ABSTRACT

Title of dissertation: DETECTING LOCAL ITEM DEPENDENCE IN POLYTOMOUS ADAPTIVE DATA

Jessica L. Mislevy, Doctor of Philosophy, 2011

Dissertation directed by: Professor Jeffrey R. Harring
Professor André A. Rupp
Department of Measurement, Statistics, and Evaluation

A rapidly expanding arena for item response theory (IRT) is in attitudinal and health-outcomes survey applications, often with polytomous items. In particular, there is interest in computer adaptive testing (CAT). Meeting model assumptions is necessary to realize the benefits of IRT in this setting, however. Although initial investigations of local item dependence (LID) have been studied both for polytomous items in fixed-form settings and for dichotomous items in CAT settings, there have been no publications applying LID detection methodology to polytomous items in CAT despite its central importance to these applications. The research documented herein investigates the extension of widely used methods of LID detection, Yen's Q_3 statistic and Pearson's Statistic X^2, in this context, via a simulation study. The simulation design and results are contextualized throughout with a real item bank and data set of this type from the Patient-Reported Outcomes Measurement Information System (PROMIS).

DETECTING LOCAL ITEM DEPENDENCE IN POLYTOMOUS ADAPTIVE DATA

by

Jessica L. Mislevy

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:

Professor Jeffrey R. Harring, Co-Chair
Professor André A. Rupp, Co-Chair
Professor Gregory R. Hancock
Professor Hong Jiao
Professor Frauke Kreuter

DEDICATION

To my parents, because your endless support,
continual encouragement, and sound
guidance, made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Item Response Theory

Models in item response theory (IRT) mathematically define the probabilistic relationship between individuals' observed responses to a series of items and their location on the unobservable latent variable continua reflecting the constructs being measured (see, e.g., De Ayala, 2009; de Boeck & Wilson, 2004; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Reckase, 2009, for overviews). Designed in the context of educational assessment, IRT has often been used to measure constructs such as math proficiency or reading comprehension. However, interest in the technique has been rapidly increasing beyond educational measurement into the areas of psychological and health-outcomes assessment due to its methodological sophistication and recent technological advances (Chang & Reeve, 2005; Fries, Bruce, & Cella, 2005; Hays, 2004; Reeve & Mâsse, 2004). IRT techniques are now being applied in these fields to measure health status variables reflecting constructs such as depression and fatigue.[1] For example, the National Institutes of Health (NIH) and health-outcomes scientists at institutions across the country have formed a cooperative network to develop the Patient-Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org).

---

[1] For simplicity, the term "individual" will be used throughout the document to denote the person who provides the responses to a data-collection instrument; the term is assumed to be synonymous with alternative terms such as "examinee" and "respondent," though it is acknowledged that each has domain-specific connotations. Along these lines, the data-collection instrument will be referred to as an "instrument," (representing alternative terms such as "test" and "questionnaire") and the term "trait" is used to denote the construct measured by the items (representing alternative terms such as "ability" and "proficiency"). Exceptions are made in rare instances when necessary to remain consistent with the literature.

This initiative focuses on more accurate and efficient measurement of patient-reported symptoms and aspects of health-related quality of life. A primary goal of PROMIS is to develop instruments based on IRT methodologies in these domains that are publically available for the clinical research community.

In IRT, estimates of respondents' traits (θ) are based not only on the responses they provide, but also the characteristics (i.e., parameters) of the items they are administered such as their difficulty – reflected by category boundary parameters (b) – their ability to differentiate among respondents – reflected by slope parameters (a) – and their susceptibility to guessing – reflected by lower asymptote parameters (c).

One unidimensional IRT model frequently applied in health-outcomes settings is the graded response model (GRM; Samejima, 1969). The GRM is appropriate for item responses that fall in multiple ordered categories. It predicts the conditional probability of an individual responding in a particular category as a function of an individual's latent trait value and several item parameters.

The GRM is considered a "difference model" (Thissen & Steinberg, 1986) or "indirect IRT model" (Embretson & Reise, 2000) because the probabilities are computed in two stages. Following Dodd, De Ayala, and Koch (1995), the probability that individual i will produce a response in category k or higher for item j is first computed as:

$$P_{jk}^{*}(\theta_i) = \frac{\exp[a_j(\theta_i - b_{jk})]}{1+\exp[a_j(\theta_i - b_{jk})]}, \qquad (1)$$

where θ_i is the individual's trait level, a_j is the discrimination parameter for item j, and b_{jk} is the category boundary for category k for item j. In the second step, the actual category

response probabilities for each response are computed by subtracting the cumulative probabilities from adjacent response categories conditional on θ using:

$$P_{jk} = P^{*}_{jk}(\theta_i) - P^{*}_{j,k+1}(\theta_i). \quad (2)$$

There are a total of $K - 1$ boundary parameters that need to be estimated for an item with K score categories.

Figure 1 illustrates the probability of choosing each of the response options offered with an item from PROMIS designed to measure fatigue impact and experience according to the GRM. It shows that a person with a low level of fatigue would have a high probability of indicating that his fatigue made it "not at all hard to carry on a conversation." In contrast, an individual with a high level of fatigue would be likely to say "very much" when asked about how hard her fatigue made it for her to carry on conversations within the last week.

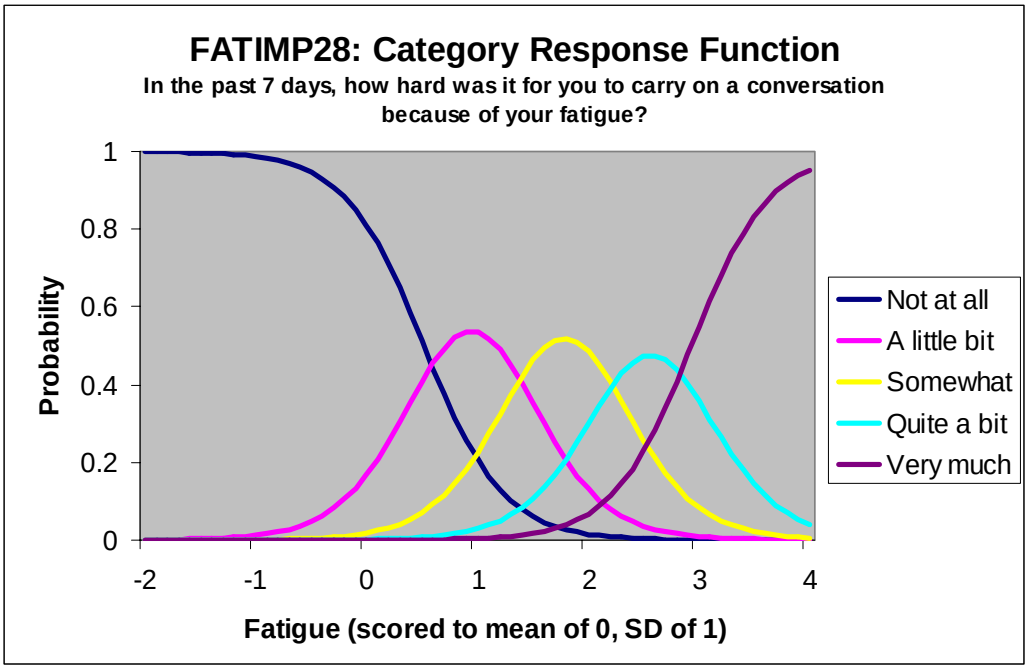


Figure 1. Category response function for an item from the PROMIS Fatigue Bank

One of the biggest advantages associated with IRT is that, if an appropriate model holds, researchers can compare individuals who have answered different items on different forms of an instrument as well as the items across the different instruments on a common scale (Embretson & Reise, 2000). As exemplified in Equation 1, the operating characteristics of items such as their difficulty and latent traits of individuals are estimated via separate parameters in the model. Thus, item parameters are theoretically invariant across different groups of individuals and person parameters are invariant across different sets of items under conditions of perfect model fit.

Computer Adaptive Testing

Coupled with modern information technology, IRT models are particularly well suited for computer adaptive testing (CAT) due to their ability to obtain comparable trait estimates independent of the set of items administered (Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990; Hambleton, Swaminathan, & Rogers, 1991). Given a bank of items with known properties, CAT algorithms identify an individually-tailored set of items for each individual that provides the most precise information about that particular individual's location along the latent dimension(s). In CAT, an individual's trait level is iteratively estimated during the administration process, continually updated based on the responses he or she provides. The next item for administration is selected based on the current trait estimate. This process is repeated until a stopping criterion is met, such as a predetermined level of precision, maximum number of items, or time limit. Because individuals are given only those items that are most relevant to their currently estimated trait level at any given point in time, instruments administered with CAT algorithms generally obtain a higher level of precision with fewer items than instruments

administered in non-adaptive settings (e.g., Embretson & Reise, 2000, p.268; Meijer & Nering, 1999; Moreno & Segall, 1997; Weiss, 1982).

Model Assumptions

To realize the potential that CAT offers, however, critical assumptions of the IRT model it is based on must be reasonably met. Local / conditional item independence (LII) is one of the most fundamental assumptions of IRT. LII assumes that, holding trait level constant, responses to any pair of items are statistically independent (Hambleton & Swaminathan, 1985). When LII holds, the probability of observing a particular response pattern for an individual is equal to the product of the probabilities of the observed responses on individual items.

When LII does not hold, item pairs or groups are said to exhibit local / conditional item dependence (LID). In this case, item responses are interrelated even after the latent trait is taken into account. There are a number of reasons responses to pairs or groups of items will violate the LII assumption (e.g., Yen, 1993) and cause LID. Items may exhibit LID for content reasons; that is, they exhibit LID because they are based on a common stimulus, contain similar, finely specified content, or give away the answer to one another. Items may also exhibit LID because of the order in which they are administered. For example, responses may be subject to practice or fatigue effects, or speededness.

Violations. IRT models may not be robust to violations of the LII assumption. When LID is present, studies have shown parameter estimates and test statistics produced by IRT models can be negatively impacted (e.g., Chen & Thissen, 1997; Chen & Wang, 2007; Kingston & Dorans, 1984; Thompson & Pommerich, 1996; Tuerlinckx & De Boeck, 2001; Yen, 1984; Zenisky, Hambleton, & Sireci, 2001). LID typically leads to an

overestimation of (1) item discrimination parameters, (2) the amount of information provided by the test, and (3) estimates of score accuracy.

The overly optimistic estimation of score precision is a particularly serious problem in CAT when the stopping rule is based on the standard error of the scores. That is, a CAT may be programmed to stop when an individual's trait estimate is measured with a desired degree of precision or is "far enough away" from a pass-fail criterion, taking estimation uncertainty into account. The CAT will be terminated prematurely when test information is overestimated and the standard error of the score is underestimated (Fennessy, 1995); individuals will be measured and decisions will be made with less precision than is assumed.

Accommodations. A number of approaches have been suggested to accommodate or effectively eliminate the impact of LID in adaptive settings. In the development stages of a CAT, researchers can use pre-calibration response data to identify problematic items and remove them from the bank so they are excluded from operational administrations (e.g., Bjorner, Chang, Thissen, & Reeve, 2007). Alternately, dependent items can remain in the bank and constraints can be placed on the CAT algorithm to maintain more consistency in content across administrations and prevent LID items from being administered together (e.g., Stocking & Swanson, 1993; van der Linden & Reese, 1998). Or, the CAT can be based on an IRT model that allows for LID among subsets of items (e.g., Wainer, Bradlow, & Du, 2000; Wainer & Kiely, 1987). However, the success of these techniques in terms of controlling LID is dependent on the ability to determine which items exhibit it.

Detecting Local Item Dependence

Exploratory approaches based on pairwise statistics have been proposed to detect LID. Such statistics include the Q_3 statistic (Yen, 1984; 1993) and the X^2 statistic (Chen & Thissen, 1997), among others. Generally speaking, LID statistics for item pairs are based on a comparison of observed performance and expected performance as predicted by the IRT model. Differences between observed and expected performance are summarized in the LID statistic, and item pairs with values above those expected under the null condition (i.e., when LII holds) can be flagged for LID.

The pairwise diagnostic tools that have been proposed to detect LID in item response data have traditionally been studied and applied in non-adaptive settings, with both dichotomous and polytomous response data (e.g., Yen, 1984, 1993; Chen & Thissen, 1997; Ip, 2001; Kim, De Ayala, Ferdous, & Nering, 2007; Levy, Mislevy, & Sinharay, 2009; Lin, Kim, & Cohen, 2006; Zenisky, Hambleton, & Sireci, 2002). However, the data resulting from an adaptive instrument differs from data from a fixed-form instrument in two notable ways (Pommerich & Ito, 2008). First, consistent with the purpose of adaptive instruments, each individual typically answers far fewer items in an adaptive setting than in a fixed setting. Put differently, not all individuals answer all items, resulting in a sparse data matrix. Second, unlike fixed-forms, each item is administered to individuals with a restricted trait range in adaptive settings. That is, only highly-skilled individuals are administered the most difficult items and vice versa.

Purpose

It is relatively unknown whether LID statistics can reasonably be applied to adaptive data[2] because they must operate on a sparse data matrix and responses to each item come from a sample with a restricted trait range. Initial investigations (Pommerich & Ito, 2008; Pommerich & Segall, 2008) have begun to consider the performance of popular LID statistics when applied to dichotomous adaptive data. However, instruments in fields with a growing interest in CAT such as psychological and health-outcomes assessment often utilize items with (multiple) ordered-response categories (e.g., Chang & Reeve, 2005; Fries, Bruce, & Cella, 2005; Reeve & Mâsse, 2004). Thus, polytomous IRT models are likely more relevant than dichotomous IRT models in these contexts. The current investigation extends the LID detection literature to examine properties of popular LID statistics when applied specifically to polytomous adaptive data.

