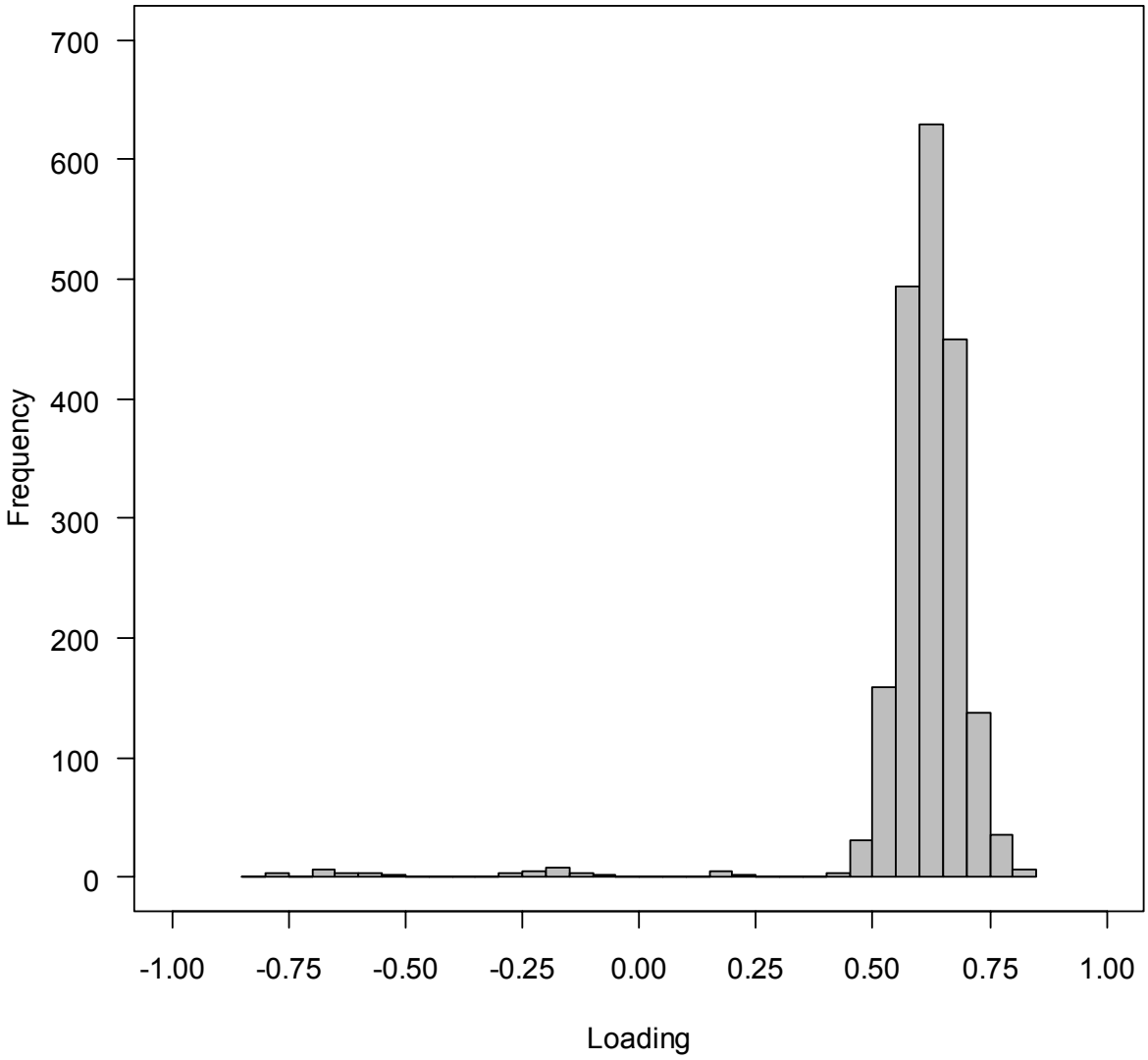

Figure 27

Distribution of Factor Loadings for Item 12: Condition With 7 Exposed Items and 10 Cheaters