Evidence suggesting the LID statistics function as expected with polytomous adaptive data has critical applications for practitioners in education and other social science fields. First, the LID statistics could be applied to reliably gauge the effectiveness of strategies used to prevent LID in adaptive administrations. For example, one could determine if LID is no longer present in adaptive data once constraints have been placed on the CAT algorithm preventing LID items from being administered together. Second, they could be used to detect LID among items that were not necessarily expected to

---

[2] Note that the phrase "adaptive data" is used throughout the document to indicate response data collected in adaptive settings/from CAT administrations. Technically, the data are realizations of the response process and are themselves not adaptive. However, this terminology was used by Pommerich and Ito (2008) to refer to data from a CAT administration, and will be used throughout the current study as well for the sake of consistency.

exhibit LID, or did not exhibit LID in fixed-form settings. These items may not share obvious content or contextual similarities but may yield dependent responses due to the mode of administration, order in which they are presented, or other factors. Third, the reliable detection of large and unavoidable (for substantive reasons) LID could indicate the need to based the CAT on an IRT model that incorporates such dependencies.

Organization

This dissertation follows a five-chapter structure. The preceding, Chapter 1, introduced the context and purpose of the research. In Chapter 2, an overview of the literature related to the development of an item bank and mechanics of CAT are first presented. Then, the LID literature is reviewed with regards to its definition, causes, impact on parameter estimates, detection methods, and accommodation approaches. The chapter concludes with a detailed review of the literature focused on the properties of popular pairwise statistics used to detect LID. Chapter 3 focuses on the methods of the current research. The objectives of the inquiry are reiterated and research questions are defined. Then, the simulation design, factors, and evaluation criteria are described. In Chapter 4, results are presented and compared with those observed in previous research. Lastly, Chapter 5 includes a summary of the key findings, theoretical and practical implications of the results, limitations of the current research, and suggestions for future directions.

Chapter 2: Literature Review

To provide context for the current study, this chapter begins with an introduction to item types and classification schemes. An overview of the steps in developing a CAT and description of how a CAT operates are then offered. Each of these sections includes an illustrative example based on the PROMIS Fatigue item bank and CAT. The chapter continues with an overview of the LID literature with regards to its definition, causes, impact on parameter estimates, and detection methods, along with approaches for modeling LID. A detailed review of the literature focused on the properties of pairwise statistics used to detect LID is then presented. The current study draws heavily upon this particular line of research and extends it to examine the properties of common LID measures when applied to polytomous adaptive data.

Item Classification Schemes

The following sections discuss several dimensions that can be used to define classification schemes for types of items found on surveys for educational and psychological measurement; dimensions include (1) the nature of information they solicit, (2) the response format they utilize, and (3) the type of data they produce. Implications of these classification dimensions for the response process, scoring, and / or analysis are addressed. Finally, an illustrative example based on an item from the PROMIS Fatigue Bank is provided.

Nature of information solicited. The survey literature makes a distinction between factual and attitudinal questions, or questions designed to gather factual or behavioral data versus questions designed to measure subjective states (e.g., Converse & Presser, 1986; Fowler, 1995; Tourangeau, Rips, & Rasinski, 2000). As described by

Fowler (1995), factual questions gather information on facts and events that, in theory, can be objectively verified. They collect temporal data about dates and durations, count and describe behaviors, and test knowledge of facts and information. Attitudinal questions, on the other hand, gather information about subjective states including individuals' opinions, perceptions, feelings, and judgments. These questions have no "right" or "wrong" answers, nor is there an objective standard against which responses can be evaluated.

The distinction between items measuring facts and items measuring subjective traits is important because the way in which subjective information is stored in and retrieved from the respondent's memory differs from that of factual information (Tourangeau, 1984, 1987; Tourangeau & Rasinski, 1988). As discussed by Tourangeau, Rips, and Rasinski (2000, Chapter 6), respondents have stored in their memory a determinate set of facts that are relevant to consult in answering factual items. In contrast, attitudinal items generally do not directly reference a well-defined set of facts. Attitudes are not viewed as pre-existing evaluations that remain stable, but instead as a collective memory structure containing vague impressions, general values, and relevant feelings and beliefs, only a subset of which are retrieved when prompted by an item. In other words, there is a dynamic component to the retrieval of information when responding to items measuring subjective traits that are heavily dependent on the context within which the information is requested (Tourangeau & Rasinski, 1988).

Response format. Items can also be classified by their response format (Fowler, 1995; Masters & Evans, 1986; Tourangeau, Rips, & Rasinski, 2000). An item consists of a minimum of two key components: a stem and response options. The item stem solicits

information and the response options specify the form the answer should take (Fowler, 1995). Item stems can be in interrogative form or declarative form (Tourangeau, Rips, & Rasinski, 2000). Interrogative stems are phrased in the form of a question and request information from the respondent whereas declarative stems are informing sentences or assertions to which the respondent is asked to react. Items may also provide a context, or "super-stem", that applies to a group of items, and a time frame, or reference period (Fries et al., 2005; Tourangeau, Rips, & Rasinski, 2000).

In terms of response format, items can be open-ended or closed-ended. Close-ended / fixed-choice / selected-response items provide a pre-established list of response options from which individuals are instructed to select a response. In contrast, for open-ended / free-response / constructed response items, individuals are allowed to generate their own responses in a format of their choice (usually within certain parameters).

Analysts can subsequently code narrative or free responses to open-ended questions and assign them to numerical categories (Fowler, 1995) so data can be used in quantitative analyses. Masters and Evans (1986) describe several such instances. For example, an examinee may supply an essay response to an open-ended item and a rater may score his performance on the task according to various criteria. Or, on a math or science assessment, examinees may be asked to complete a multi-step problem and credit is awarded based on the number of steps correctly completed.

The distinction between open-ended and close-ended items is important because close-ended items result in more readily-analyzable answers (Fowler, 1995). Close-ended items are more specific than open-ended items because the response options are pre-determined leading to a more consistent frame of reference for interpretation across

respondents (Converse & Presser, 1986). Furthermore, when respondents answer close-ended questions, their answers are already in numerical format for processing or can be easily converted from letter codes to numeric codes (Fowler, 1995). For open-ended items, responses must be coded before they can be used in analyses, introducing another potential source of error and variability. Coding rules must be standardized and implemented consistently across coders or raters so that the data are not subject to uncontrolled rater effects, such as halo or leniency effects (Yen, 1993).

Type of data produced. Mellenbergh (1995) classified items by the type of data they produce. Score variables for items can be continuous with an infinite number of values; this is approximately the case in practice, for example, when respondents are asked to mark a particular position on a line or when response time is measured. Alternatively, score variables can also be discrete, which means that the item contains either a set of pre-specified response categories or distinct scores are provided by raters for open-ended responses.

Discrete items can yield dichotomous data with two response scores such as "correct" or " incorrect," "pass" or "fail," "true" or "false," "yes" or "no," and "agree" or "disagree." They can also yield polytomous data when item responses yield more than two response scores. A score variable that arises from ordered item response categories is measured on an ordinal scale. Likert scales are prototypical examples of ordered polytomous response options, where the stem of the item is a declarative sentence and the accompanying response options indicate different levels of agreement or endorsement of the statement (DeVellis, 2003). Other popular examples include sets of ordered response options reflecting frequency, ranging, for example, from "not at all" to "very often." A

score variable that arises from unordered item response categories (e.g., respondents are asked about qualitatively different behaviors that they would engage in if faced with a particular situation) is measured on a nominal scale.

The type of response data produced has implications for the selection of an appropriate IRT model. Given dichotomous response data, dichotomous IRT models such as the 1-, 2-, and 3-parameter logistic models are appropriate (Embretson & Reise, 2000). Accordingly, polytomous models should be applied to ordered polytomous response data. In addition to the GRM (Samejima, 1969) mentioned earlier, popular polytomous IRT models for ordinal response data include the modified graded response model (M-GRM; Muraki, 1990), partial credit model (PCM; Masters, 1982), generalized partial credit model (G-PCM; Muraki, 1992), the rating scale model (RSM; Andrich, 1978), and their multidimensional extensions (e.g., Bock, Gibbons, & Muraki, 1988; McKinley & Reckase, 1982; Reckase, 1997). Some polytomous IRT models, such as the RSM, require that all items on the instrument have the same number of response categories, while other models, such as the GRM, can accommodate items with different response formats. Bock's (1972) nominal response model (NRM) is specifically appropriate when response variables are measured on a nominal scale.

Illustrative example. A sample item from the PROMIS Fatigue Bank, HI7, is discussed with respect to the preceding classification dimensions. This item begins with a super-stem reading, "In the past 7 days," which establishes the reference period for the item. The item stem then reads, "I feel fatigued." This stem is presented in declarative form, in that it offers an informing sentence or assertion to which the respondent is asked to react. Because the item asks the respondents to describe how they feel, it can be

considered a question designed to measure a subjective state, as opposed to a factual state. The item is then accompanied by five response options: not at all, a little bit, somewhat, quite a bit, and very much. Thus, the item is considered a close-ended item. Lastly, with five ordered response options, the item yields ordinal polytomous response data.

CAT Development Process

The preceding section described the types of items that may be included in an item bank. This section now focuses on how the item bank itself is developed and utilized as the platform for a CAT. Table 1 outlines the steps in the construction of an item bank and development of an adaptive instrument, adapted from Bjorner et al. (2007). It provides an overview of the key components at each stage and select resources which describe the process in greater detail. An overview of each step follows, using information about the actual development of the PROMIS Fatigue item bank and CAT as an illustrative example.

It is important to note that pairwise LID statistics have traditionally been applied in Step 4 of the CAT development process outlined in Table 1, during item calibration and tests of model fit. At this point, because the CAT is under development, the calibration data are not adaptive. Instead, the current research focuses primarily on the application of the Q_3 and X^2 in Step 7 of the CAT development process, during item bank maintenance. In this stage, because the CAT is operational, the available data are adaptive; unlike the pre-calibration data, responses to a given item pair come from a sample with a restricted trait range and individuals see different subsets of items from the bank in different orders.

Table 1. Steps in the CAT development process

Step	Description	Select Resources
1. Construct definition	• Clearly define the construct of interest by specifying all its relevant aspects or sub-domains and the domains that are not part of the construct • Have experts make conceptual and qualitative decisions about the framework based on theory • Use pre-existing empirical data to confirm or deny the viability of the framework	Bjorner et al. (2007) DeVellis (2003) Embretson & Reise (2000) Fries, Bruce, & Cella (2005)
2. Item development	• Construct an initial pool of newly developed items and/or items from established measures that reflect the construct of interest • Ensure items meet test specifications, adhere to the basic rules of item writing, and reflect a spread of difficulty along the latent continuum • Subject items in initial pool to an expert item review and pre-test items in cognitive interviews and/or focus groups to identify problematic items • Revise or remove problematic items from the pool	Bjorner et al. (2007) Converse & Presser (1986) DeVellis (2003) Flaugher (1990) Fowler (1995) Fries et al. (2005) Touragenau, Rips, & Rasinski (2000) Willis (2005)
3. Data collection	• Collect data from a large sample of respondents representing the target population • Over-sample individuals at extreme trait ranges if necessary so sufficient data are available to estimate all category thresholds for all items with reasonable precision • Collect pre-calibration data in the same mode as operational administrations if possible to support comparability and stability of estimates	Bjorner et al. (2007) Embretson & Reise (2000) Thissen et al. (2007) Wise (1997)

4. Item calibration and test of model fit	• Fit a non-parametric IRT model to the pre-calibration data and examine item characteristic curves to identify poor items and/or response options • Fit parametric IRT model(s) and examine fit statistics to identify the best-fitting model • Test model assumptions, including dimensionality and local independence • Identify any poorly fitting items and/or items that exhibit differential item functioning • Remove problematic items from the pool and/or use a more general IRT model • Calibrate item parameters according to selected IRT model	Ackernan (1994) Embretson & Reise (2000) Jang & Roussos (2007) Lee (2007) Mellenbergh (1995) Steinberg & Thissen (2006) Stout (2002)
5. Norming, benchmarking, and interpretation guidelines	• Define the metric or IRT score relative to the target population • Create scores and score reports that are easy for users to interpret • Provide benchmark scores to help users determine where they stand compared to various population subgroups • Produce cross-calibration tables to show how scores on the CAT compare to those on traditional, non-adaptive measures	Bjorner et al. (2007) Thissen et al. (2007)
6. CAT design and pretesting	• Conduct simulations of CAT administrations to evaluate the impact of various parameters on test length, precision, and validity • Establish parameters for operational administrations • Build computer program and operating system for administering, scoring, and reporting on the CAT • Pilot the tool to ensure technical specifications are met and programming errors are identified and corrected • Solicit and incorporate feedback from users on usability and acceptability	Gershon et al. (2010) Jansky & Huang (2009)
7. Final item bank and bank maintenance	• Investigate the longitudinal performance of items to address any parameter drift • Integrate new items into the bank as they are developed using equating methods • Maintain the integrity and security of the item bank by retiring over-exposed items • Discard existing items that become irrelevant	Masters & Evans (1986) Wise (1997)

Construct definition. The PROMIS network first developed a construct map that portrayed the structure of each target domain and its conceptual framework (Cella et al., 2007). As a part of this process, three independent literature reviews were conducted by experts at three of the funded PROMIS entities. In the literature review, the explicit or implicit frameworks that formed the basis for existing outcome assessment questionnaires were considered, along with well-accepted models of health, including the model adopted by the World Health Organization (Fries et al., 2005). PROMIS also completed statistical analyses of available data from more than 50,000 respondents to investigate the dimensionality of health status assessment (Reeve et al., 2007). Using both the theoretically- and empirically-driven models, PROMIS reached consensus on a framework with three broad health domains: physical health, mental health, and social health.

Within these, sufficiently unidimensional sub-domains were defined, with fatigue falling under the physical health domain. The dimensionality of the Fatigue sub-domain was explored using pre-existing data from PROMIS's Statistical Coordinating Center (Lai & Chen, 2006). Seventy-two fatigue-related items from the CORE Cancer Fatigue instrument and 13 items from the UBC Fatigue instrument were considered, using data from 555 cancer treatment patients and 1,225 chronic hepatitis C patients, respectively. Descriptive statistics including psychometric properties, like Cronbach's alpha, item-total correlations, exploratory factor analyses, and one-factor confirmatory factor analyses yielded support for sufficient unidimensionality.

For the fatigue sub-domain or construct, PROMIS (www.nihpromis.org/Web%20Pages/Domain%20Definitions.aspx) offers the following

definition: “The PROMIS Fatigue item bank assesses fatigue from mild subjective feelings of tiredness to an overwhelming, debilitating, and sustained sense of exhaustion that is likely to decrease one’s ability to carry out daily activities, including the ability to work effectively and to function at one’s usual level in family or social roles. Fatigue is divided conceptually into the experience of fatigue (such as its frequency, duration, and intensity), and the impact of fatigue upon physical, mental, and social activities.”

Item development. PROMIS began the process of item development by cataloguing items from well-established instruments in health-outcomes domains (DeWalt, Rothrock, Yount, & Stone, 2007). Investigators conducted both electronic database searches and manual file searches to locate such instruments. Candidate items were pulled from these instruments into an initial pool of domain-relevant items. At this point, items were only excluded from the pool if their content was not aligned with the domain definition, and no judgments were made regarding the reference population or item quality.

Once an initial pool of items was established, PROMIS began the item review process. Content experts “binned and winnowed” the items by placing items with a common content in “bins” (i.e., sets) and “winnowing out” (i.e., removing) items that were redundant or inferior to other items. Items were removed if the content was inconsistent with the domain definition, the item was semantically redundant with another item, the item content was too narrow or disease-specific to be universally applicable, or the item was confusing. Thus, the goal was to identify the best potential items based on their qualitative characteristics and identify sets of items that adequately represent the facets of the latent trait.

Following the expert item review and revision, PROMIS conducted a series of focus groups and cognitive interviews to further pre-test the items (DeWalt et al., 2007). The focus groups were used to gather patient input regarding conceptual gaps in the domain definition, leading to the development of new items when existing items did not provide adequate coverage. The cognitive interviews helped ensure that items were understood and interpreted as intended, particularly by respondents with low levels of literacy. At the end of this process, the fatigue item pool consisted of 58 items measuring fatigue impact, 54 items measuring fatigue experience, and 17 "legacy" items (items from widely-used fixed length measures) related to fatigue.

Data collection. As described on the PROMIS website (www.nihpromis.org/Web%20Pages/PSYCHO%20Metricians.aspx), Wave I data were collected in 2006 and 2007 from the U.S. general population and multiple disease populations primarily by the polling firm YouGovPolimetrix (www.polimetrix.com). Subjects were selected from a panel of over one million respondents maintained by YouGovPolimetrix. Individuals in the panel regularly participate in YouGovPolimetrix Internet surveys and have provided YouGovPolimetrix with their names, physical addresses, email addresses, and other information. Subjects were recruited by a variety of methods, including e-random digit dialing, invitations via web newsletters, and Internet poll-based recruitment. A small number of subjects were also recruited from primary research sites associated with PROMIS network sites.

Subjects were selected to meet specified targets in terms of gender (50% female), age (20% in each of 5 age groups: 18-29, 30-44, 45-59, 60-74, 75+), race/ethnicity (10% black and Hispanic), and education (10% less than high school graduate). Persons who

self-reported as currently being diagnosed with a given condition were included in the clinical sample associated with that condition. The selection of subjects was made on this basis so that responses could be collected on the candidate items from the targeted PROMIS domains for both (1) the general U.S. population and (2) specific disease subpopulations. The general population subsample was primarily used to establish U.S. population norms. Overall, the PROMIS Wave I sample included 21,133 participants. The item calibration sample for the Fatigue bank specifically included 14,931 cases in total.

Subjects were recruited by YouGovPolimetrix online via e-random digit dialing, invitations via web newsletters, and Internet poll-based recruitment or on-site at PROMIS research sites. To insure the comparability of results, pre-calibration data were collected on an Internet survey platform so that parameters can safely be considered valid for Internet or personal computer-based applications with screen presentations of individual items. YouGovPolimetrix sample data were collected using their website on a secure server, while data from the research sites were collected using the PROMIS Assessment System.

Given the large number of candidate items from the targeted PROMIS domains, it was not possible for each participant to respond to every item in the pool. Instead, two data collection designs – full bank and block administration – were used. In the full bank administration, individuals were administered full banks of items for the relevant PROMIS subdomains. In the block administration, individuals were administered a subset, or block, of items from each domain. These administration approaches limited the number of items administered to any respondent to roughly 150 and the administration

time to less than 40 minutes. In addition to the candidate items, participants were asked to answer about 20 auxiliary items consisting of global health rating items and sociodemographic variables such as age, income, gender, and race/ethnicity. They were also asked a series of questions about the presence and degree of limitations related to chronic medical conditions.

Item calibration and tests of model fit. Once data from the calibration sample had been collected, they were fit to an IRT model. For the PROMIS Fatigue bank, the full bank sample was used to determine the dimensionality of the fatigue items; both full bank and block data were used for item parameter estimation (Lai, 2007). The full bank data included 803 cases in total while the item calibration sample included 14,931 cases in total with the number of cases per item ranging from 2,209 to 2,893. The pool of 58 fatigue impact items and 54 fatigue experience items were initially analyzed separately and then together to determine if they could be combined.

The scalability of the items was investigated using classical test theory and non-parametric IRT techniques, including Spearman's rho, item-total score correlations, Cronbach's alpha, and Mokken scaling (Lai, 2007). As a result, two of the fatigue impact items were removed because they failed to meet most of the inclusion criteria. After reviewing measures of item fit, two fatigue experience items were removed due to poor fit as well. Local dependency was also examined at this step in the CAT development process. Although researchers in the PROMIS network did not use the Q_3 and / or X^2 specifically, they used a similar kind of pairwise statistic based on the polychoric correlation coefficients resulting from the observed and expected tables of responses across two items (Bjorner, Smith, Stone, & Sun, 2007). Residual correlations were

calculated as the difference between the expected and observed correlation coefficient; five fatigue impact and 11 fatigue experience items were set aside because of residual correlations greater than 0.20. Unidimensionality of the remaining items was investigated via confirmatory factor analysis for the fatigue impact pool, fatigue experience pool, and combined pool; in all three cases, multiple goodness-of-fit statistics supported sufficient unidimensionality. As a result, one fatigue item bank was developed instead of two separate banks.

Item parameters for the combined item pool were then estimated using the GRM (Lai, 2007). In initial analyses with pre-existing health-outcomes datasets, the PROMIS network evaluated several alternative polytomous IRT models, including the GRM, PCM, and RSM (Reeve et al., 2007). Based on these analyses and discussions with measurement experts about the strengths and limitations of each model, the development team decided to focus on the GRM (J-S. Lai, personal communication, July 7th, 2010; B. Reeve, personal communication, July 9th, 2010). Primary reasons for selecting the GRM included its flexibility and interpretability (Reeve et al., 2007).

With the number of responses per item greater than 2,000, Reise and Yu's (1990) minimum sample size recommendation of 500 for adequate item parameter calibration under the GRM was well-met. After fitting the data to the GRM, seven additional items were removed due to poor item fit. Differential item functioning (DIF) with regards to gender and age was investigated using ordinal logistic regression analyses. Four items were found to exhibit DIF and were not calibrated as a result. The remaining 98 items were calibrated under the GRM (note that three items were later removed for content reasons, resulting in a final Wave 1 bank of 95 items).

Norming, benchmarks, and interpretation guidelines. A subset of the PROMIS general population sample (n = 5,239) was selected to match the marginal distributions of race/ethnicity and education from the 2000 U.S. Census (Lai, 2007). This subsample was used to center and norm the item parameters so that they would align with performance characteristics of the general U.S. adult population. In other words, this subsample of respondents formed the norming sample for the PROMIS item banks. On the latent variable scale, the mean fatigue level was set at zero and the standard deviation to one. Under this parameterization, slope parameter estimates ranged from 1.17 to 4.77 and category boundary parameter estimates from -2.48 to 3.67 across the fatigue items.

The metric was then transformed so that the mean was set to 50 and the standard deviation at 10 (Bjorner et al., 2007; Lai, 2007). This was done in part to help make scores easier to interpret for instrument users, such as physicians and patients themselves. PROMIS also calculated "benchmark" scores to show the average level of fatigue for various population subgroups based on gender and age. These benchmarks can be used to inform patients about how their level of fatigue compares not only to that of the general population, but to someone of their gender, and of their age group.

CAT design and pretesting. One of the PROMIS objectives was to create a web-based software system, now known as the Assessment Center, which allowed researchers to create study-specific websites that could administer PROMIS CATs (Gershon, Rothrock, Hanrahan, Jansky, Harniss, & Riley, 2010). To develop the software, the design team first met with researchers in the PROMIS network to gather information about the desired requirements and functional specifications of the system, including features, performances, and interface considerations (Gershon et al., 2010; Jansky &

Huang, 2009). A prototype was developed based on these specifications and tested by experts and PROMIS stakeholders to ensure the requirements were met, features worked properly, and that errors encountered in the system were identified. The team then finalized the programming and conducted quality assurance to identify and resolve the issues noted by users.

The software was also tested by end-users, including those with a low level of computer literacy and those with disabilities such as visual, motor, or reading impairments (Gershon et al., 2010; Jansky & Huang, 2009). One of the testing activities involved participant observation; researchers used cameras to link participants' facial expressions to screen shots. Another activity involved think-alouds in which the participants verbalized their thought processes while navigating the system. Lastly, the development team conducted focus groups and semi-structured interviews in which they asked participants to comment on topics including the functionality and ease of use of the system and the interface design. They were also asked about current barriers and suggestions for system modifications. These changes were prioritized based on the frequency with which they were mentioned, the degree of fit with the intended scope of the Assessment Center, and the amount of time needed for completion, and either incorporated or included in feature lists for future release (Gershon et al., 2010).

PROMIS also supported the development of programs that simulate CAT with polytomous items so researchers could evaluate the impact of various CAT parameters on test length, precision, and validity. Two such programs include FIRESTAR (Choi, 2009) and SIMPOLYCAT (Chen & Cook, 2009). Both programs support multiple polytomous IRT models, including the GRM and PCM, and allow users to apply various item

selection techniques, stopping criteria, and theta estimation techniques. Researchers in the PROMIS network have utilized simulation programs to determine appropriate CAT parameters. For example, Gershon, Choi, Lai, Wee, Yoo, and Hambleton (2009) compared the average test length of the Fatigue CAT under different polytomous IRT models. They found that, on average, an additional one to two items were needed to meet the precision requirements when a version of the GRM that restricted the slope parameter across items was used as opposed to the unrestricted GRM with unique slope parameters for each item. As a result, the default parameters of the PROMIS Fatigue CAT are based on the unrestricted GRM.

Final item bank and bank maintenance. Because PROMIS is still in the initial phases of the project, long-term maintenance of the final item bank is not yet a primary focus. As described on the PROMIS website, in Wave I of the project item banks were developed. Currently, in Wave II, the focus is on validating the existing item banks by comparing them with "gold standard" instruments in the health-outcomes field, evaluating the responsiveness of instruments under conditions of known change in an underlying chronic disease, and investigating mode-of-administration effects. As part of the next phase of PROMIS advancement, the network plans to develop new items and improve PROMIS tools to improve outcomes measurement. To achieve these goals, item bank maintenance will be necessary. For example, after the item banks have been operational for a period of time, the longitudinal properties of items may be examined for evidence of parameter drift. Or, equating methods may be used so that new items can be included in the bank. It is less likely, however, that PROMIS will retire over-exposed

items from the bank; the content of items need not be secure because the concept of "cheating" is not applicable in health-outcomes settings (Bjorner et al., 2007).

How CAT Operates

An instrument provides the most precise measurement of an individual when the difficulty of the instrument matches the individual's trait level (Hambleton, Swaminathan, & Rogers, 1991). Based on this premise, CAT uses computer technology to produce an instrument that is uniquely tailored or adapted to each individual completing it. The power of computers to store test information and efficiently produce, administer, and score tests, makes such adaptive testing feasible.

A CAT is built on a pool of items, or item bank, with known statistical properties. IRT is particularly well-suited for CAT because it yields parameter estimates for items that are independent of the individuals completing the instrument and trait estimates for individuals that are independent of the items they answered (Embretson & Reise, 2000). Importantly, this is only true when the model fits perfectly to data from a homogenous population and is only approximately true once model fit becomes imperfect or populations become heterogeneous.

Thus, using IRT, the statistical properties of the items in the bank can be pre-calibrated using a representative set of individuals and administered to new individuals. From this pool, suitable items can be selected and administered to respondents following certain criteria. These criteria establish rules for (1) selecting the first item, (2) selecting subsequent items, and (3) stopping the CAT.

Selecting the first item. In CAT, the sequence of items administered to an individual is dependent on the responses he or she provided to prior items. However, it is

not possible to select the first item based on previous responses because no such responses are available. In this case, an initial trait estimate must be supplied so that the algorithm can select an appropriate item. This provisional trait estimate can be set at the mean of the population trait distribution or randomly selected from a range of plausible trait estimates. If more information is known about the respondent, such as demographic characteristics or scores on another assessment, this auxiliary information can be used to supply a more reasonable guess based on a more narrowly defined population subgroup (Thissen & Mislevy, 1990).

In certain settings, particularly in educational testing, it may be desirable to present an initial item with a difficulty below the average proficiency level of the population or subgroup so that the examinee is more likely to get the answer right than wrong. Starting with such an item is not optimal from a statistical perspective but from a psychological perspective; ensuring the examinee has a successful first experience may help reduce test anxiety (Gershon, 1989).

The selection of the initial item may also be governed by item exposure controls. If, for example, the provisional trait estimate for every individual was set as the mean of the population distribution, the same item – specifically the item in the bank with the highest discrimination parameter and a difficulty level closest to that of the population mean – would be selected first in every administration. In the context of high-stakes educational testing, the content of this item would be exposed to the public and test security would be compromised, which is why unconstrained selection of the first item is not done in practice. To prevent this problem, exposure controls may be applied so that individuals

randomly receive one out of a (small) number of most appropriate possibilities (Embretson & Reise, 2000, p. 266; Thissen & Mislevy, 1990, p. 121).

Once the first item is administered and a response is received, the individual's initial trait estimate must be updated using the IRT model upon which the CAT is built. One commonly utilized estimator, maximum likelihood estimation (MLE), maximizes the likelihood function, or $L(\theta \mid \mathbf{x_n})$, where $\boldsymbol{x}_n$ is the set of responses obtained thus far (Thissen & Mislevy, 1990). A popular alternative to the MLE estimator is the expected a posteriori (EAP) estimate (Bock & Mislevy, 1982), which is based on numerical evaluation of the mean of the posterior distribution. Broadly speaking, if the respondent gets the first item "right" (or falls in a high response category) in CAT, the estimate of his or her trait level will increase, but if he or she gets the first item "wrong" (or falls in a low response category), his or her estimated trait level will decrease. Additionally, as more items are administered, the precision with which the trait level is estimated will increase, and the confidence interval surrounding the estimate will become narrower.

Selecting the subsequent items. After the first item is administered and scored, the next item must be selected for administration based on the updated trait estimate. The CAT algorithm is instructed to select an unused item from the bank that best meets specified criteria. A commonly implemented strategy is to select the item that provides the maximum information at the currently-estimated trait level. Known as the maximum item information (MII) approach (Lord, 1980), the item that maximizes the item information function at the estimated trait level for respondent *i*, $\hat{\theta}_i$, is chosen. Bayesian approaches to item selection are also available in which the item that maximizes the posterior precision / minimizes the posterior variance of the trait estimate is selected

(Owen, 1975; Thissen & Mislevy, 2000; van der Linden & Pashley, 2000). For example, maximum posterior-weighted information (MPWI) is a Bayesian approach to item selection (Penfield, 2006). MPWI is based on the expected information for the individual which incorporates his or her posterior distribution of θ.

In addition to selecting an item that is statistically optimal, other practical constraints such as item exposure and content balancing may govern the selection of items in CAT (e.g., Chang & Ying, 1999; Stocking & Swanson, 1993; van der Linden, 2000; van der Linden & Chang, 2003). Without constraints, the selection algorithm is drawn to the highest quality, or most discriminating, items in the bank. This leads to consistent use of just a small subset of items from the bank compromising test security. Also, without constraints the algorithm may select collections of items across administrations that reflect too much variation in content. In a verbal reasoning test, for example, one examinee could receive a test consisting entirely of reading passages while another examinee could receive a test consisting of only sentence completions. Thus, constraints may be placed on the selection algorithm so that the statistically optimal item that best meets additional criteria is selected from the bank.

After the second item is administered, the individual's trait estimate must again be updated, along with the confidence interval surrounding the score. At this point, the individual's current status is compared against the stopping rules for the CAT to determine if they have been met or if more items need to be selected and administered. Stopping rules specify the necessary requirements for terminating the CAT. Frequently implemented stopping rules require that a set measurement precision has been attained, a minimum or maximum number of items has been administered, and / or a time limit has

been reached (Thissen & Mislevy, 1990). With regards to precision, this may mean the standard error of the estimate is at or below a predetermined level or the trait estimate is "far enough away" from a pass-fail criterion, taking estimation uncertainty into account. A maximum number of items and time limits are typically set to ensure the CAT does not become too long, unnecessarily exposing items or leading to respondent burden and fatigue.

It is also important to note that stopping rules govern whether a CAT is fixed length or variable length. In a fixed-length CAT, all individuals are administered the same number of items. In variable-length CAT, individuals may be administered different numbers of items depending on how many items it takes for the stopping rules to be met. For example, if the stopping rule is based on a time limit, individuals who have not mastered the concepts or read at a slower pace will be administered fewer items than individuals who are able to respond at a faster rate. Or, if the stopping rule is based on a predetermined level of precision, trait estimates for individuals who provide highly consistent responses across a series of items will be more precise with fewer items than for individuals who offer aberrant response patterns, allowing the CAT to terminate sooner. In educational settings, examinee perceptions of fairness may lead test administrators to use fixed-length CATs (or at least require a comparably high minimum number of items), even if the trait estimates of some examinees could be estimated with adequate precision with fewer items.

Stopping the CAT. Once the stopping rules have been satisfied, the CAT is terminated. At this time, a final trait estimate is calculated, either in the same manner as the provisional trait estimates or using an alternative approach which focuses less on

expediency and more on factors such as test fairness (Thissen & Mislevy, 1990). For example, prior information that may have been incorporated in initial item selection and provisional trait estimates in a Bayesian estimation approach may not be incorporated in the final estimation process so that final estimates for individuals providing identical response sets do not differ based on other characteristics. Administrators must also consider the reporting metric so that the meaning of scores is easily understood and interpreted by respondents and instrument users. For example, instead of reporting latent variable estimates that function akin to standardized z- scores, transformed scores with a mean of 50 and a standard deviation of 10 may be presented to members of the general public (Bjorner et al., 2007).

Comparison of dichotomous and polytomous CAT. Whether the CAT is based on an IRT model for dichotomous or polytomous data is actually of limited consequence. In fact, researchers note that CAT with polytomously scored items proceeds in exactly the same way as CAT for dichotomously scored items (Masters & Evans, 1986; Weiss, 1982). The most critical distinction, however, is that polytomously scored items are more informative than dichotomously scored items, which has implications for the size of the item bank and test length (Bjorner et al., 2007; Dodd, Koch, & De Ayala, 1989, 1993; Singh, Howell, & Rhoads, 1990; Thissen et al., 2007).

Polytomously scored items accompanied by ordinal response options can, in a sense, be considered "self-adapting" (Thissen et al., 2007). That is, individuals with trait levels towards the lower end of the continuum can chose amongst the lower response options (e.g., strongly disagree and disagree), while individuals with higher trait levels can select higher response options (e.g., agree and strongly agree). Thus, polytomous

items provide information about respondents across a relatively wide range of the latent continuum. In contrast, dichotomously scored items are informative over a narrow range of the continuum and uninformative at other levels (Bjorner et al., 2007). Prototypical item information functions support this fact. Generally speaking, the information function for a dichotomously scored item is highly peaked and unimodal. The information function for a polytomous item is generally broader and flatter, relatively speaking, and may even be multimodal.

The greater amount of information associated with polytomously scored items has several implications for CAT. First, item banks consisting of polytomously scored items can be smaller (i.e., contain fewer items) than banks consisting of dichotomously scored items (Dodd, Koch, & De Ayala, 1989, 1993). To implement a CAT based on dichotomously scored items, Urry (1977) recommends that the bank contain at least 100 items (ignoring other test specifications such as exposure controls and content balancing). Banks of polytomously scored items can be smaller, and a CAT can be implemented successfully on banks containing as few as 30 or 40 items that are sufficiently spread along the latent continuum (Craig & Harvey, 2004; Dodd, Koch, & De Ayala, 1989, 1993; Singh, Howell, & Rhoads, 1990). Second, assuming the stopping rule is based on the standard error of the trait estimate, adaptive instruments containing polytomously scored items can be shorter than those containing dichotomously scored items. This is because the same level of precision can be obtained with fewer items when multiple response options are used (Bjorner et al., 2007).

Illustrative example. Adapted from Thissen and Mislevy (1990), Figure 2 depicts the logic of a computer adaptive test in the form of a flowchart.

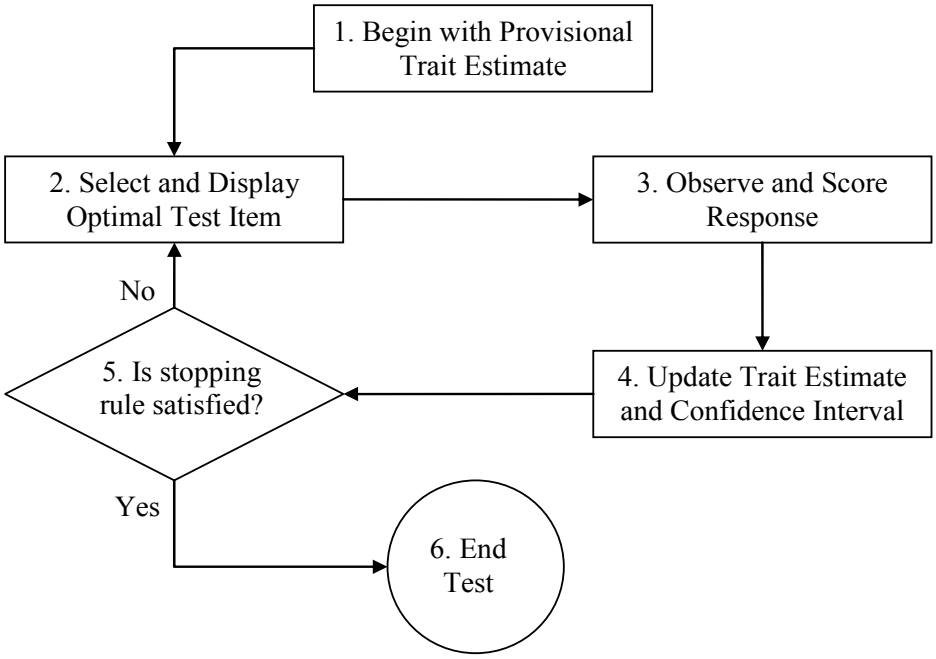
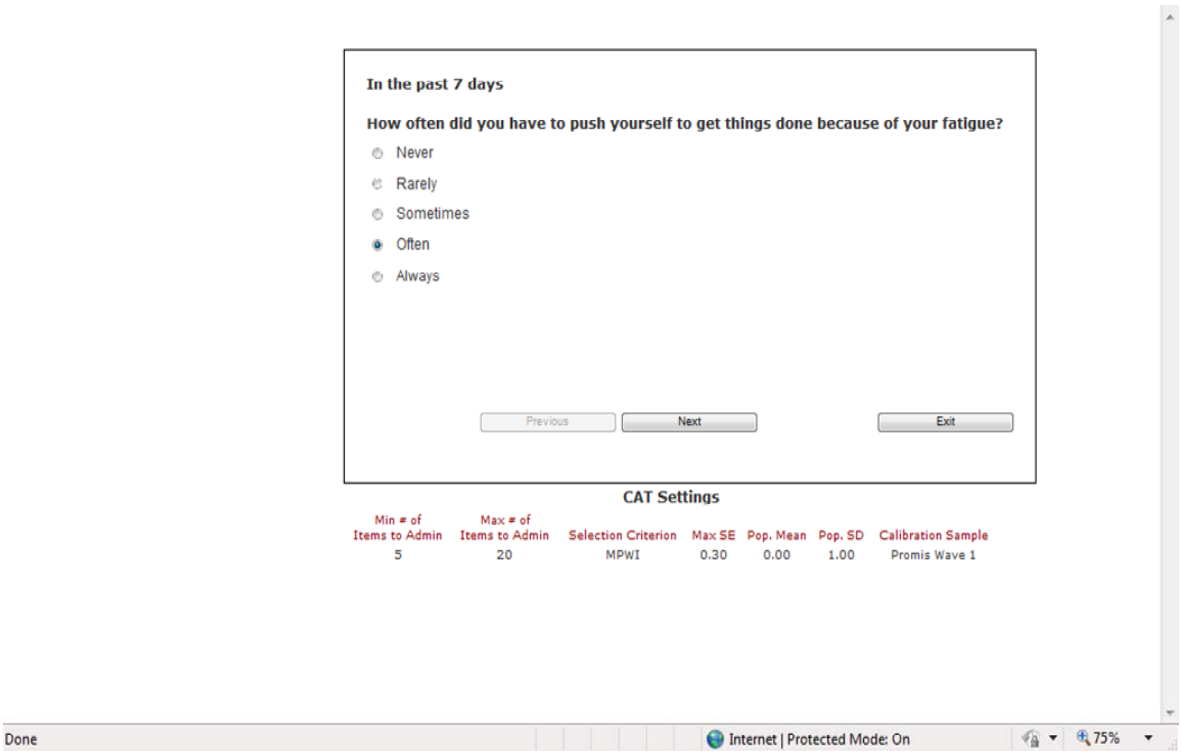


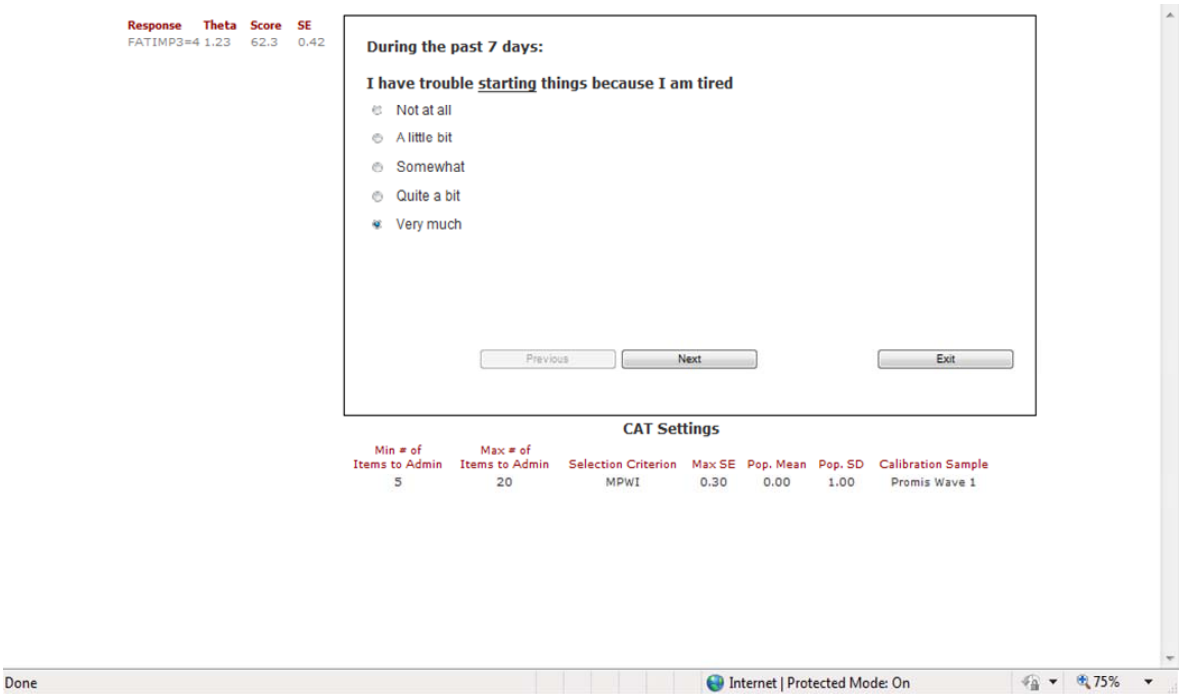
Figure 2. Flowchart describing logic of a computer adaptive test

To illustrate these steps, a sample CAT administration of the PROMIS Fatigue instrument is presented. The bank includes 95 items accompanied by five-point response scales, calibrated under the GRM. The default CAT parameters for the instrument are modeled in the CAT illustration. Specifically, the parameters specify that between five and 20 items are administered, a standard error of 0.30 or below must be achieved, and items are selected using the MPWI criterion. The hypothetical individual responding to the instrument, a 50-year-old male, belongs to the population against which the PROMIS Fatigue bank was calibrated and normed, namely the U.S. general adult population. The mean level of fatigue in the population is 0 and the standard deviation is 1. The illustrative example is carried out using the CAT demonstration and simulation features on the PROMIS Assessment Center website (www.assessmentcenter.net). Figure 3 includes the screen shots accompanying the example.

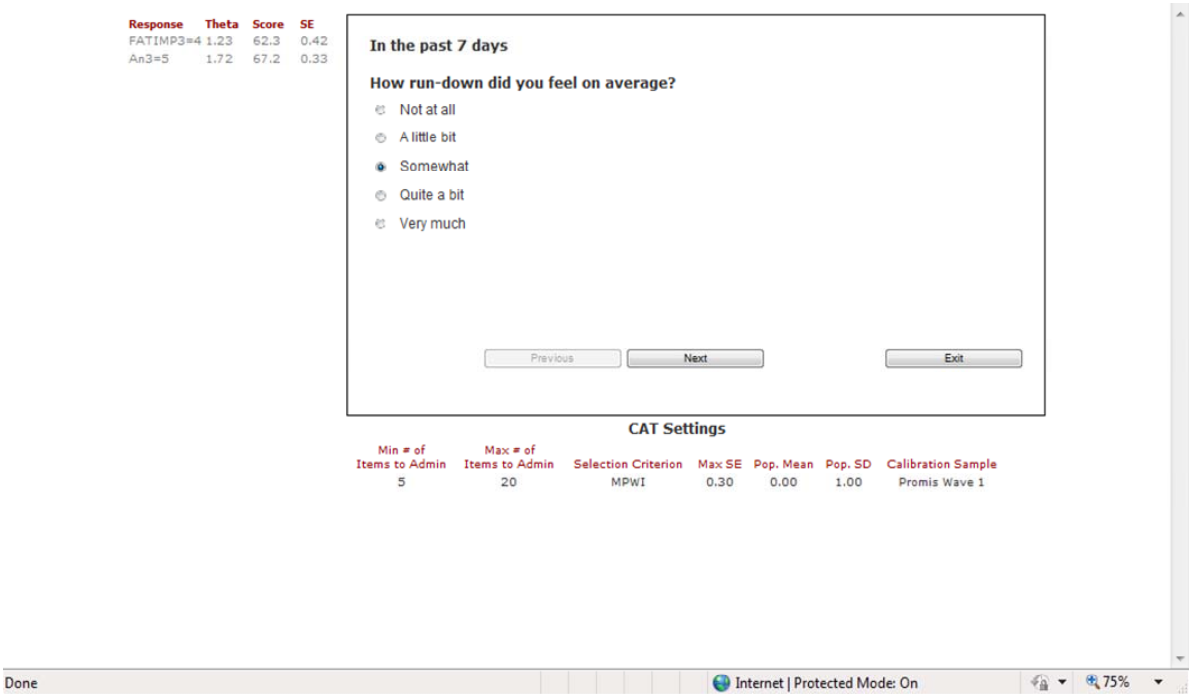
Item1


In the past 7 days
How often did you have to push yourself to get things done because of your fatigue?
Never
Rarely
Sometimes
Often
Always
Previous
Next
Exit
CAT Settings
Min # of Items to Admin
Max # of Items to Admin
Selection Criterion
Max SE
Pop. Mean
Pop. SD
Calibration Sample
5
20
MPWI
0.30
0.00
1.00
Promis Wave 1
Done
Internet | Protected Mode: On
75%

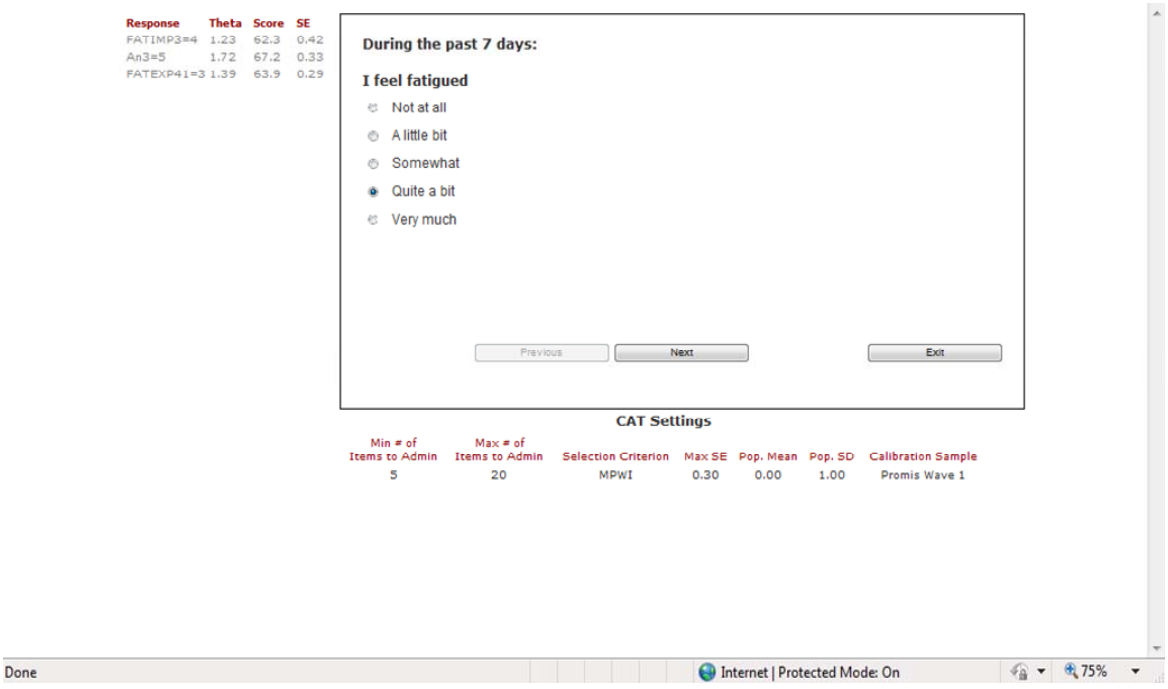
Item2


Response
Theta
Score
SE
FATIMP3=4 1.23 62.3 0.42
During the past 7 days:
I have trouble starting things because I am tired
Not at all
A little bit
Somewhat
Quite a bit
Very much
Previous
Next
Exit
CAT Settings
Min # of Items to Admin
Max # of Items to Admin
Selection Criterion
Max SE
Pop. Mean
Pop. SD
Calibration Sample
5
20
MPWI
0.30
0.00
1.00
Promis Wave 1
Done
Internet | Protected Mode: On
75%

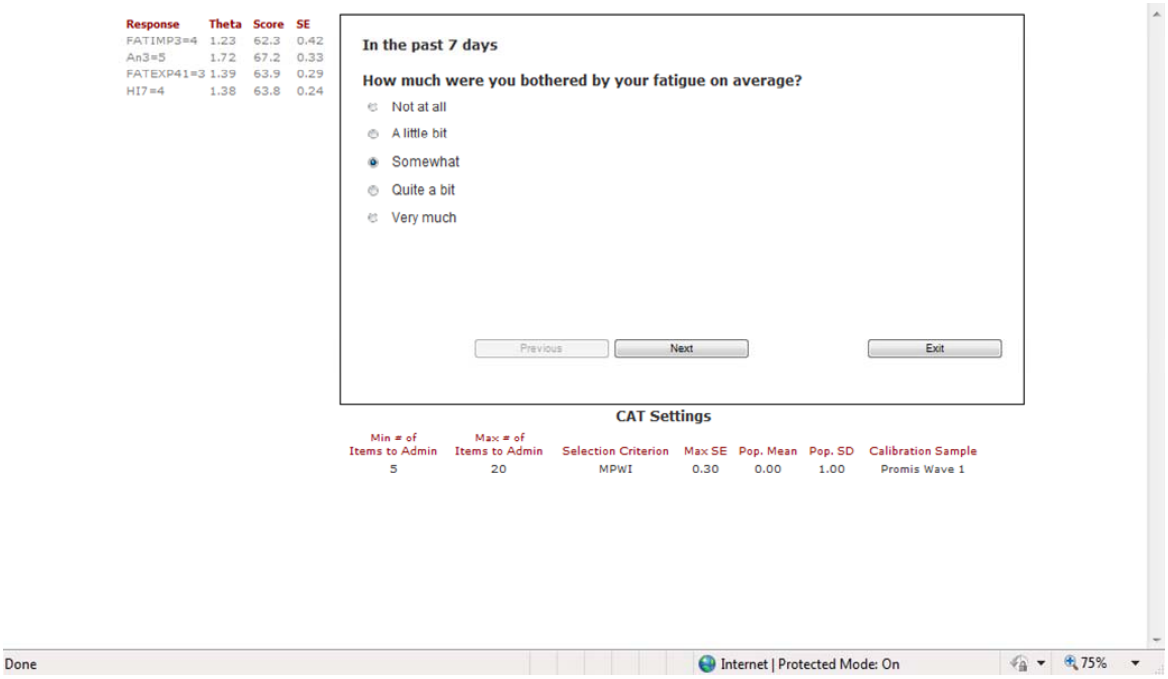
Item3


Response Theta Score SE
FATIMP3=4 1.23 62.3 0.42
An3=5 1.72 67.2 0.33
In the past 7 days
How run-down did you feel on average?
Not at all
A little bit
Somewhat
Quite a bit
Very much
Previous
Next
Exit
CAT Settings
Min # of Items to Admin
Max # of Items to Admin
Selection Criterion
Max SE
Pop. Mean
Pop. SD
Calibration Sample
5
20
MPWI
0.30
0.00
1.00
Promis Wave 1
Done
Internet | Protected Mode: On
75%

Item4


Response Theta Score SE
FATIMP3=4 1.23 62.3 0.42
An3=5 1.72 67.2 0.33
FATEXP41=3 1.39 63.9 0.29
During the past 7 days:
I feel fatigued
Not at all
A little bit
Somewhat
Quite a bit
Very much
Previous
Next
Exit
CAT Settings
Min # of Items to Admin
Max # of Items to Admin
Selection Criterion
Max SE
Pop. Mean
Pop. SD
Calibration Sample
5
20
MPWI
0.30
0.00
1.00
Promis Wave 1
Done
Internet | Protected Mode: On
75%

Item5



End

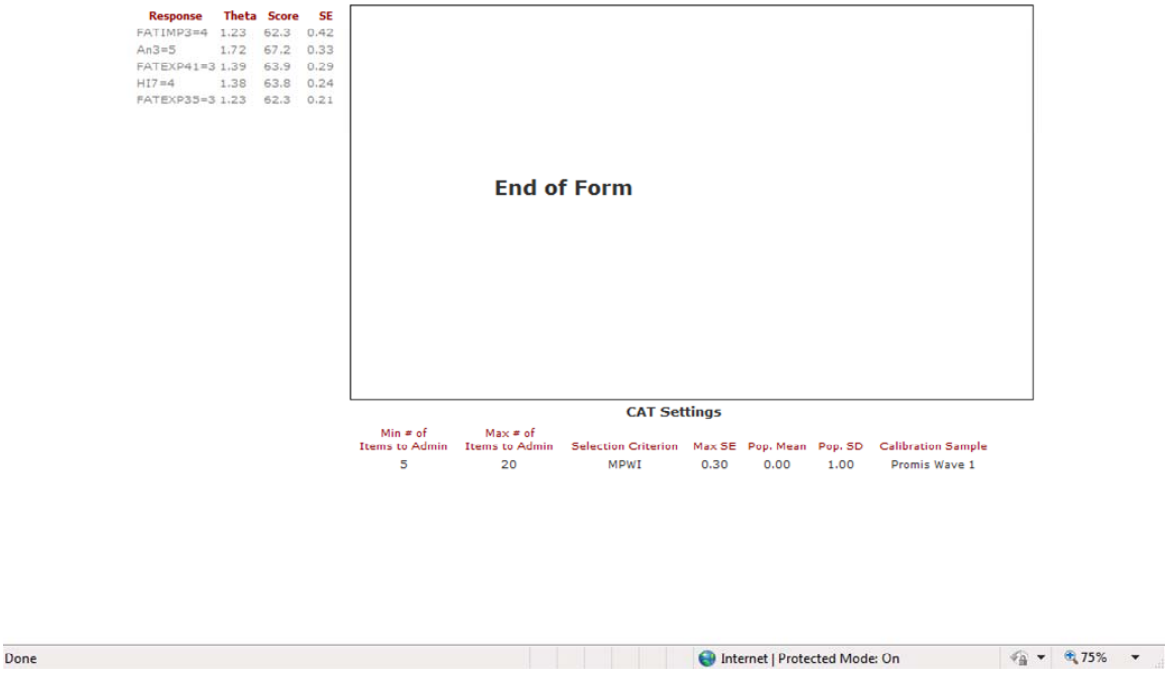


Figure 3. CAT screen shots

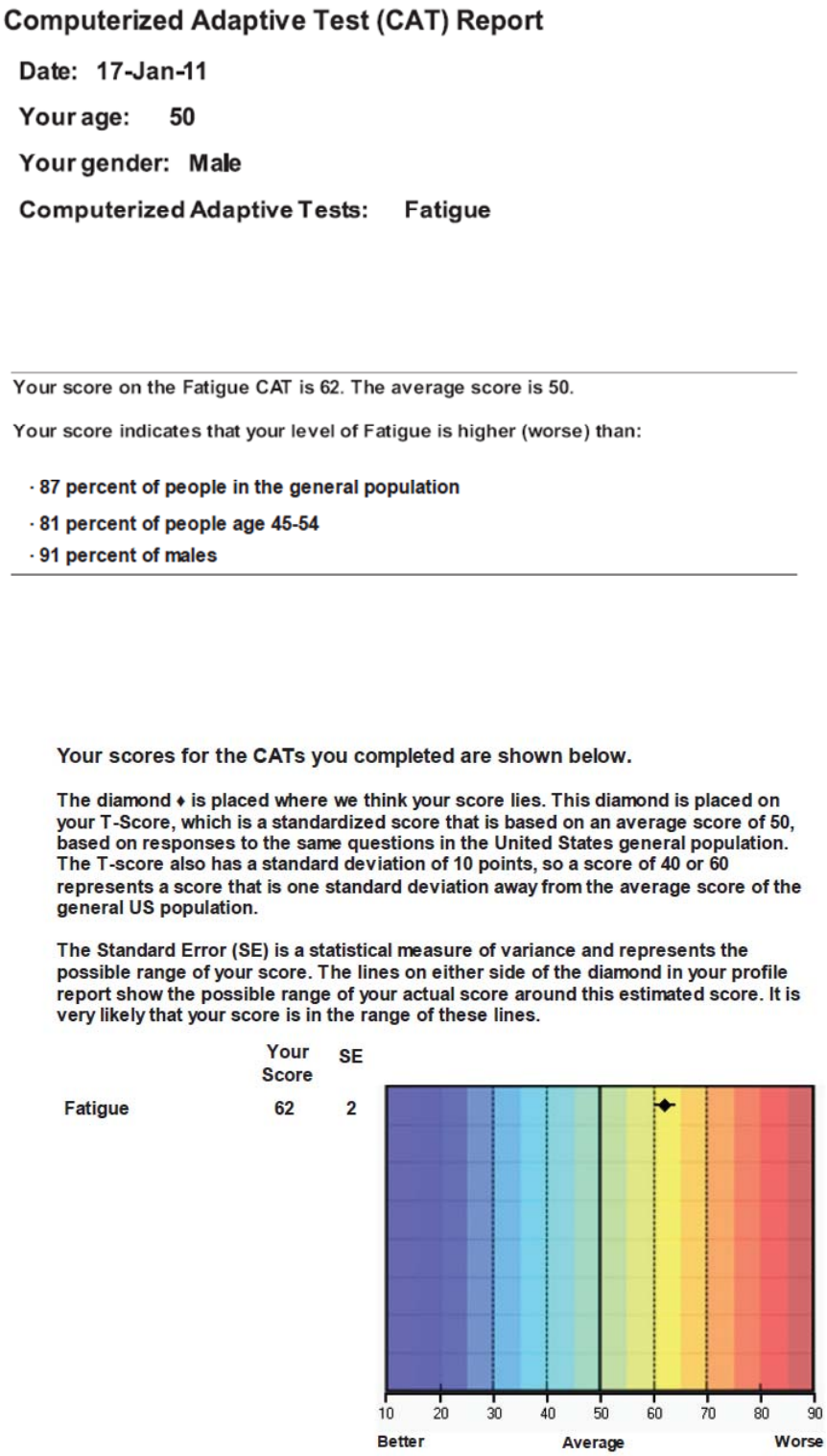
As depicted in Figure 2, Step 1 requires that a provisional trait estimate for the respondent be provided. In this case, no auxiliary information is offered, so the initial trait is set at 0, or the mean of the population distribution. As a result, in Step 2, the most informative item at that trait level, FATIMP3, is selected for administration. The maximum of the item information function for this item is at $\theta = 0.10$, and it is a highly discriminating item at this trait level with a slope parameter (a) equal to 4.77. Assume that when presented with this item about the frequency with which one has to push himself to get things done because of fatigue, the respondent selects "Often" (Category 3) as his response in Step 3. In Step 4, his trait estimate is then updated based on this response. Because he scored in one of the higher response categories, his trait estimate is increased from 0.00 to 1.23 and the standard error associated with the score is 0.42. In Step 5, the respondent's performance is compared against the stopping rules. At this point, the respondent has only answered one item, falling short of the five-item minimum, and the standard error associated with his score is still above the 0.30 maximum. Thus, the CAT returns to Step 2 and selects the best remaining item in the CAT bank.

An3 is selected as the second item for administration, as it is maximally informative at $\theta = 1.20$ and is also a highly discriminating item ($a_{An3} = 4.34$). Suppose the respondent selects "Very much" (Category 5) in response to the statement "I have trouble starting things because I am tired." After selecting the highest category, his trait estimate increases to 1.72 and the standard error decreases to 0.33 with the additional response. Again, the stopping rules are not yet met, so another item must be administered. With a high discrimination parameter ($a_{FATEXP41} = 4.32$) and maximum information at $\theta = 1.10$, FATEXP41 is selected as the third item. Here, the respondent selects "Somewhat"

(Category 3) when describing how run-down he felt on average. In turn, his trait estimate is decreased to 1.39 with a standard error of 0.29. At this point, the standard error is below the 0.30 requirement, but the minimum number of items has not been reached, forcing the CAT to return again to Step 2.

HI7 is the fourth item selected for administration, with a high discrimination parameter of 4.32 and maximum information at $\theta = 1.00$. To this item, the respondent indicates he feels fatigued "Quite a bit" (Category 4), leading to a trait estimate of 1.38 with a standard error of 0.24. One final item is needed to meet the minimum test length of five items. FATEXP35 is the best remaining item, with $a_{FATEXP35} = 4.23$ and maximum information at $\theta = 1.00$. Suppose the respondent indicates that he is bothered "Somewhat" (Category 3) by his fatigue on average. In turn, the trait estimate becomes 1.23 and the standard error is 0.21. At this point, the stopping rules have been met, so the CAT can proceed from Step 5 to Step 6, terminating the assessment.

Once the CAT is complete, the respondent is given a score report produced by the computer (see Figure 4). This report displays his standardized score on the Fatigue CAT, which is 62. The report also informs him that the average fatigue score is 50, and that his level of fatigue is higher (worse) than 87% of people in the general population, 81% of people age 45-54, and 91% of males. Additionally, his location along the fatigue continuum is depicted graphically. A diamond shows where his score of 62 falls and the lines on either side of the diamond show the possible range of his actual score based on the standard error associated with the estimate. The individual falls above the average line in the yellow range, which is another way of indicating that his level of fatigue is higher than most.



Computerized Adaptive Test (CAT) Report

Date: 17-Jan-11

Your age: 50

Your gender: Male

Computerized Adaptive Tests: Fatigue

Your score on the Fatigue CAT is 62. The average score is 50.

Your score indicates that your level of Fatigue is higher (worse) than:

- 87 percent of people in the general population
- 81 percent of people age 45-54
- 91 percent of males

Your scores for the CATs you completed are shown below.

The diamond ♦ is placed where we think your score lies. This diamond is placed on your T-Score, which is a standardized score that is based on an average score of 50, based on responses to the same questions in the United States general population. The T-score also has a standard deviation of 10 points, so a score of 40 or 60 represents a score that is one standard deviation away from the average score of the general US population.

The Standard Error (SE) is a statistical measure of variance and represents the possible range of your score. The lines on either side of the diamond in your profile report show the possible range of your actual score around this estimated score. It is very likely that your score is in the range of these lines.

	Your Score	SE
Fatigue	62	2

Figure 4. Sample PROMIS fatigue score report

Defining LID

To realize the benefits of CAT – that is, achieving precise measurement of individuals' traits with the administration of just a small number of targeted items – the assumptions of the IRT model it is built on must be reasonably met. LII is one of the most fundamental assumptions of IRT. The LII assumption states that an individual's responses to different items are statistically independent after taking trait level into account. That is, conditional on the latent trait and, technically, the item parameters, the probability of observing a particular response to an item is independent of the probability for other items (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). When LII holds, the probability of observing a particular response pattern for an individual is equal to the product of the probability of the observed response on each item, multiplied over all items taken. Formally, this assumption can be expressed by:

$$P(X_1 = x_1, X_2 = x_2, ..., X_J = x_J \mid \theta_i) = \prod_{j=1}^{J} P(X_j = x_j \mid \theta_i), \quad (3)$$

with θ_i representing the *i*th individual's latent trait and x_j being the value of the item response variable X_J (i.e., the response of individual *i* to item *j*).

When LII does not hold, item pairs or groups are said to exhibit LID. In this case, item responses are correlated even after the latent trait is taken into account. Therefore, the probability of observing a particular response pattern is no longer a product of the individual probabilities of the individual item responses, conditional on the latent trait. LID can be positive or negative in direction (Yen, 1993). Given a pair of items, positive LID occurs when an individual performs higher (or lower) than expected on both of the

items. Negative LID occurs when an individual performs higher than expected on one of the LID items, but lower than expected on the other.

Causes of LID

There are a number of reasons pairs or groups of items will exhibit LID. Although researchers have used a variety of different terms to classify types and causes of LID (e.g., Hoskens & De Boeck, 1997; Wainer & Kiely, 1987; Yen, 1993), they all appear to fall into two broad categories: content dependencies and order dependencies.

Content dependences. Items may exhibit LID because they share a common stimulus. For example, respondents may be presented with a series of items based on a reading passage on a verbal test or a chart on a mathematics test. Their responses to the series of items rely not only on their ability level in this case but also on how well they understand the common stimulus. Such sets of items are typically referred to as item bundles or testlets in the IRT literature (Wainer & Kiely, 1987). Other contextual characteristics shared by items, such as a common set of directions or response requirements, may also form LID item sets. These similar features may be incidental and not necessarily related to the purpose of the test (Stocking & Swanson, 1993).

Items may also exhibit LID because they contain similar, finely specified content. In a personality inventory, for example, the same question may essentially be rephrased in different terms. In an educational assessment, items may test the same concept using the same or similar notation but different numerical values. Or, the same content may simply be reversed across the stem and responses. Such items have been referred to as item clones or alternates (Pommerich & Segall, 2008). Lastly, items may exhibit LID

because they give away the answer to one another. Such items are referred to as cross-informational items (Wainer & Kiely, 1987).

Order dependencies. Items can also exhibit LID because of the order in which they are administered. That is, responses to earlier items affect responses to subsequent items. In an educational setting, a respondent may struggle when first presented with a complicated set of directions and items of an unfamiliar type. Once the directions are understood and the respondent finds the key to solving the items, the respondent carries over this learning to the subsequent items (Hoskens & De Boeck, 1997). Alternately, respondents may respond similarly to items towards the end of a test. They may be unable to solve items correctly because they experience fatigue, or, in the case of speeded tests, they may not even reach the final items.

In a survey context, earlier items may also provide a framework for interpreting later items and selecting a response (Tourangeau & Rasinski, 1988). Prior items may become a standard of comparison for making judgments about subsequent items; earlier items can make later items appear more or less extreme. For example, in a survey designed to measure support for abortion under various circumstances, presenting an item stating there is a strong chance of a birth defect may make an item stating that a married woman simply does not want any more children seem more extreme (Schuman & Presser, 1981). In other words, the birth-defect item makes it more difficult to endorse the married-woman item for respondents with identical attitudes towards abortion, violating the assumption of LII.

LID is also well-aligned with what is termed a "carryover effect" in the survey literature on context effects (e.g., Tourangeau & Rasinski, 1988; Tourangeau, Rips, &

Rasisnski, 2000). A carryover effect occurs when judgments of one stimulus are assimilated to earlier judgments of other stimuli. One item can be seen as belonging to the same category or relating to the same general issue as earlier items. Individuals may simply reuse the information that was used to form a response to a previous item instead of thoughtfully assessing the relevance of the material primed by the prior item. Additionally, the same standards or dimensions used to form a judgment about a previous item may be applied to later questions. Or, respondents may feel pressured to provide responses that are consistent with their previous responses to seemingly similar items. Thus, for one or more of these reasons, the individual simply provides a similar response to a later item because of a previous item, violating the LII assumption.

Impact of LID

When items exhibit LID but LII is assumed, the IRT model is misspecified. If not taken into account, LID leads to inaccurate estimation of item parameters, test statistics, and individuals' trait levels (e.g., Chen & Wang, 2007; Kingston & Dorans, 1984; Wainer & Thissen, 1996; Zenisky, Hambleton, and Sireci, 2001). Technically speaking, this bias occurs because the estimation of IRT model parameters is typically based on identifying values of the parameters that maximize the likelihood function (Embretson & Reise, 2000, Chapter 8). When LII is violated, the values that maximize the product of the response probabilities will generally not be those that maximize the true probability function for the response pattern.

Bias in item parameter estimates. Several studies have shown inaccurately high item discrimination parameter estimates for LID items suggesting items are more strongly related to the latent trait than they really are; in other words, LID can result in

misleading assessments of item quality. Tuerlinckx and De Boeck (2001) examined the effects of ignoring LID in the form of positive item interactions on item discrimination parameters by fitting the 2PL model to data generated under a 2PL interaction model. These authors found an overestimation of item discriminations suggesting the items are better able to discriminate between individuals at trait levels near the items' discrimination parameters than they really are. Expanding upon this research, Chen and Wang (2007) found that LID in the form of positive item interactions led to an overestimation of item discrimination parameters and an underestimation of item difficulty parameters in their work with the 3PL (Birnbaum, 1968) and the GPCM (Muraki, 1992). The overestimation of the discrimination parameters became more severe as the guessing parameters approached zero. Using TOEFL listening and reading comprehension data, Wainer & Wang (2000) compared the item parameter estimates obtained with a model that allowed for LID among subsets of items, namely a random-effects testlet model, with those obtained through traditional modeling. They found estimates of item difficulties were comparable across the two modeling approaches, but item discriminations were underestimated for listening comprehension items and overestimated for reading comprehension items, and guessing parameters were overestimated for both types of items when LID is ignored.

Bias in test statistic estimates. Several studies have compared estimates of test information and score reliability using a model that ignores local dependence and one that accommodates local dependence. Thissen, Steinberg, and Mooney (1989) showed that test information curves were artificially inflated due to LID in a reading comprehension test comprised of four passages with several questions following each

passage. These results were obtained by fitting a traditional IRT model for dichotomously scored items as well as an IRT model for polytomously scored items to rescored data where the dichotomously scored LID items had been rescored as a single polytomously scored item.

Using this same approach, Yen (1993) identified artificially inflated information curves when LID items from language arts and mathematics performance assessments were treated as independent dichotomous items. Sireci, Thissen, and Wainer (1991) and Zenisky, Hambleton, and Sireci (2002) found inflated reliability estimates for an experimental SAT verbal form and the Medical College Admissions Tests (MCAT), respectively, both of which showed evidence of passage-based LID. These results were again based on a comparison of results from IRT models for dichotomously and polytomously scored items.

Wang, Bradlow, and Wainer (2002), via simulations and the analysis of real data, demonstrated that estimates of tests' precision are overly optimistic when LII is incorrectly assumed. Their analyses considered tests in which some or all of the items are polytomously scored, and compared results across a traditional IRT model and a random-effects testlet model which models LID among subsets of items. Using TOEFL data, Wainer and Lukhele (1997) and Wainer and Wang (2000) also found reliability and test information to be substantially overestimated.

Accommodating LID

If LID is present, a number of approaches have been applied to reduce its impact on parameter estimates, some of which have been extended to adaptive settings. Broadly speaking, these approaches either (1) focus on test construction practices on the front end

to prevent LID from occurring or (2) use expanded IRT models on the back end to properly account for LID in score estimates.

Prevent LID through test construction. Researchers have advocated the use of extensive test specifications to help reduce the impact of LID by preventing it on the front end (Kingsbury & Zara, 1989; Lord, 1977; Stocking & Swanson, 1993). Test specifications provide rules for developing instrument forms in terms of key item characteristics such as the domain and cognitive processes they are tapping and the number of items required of each type. More extensive sets of item characteristics can include format and appearance, similarity to other items, and statistical properties.

Test specifications have been incorporated in adaptive settings by placing constraints on the CAT selection algorithm. The weighted-deviations model (Stocking & Swanson, 1993) and constrained CAT using shadow testing (van der Linden & Reese, 1998) are popular examples of CAT variations with constraints. Instead of simply selecting the maximally informative item, these algorithms select the most informative item that best satisfies the remaining test specifications. For example, researchers can constrain the selection algorithm to prevent cross-informational items from appearing in the same administration. Additionally, constraints can be placed on item content, ensuring each administration contains the same "mix" of items. Thus, constrained algorithms reduce the impact of LID because they maintain more consistency in item content and order across administrations.

Stocking and Swanson's (1993) weighted deviations model and associated algorithm for severely constrained item selection in adaptive testing is particularly useful when one or more of the item constraints cannot be satisfied simultaneously with another.

Constraints are viewed as desired – but not required – properties for a selection algorithm. Test developers can even weight the desired properties to maintain some level of control over which constraints are imperative and which can be relaxed.

Constrained CAT using shadow testing (van der Linden & Reese, 1998) is another general method used to control test quality by introducing constraints in the item selection process. The key difference between this linear-programming-based technique and traditional CAT algorithms is that items are selected from a shadow test instead of the item bank directly. Shadow tests are tests assembled prior to the selection of each item that contain all items already administered to the examinee, are optimized at the individual's current trait estimate, and meet all required test specifications (van der Linden & Chang, 2003). The most informative item that has not yet been administered is then selected from the shadow test, and the process is repeated. Because the shadow test from which the item is selected meets all test specifications, so does the adaptive test.

Model LID through expanded IRT models. Instead of trying to prevent LID on the front end, researchers have proposed ways to model or score response data on the back end to reduce the impact of LID. One early approach proposed to accommodate LID among subsets of dichotomous items consisted of grouping LID items into a super-item known as a testlet, and modeling the number correct using an IRT model for polytomously scored items (e.g., Masters & Evans, 1986; Sireci, Wainer, & Thissen, 1991; Wainer & Kiely, 1987). The LID among items within a testlet is absorbed into that testlet score, and the assumption of LII can be met between testlets.

This approach to mitigate the effects of LID has been applied in CAT settings as testlet-based adaptive testing (Wainer & Kiely, 1987). Testlets are substituted for single

items as the unit of administration in a CAT, meaning the algorithm uses the testlet rather than an individual item as a branching point. This approach helps to minimize LID because each item is embedded in a pre-determined testlet, carrying its own context with it. Item order and content dependencies are localized because individuals see items sharing content together and in the same order within each testlet. However, this approach has been criticized in terms of its efficiency; longer tests are required to achieve the same level of precision because only the total testlet score is modeled and the information contained in the precise pattern of responses within each testlet is lost (Kingsbury & Zara, 1989).

Additional item parameters can also be incorporated into traditional IRT models that assume LII to model LID. Several researchers proposed the inclusion of fixed item-interaction terms (e.g., Chen & Wang, 2007; Hoskens & De Boeck, 1997; Jannarone, 1986). These models express the likelihood as a product of the probabilities of the responses to subsets of LID items, as opposed to individual item responses. The item interaction term associated with two or more items represents the additional difficulty (or easiness) of jointly solving all of the items in the set correctly (Hoskens & De Boeck, 1997). This interaction term can also be specified as a linear function of the latent trait, such that the interaction between the items depends on the individual's trait level.

Random-effects testlet models (e.g., Bradlow, Wainer, & Wang, 1999; Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005) expand standard IRT models to include an additional parameter that represents the interaction between an individual and a given item nested within a testlet. The greater the variance associated with this parameter, the greater the amount of LID among the items. This testlet effect essentially

"absorbs" unwanted LID among items. LII can be achieved by conditioning on the testlet effect parameter in addition to the latent trait and standard item parameters. The random-effects testlet model can then be used during operational CAT administrations to update an individual's trait estimate and calculate a final score, automatically accounting for LID (Wainer, Bradlow, & Du, 2000).

Researchers have also re-parameterized traditional IRT models under a hierarchical generalized linear modeling framework (HGLM) to accommodate LID (e.g., Adams, Wilson, & Wu, 1997; Jiao, Wang, & Kamata, 2005; Kamata, 1999, 2001). Hierarchical linear models relax the assumption of independence of observations across respondents that is made in traditional linear models, permitting the modeling of nested data structures (Bryk & Raudenbush, 1992). Reformulating the IRT model as a hierarchical model allows a subset of LID items to be nested within a testlet, and testlets to be nested within individuals. As a result, the dependencies among responses to items within a testlet can be appropriately accounted for under the HGLM framework.

Performance of accommodation approaches in CAT settings. Several studies have compared the performance of various approaches used to accommodate LID in CAT settings, against one another and / or against a traditional approach in which LID concerns are ignored (Boyd, 2003; Keng, 2008; Schnipke & Reese, 1997; Stocking & Swanson, 1993; van der Linden, 2005; van der Linden & Reese, 1998). Results generally show that although the alternative approaches are slightly less efficient in terms of measurement precision, the differences are often of practical negligence. Furthermore, the alternative approaches are better able to meet non-psychometric criteria and test specifications.

Schnipke and Reese (1997) compared several testlet-based CAT designs against the standard item-level CAT design and a traditional paper-and-pencil design via simulation study. They found that, because it adapts the difficulty after each item, the item-level CAT produced the least error and bias in trait estimates, while the non-adaptive paper-and-pencil produced the most. All testlet-based designs resulted in improved precision over the paper-and-pencil test, achieving nearly the same level of precision in half the number of items. Although the item-level CAT design was optimal from a statistical perspective, the testlet-based CAT designs performed at an acceptable level and offered non-psychometric advantages. Also via simulation, Keng (2008) compared a CAT which adapted at the item level against one which adapted at the testlet level. He found both yielded similar and good measurement accuracy, though the precision of the item-level CAT was slightly better.

Via simulation, Boyd (2003) compared the performance of a CAT based on the random-effects testlet model (Wainer, Bradlow, & Du, 2000) against one where LID items were grouped together and scored using a polytomous IRT model. The two modeling approaches actually performed similarly in terms of measurement precision, but, unlike the polytomous IRT model, the random-effects testlet model was able to provide information about individuals' item-level response patterns within testlets.

Stocking and Swanson (1993) compared the performance of the weighted-deviations model against an unconstrained CAT in a simulation context. They found that, in terms of measurement precision alone, the unconstrained CAT performed best; however, the simulated tests were unsatisfactory from a content-balancing perspective. Via simulation, van der Linden and Reese (1998) showed that a CAT with linear-

programming-based constraints performed nearly as well as the unconstrained CAT in terms of measurement precision, particularly for longer tests, while also meeting content and item exposure requirements.

van der Linden and Reese (1998) and van der Linden (2005) stress the superiority of linear-programming-based constraints over Stocking and Swanson's (1993) weighted deviations model. These authors noted that the weighted deviations model requires the test developer to specify weights for all constraints, and that with the potential for hundreds of constraints, this task can become unwieldy. Also, because constraints are seen as desirable but not mandatory in the weighted-deviations model, constraints with low weights may be violated often (van der Linden, 2005). Via simulation, van der Linden and Reese (1998) showed how the linear-programming model was able to automatically meet all constraints of the model without unpredictable violations. van der Linden (2005) demonstrated the advantages of the linear-programming model through an empirical comparison of data from an adaptive version of the Law School Admission Test (LSAT).

Simulating LID

To systematically investigate the impact of LID on various outcome measures or evaluate the effectiveness of strategies used to accommodate LID, researchers frequently perform simulation studies (e.g., Chen & Wang, 2007; Jiao, Wang, & Kamata, 2007; Pommerich & Segall, 2008). These studies use (multiple) generated datasets that have known properties, such a prescribed dependency structure. There are two general modeling approaches used to simulate LID among item responses: models of underlying

local dependence and models of surface local dependence. A brief review of these approaches follows, and example scenarios to which the models apply are included.

Underlying local dependence. Underlying local dependence (ULD) models assume that there is a separate trait that is common to each set of locally dependent items but is not common to the rest of the items on the instrument (Chen & Thissen, 1997; Levy, Mislevy, & Sinharay, 2009; Thissen, Bender, Chen, Hayashi, & Wiesen, 1992). In other words, all items have a non-zero weight, or "loading" on the common trait. Then, a given item also has a non-zero weight on the non-common trait(s) it is associated with and a weight equal to zero on any remaining non-common traits. Data generated under this model is multidimensional with correlated dimensions.

Under the ULD model, the proportion of items loading on non-common traits in addition to the common trait can be manipulated, as can the correlation between the common and non-common dimensions. Also, the strength of the loadings on the associated non-common traits can be varied to reflect degrees of multidimensionality. Strong multidimensionality is represented when the average weights of the items on the common and non-common traits are equal, whereas weak multidimensionality is represented when the average weights on the common trait are larger than those on the non-common traits.

Simulating LID under the ULD model is conceptually similar to simulating LID using other multidimensional IRT models, including dimension-dependent interaction models (Hoskens & De Boeck, 1997; Turlinckx & De Boeck, 2001), random-effect testlet models (e.g., Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Zu, 2000), and hierarchical generalized linear models (Jiao, Wang, & Kamata, 2005; Kamata 2001).

Broadly speaking, these models include random effects that represent the interaction between one's trait level and an item cluster. Because random variables are included, the resulting models are multidimensional.

Using the ULD model to simulate LID would be appropriate for the following types of scenarios:

- A reading comprehension test with a set of items following each passage (Chen & Thissen, 1997)
- Math content clusters where questions in the same cluster depend on the same graph or data table (Jiao, Wang, & Kamata, 2007)
- Items tied to the same scenario in a scenario-based science assessment (Jiao, Kamata, Wang, & Jin, 2010)
- Groups of items utilizing the same response format (Yen, 1993)
- Items sharing a particular content, not necessarily relevant to the concept being tested, which an examinee has prior interest in, knowledge of, or exposure to (Yen, 1993)
- Any testlet in which a common stimulus is used for a subset of items (Lin, Kim, & Cohen, 2006)

Surface local dependence. Surface local dependence (SLD) models assume that items are so similar in content or location on an instrument that individuals respond identically to the second item without the underlying process implied by the IRT model (Ackerman & Spray, 1987; Chen & Thissen, 1997; Pommerich & Segall, 2008; Thissen et al., 1992). That is, they provide the same response to the second item without regard to θ or the statistical characteristics of the second item. Under the SLD model, the

probability with which individuals respond identically can be varied to represent the strength of the dependency. That is, the higher the π_{LID} value, the stronger the dependency. Also, the number of items on the instrument affected by SLD can be manipulated.

Simulating LID using the SLD model is conceptually similar to simulating it using constant item interaction models (e.g., Hoskens and De Boeck, 1997; Tuerlinckx & De Boeck, 2001). Under this modeling framework, the response to one item has consequences for the response probabilities for another item or items. An extra parameter that is constant across persons is added to the IRT model that reflects the increased (or decreased) chance of getting a second item right if the first item is answered correctly. Unlike the ULD or multidimensional case where item responses are affected by an additional latent variable that has not be accounted for, here the variable that affects the responses is itself an item response, or a manifest variable.

Using the SLD model to simulate LID would be appropriate for the following types of scenarios:

- A speeded test where some of the items towards the end are not reached due to the time constraint (Chen & Thissen, 1997; Lin, Kim, & Cohen, 2006)
- A lengthy test where examinees experience fatigue or low motivation and answer items towards the end incorrectly (Yen, 1993)
- A test with similarly-worded items (Chen & Thissen, 1997)

- Multiple-part or nested questions where the correct solution to one part or item must be achieved before the solution to subsequent parts can be achieved (Ackerman & Spray, 1987)
- Enemy items where one item inadvertently "cues" the correct response to other items on the same test (Ackerman & Spray, 1987)
- Items testing the same concept using similar content, wording, and/or notation (Pommerich & Segall, 2008)
- Items on a personality or attitudinal questionnaire where a respondent strives to provide responses that are consistent with his earlier responses (Hoskens & De Boeck, 1997)
- Items of an unfamiliar type where the examinee struggles with early items but learns the key to solving later items (Hoskens & De Boeck, 1997)

Detecting LID

The success of the aforementioned modeling approaches and accommodation techniques in terms of controlling LID is dependent on the ability to determine which items exhibit it. LID manifests itself by the IRT model failing to account for unmodeled associations between items. Exploratory approaches based on pairwise statistics have been proposed to detect LID by pin-pointing instances of unmodeled associations between specific pairs of items. Some of the most well-known statistics include the Q_3 statistic (Yen, 1984; 1993) and the X^2 statistic (Chen & Thissen, 1997). Others include the likelihood ratio G^2 statistic (Chen & Thissen, 1997), the absolute value mutual information difference statistic (Tsai & Hsu, 2005), the power-divergence statistic

(Cressie & Read, 1984), a Mantel-Haenszel statistic (Agresti, 2002), residual correlations in factor analysis, and modification indices in structural equation modeling.

Some indexes of multidimensionality (e.g., Hattie, 1985) can also function as indicators of LID because the concepts of multidimensionality and LID are interrelated (e.g., Hambleton, Swaminathan, & Rogers, 1991; Ip, 2001). LII is obtained when all of the latent dimensions influencing performance have been taken into account (i.e., are included) in the IRT model. Associations among items in the form of LID may be present, for example, when a unidimensional model is specified for an underlying model that is actually multidimensional. Well-known methods of multidimensionality assessment include a non-parametric clustering technique called HCA/CCPROX (Roussos, Stout, & Marden, 1998), Stout's non-parametric T statistic in the DIMTEST program (Stout, 1987), and the non-parametric DETECT index (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996).

These techniques are based on functions of conditional covariances between item pairs, which form the basis of a weakened assumption of essential local independence within a non-parametric IRT framework. However, they only offer evidence of the presence of a "nuisance dimension," or overall LID effect, and most of them do not point to the specific location of the LID (for an exception see the HCA/CCPROX routine). Since the PROMIS data bank uses parametric IRT models and uses a complex item sampling design, these non-parametric methods will not be considered further in this dissertation.

The Q_3 and X^2 are selected in this study for several reasons. First, the Q_3 and X^2 can be considered parametric IRT-based LID statistics because they require item or

person or both parameter estimates in the calculation (Kim et al., 2007). Second, unlike some LID statistics, they have been extended to accommodate polytomous response data in addition to dichotomous response data (Lin, Kim, & Cohen, 2006). Third, the Q_3 and X^2 are among the most frequently applied and well-studied LID statistics (e.g., Chen & Thissen, 1997; Chen & Wang, 2007; Habing, Finch, & Roberts, 2005; Ip, 2001; Kim, De Ayala, Ferdous, & Nering, 2007; Pommerich & Ito; 2008; Pommerich & Segall, 2008; Yen, 1984, 1993; Zenisky, Hambleton, & Sireci, 2002). The inclusion of the Q_3 and / or X^2 in software programs used to detect LID for dichotomous and polytomous response data, such as IRT_LD (Chen, 1993), LDID (Kim, Cohen, & Lin, in press), IRTFIT (Bjorner, Smith, Stone, & Sun, 2007), and LDIP (Kim, Cohen, & Lin, 2006) provides additional evidence of the popularity of these statistics.

Observed and expected frequencies. In many cases, LID statistics for item pairs are based on a comparison of observed and expected frequencies. Following Lin, Kim, and Cohen (2006), for two items, *j* and *j*', both with an equal number of *K* response categories, possible combinations of pairwise item responses can be represented in a *K* by *K* table, with $O_{kk'}$ representing the observed frequency for the k^{th} row and k'^{th} column. For example, the observed frequencies for a pair of items accompanied by a five-point response scale are displayed in Table 2.

Table 2. Observed frequencies for a pair of items with $K = 5$

		Item j'				
		k'=1	k'=2	k'=3	k'=4	k'=5
Item j	k=1	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}
	k=2	O_{21}	O_{22}	O_{23}	O_{24}	O_{25}
	k=3	O_{31}	O_{32}	O_{33}	O_{34}	O_{35}
	k=4	O_{41}	O_{42}	O_{43}	O_{44}	O_{45}
	k=5	O_{51}	O_{52}	O_{53}	O_{54}	O_{55}

Note that the sum of $O_{kk'}$ across all cells is equal to N, or the total number of individuals (in the case of no missing data):

$$N = \sum_{k=1}^{K} \sum_{k'=1}^{K} O_{kk'} \ . \tag{4}$$

The expected frequencies for a pair of item scores ($E_{kk'}$) are predicted by the combination of the category response functions for item j ($P_{jk}(\theta)$) and item j' ($P_{j'k'}(\theta)$) and the sample size for the analysis, N, as

$$E_{kk'} = N \int_{-\infty}^{\infty} P_{jk}(\theta) P_{j'k'}(\theta) g(\theta) d\theta \ , \tag{5}$$

where $g(\theta)$ is the population ability distribution, typically assumed to be $\theta \sim N(0,1)$. The expected frequencies for a pair of items are displayed in Table 3.

Table 3. Expected frequencies for a pair of items with $K = 5$

		Item j'				
		k'=1	k'=2	k'=3	k'=4	k'=5
Item j	k=1	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}
	k=2	E_{21}	E_{22}	E_{23}	E_{24}	E_{25}
	k=3	E_{31}	E_{32}	E_{33}	E_{34}	E_{35}
	k=4	E_{41}	E_{42}	E_{43}	E_{44}	E_{45}
	k=5	E_{51}	E_{52}	E_{53}	E_{54}	E_{55}

Again, the sum of $E_{kk'}$ across all cells is equal to *N*, or the total number of individuals, when there is no missing data:

$$N = \sum_{k=1}^{K} \sum_{k'=1}^{K} E_{kk'} . \quad (6)$$

In applications, expected cell counts are often computed by calculating probabilities of a given response pattern at each of the quadrature points of a normal distribution, converting them to frequencies, and integrating them over the normal distribution (e.g., Pommerich & Ito, 2008). For example, the LDIP computer program uses 41 equally spaced points from -4 to 4 with an increment of 0.20 to approximate the integral (Kim, Cohen, and Lin, 2005); similarly, IRTFIT uses numerical integration to evaluate the integral (Bjorner et al., 2007).

Here, the normal probability density function is used in the calculation to derive expected cell counts. Roughly speaking, the joint conditional probabilities are "weighted" by the relative frequencies of different values for the random variable (i.e., θ). This means that values near the mean of the normal distribution, or zero, contribute more weight than values near the tails of the distribution.

An alternate approach to calculating expected cell counts follows techniques that have been applied in the IRT linking literature (e.g., Stocking & Lord, 1983) in which trait distributions are empirically approximately from the sample rather than theoretically assumed:

$$E_{kk'} = \sum_{i}^{N_{jj'}} P_{jk}(\hat{\theta}_i) P_{j'k'}(\hat{\theta}_i) , \quad (7)$$

where $N_{jj'}$ is the number of individuals responding to the pair of items j and j', $\hat{\theta}_i$ is the trait estimate for individual i, and $P_{jk}(\hat{\theta})$ and $P_{j'k'}(\hat{\theta})$ are the estimated category response functions for item j and item j', respectively.

This approach for determining expected cell counts differs from that in Equation 5 in two critical ways. First, this approach uses estimates of individuals' trait levels in the calculations. Second, instead of integrating over a common trait distribution, this approach sums the joint probability of observing each response combination across individuals (calculated as the product of conditional response probability categories given θ for the two items assuming LII) to estimate expected cell counts. In other words, it approximates the trait distribution for individuals responding to that item pair by using the trait estimates of the individuals actually responding to it in the sample. This means, for example, that if only low-level individuals respond to a given pair of items, the joint conditional probabilities of higher-level individuals are irrelevant for the calculation of the expected cell counts. In effect, this approach removes assumptions about the trait distribution and allows the distribution to vary across each item pair. Assuming no missing data, the expected cell counts calculated using Equation 7 will converge with those calculated using Equation 5 as the sample size increases and the error associated with the trait estimates decreases.

Pearson statistic. Pearson's statistic, X^2, is defined as (Chen & Thissen, 1997):

$$X^2 = \sum_{k=1}^{K}\sum_{k'=1}^{K'}\frac{(O_{kk'} - E_{kk'})^2}{E_{kk'}}. \quad (8)$$

It tests the null hypothesis (H_0) that observed cell frequencies equal those predicted by the model. The X^2 takes the minimum value, zero, when all $O_{kk'} = E_{kk'}$. Holding sample size constant, larger differences between $O_{kk'}$ and $E_{kk'}$ yield larger values of X^2, and stronger evidence against the null hypothesis. The X^2 has approximately a χ^2 distribution with $(K\text{-}1)^2$ degrees of freedom. Note that the X^2 is a non-directional LID statistic, able to detect the presence of LID but not the direction (i.e., positive or negative).

Yen's Q_3 statistic. Yen's (1984; 1993) Q_3 statistic represents the correlation between two items after accounting for performance on all items. To calculate the Q_3 statistic, a trait estimate is first obtained for each individual from an IRT model that assumes LII. Given this estimate, residuals are calculated as the difference between the response predicted by the IRT model and the response actually observed. These residuals are then correlated among item pairs over all individuals. Formally, the statistic is defined as:

$$Q_{3_{jj'}} = r_{e_{ij}e_{ij'}}, \tag{9}$$

where *r* refers to the correlation, $e_{ij} = X_{ij} - E(X_{ij})$, X_{ij} is the observed response, and $E(X_{ij})$ is the expected response of the individual on the item, given by the item response function.

Yen (1993) showed the expected value of the Q_3 statistic when LII holds is approximately -1/(*J*-1), where *J* is test length. If item pairs do not exhibit local dependence, the value of the Q_3 statistic should be near zero in large samples; large (absolute) values suggest item pairs share some other common cause beyond the latent

trait(s) specified in the model. Note that the Q_3 is a directional LID statistic, able to identify both positive and negative LID.

In practice, a cut-off of ±0.20 has been used to screen items for LID (Chen & Thissen, 1997; Yen, 1993). Some researchers have obtained optimal cut-points or critical values empirically through the simulation of LII data (Chen & Thissen, 1997; Chen & Wang, 2007; Kim et al., 2007); in this approach, Q_3 values located beyond the 95% interval of the empirical distribution are flagged. Others have applied a Fisher transformation to the raw Q_3 values and used conventional cut-off values for the standardized normal distribution (Chen & Thissen, 1997; Kim et al., 2007; Lin, Kim, & Cohen, 2006).

Examining Properties of LID Statistics

Several studies have been conducted to specifically examine the properties of pairwise LID statistics under various conditions. Popular LID statistics have been applied to dichotomous non-adaptive data (Chen & Thissen, 1997; Kim et al., 2007; Levy, Mislevy, & Sinharay, 2009; Yen, 1984), polytomous non-adaptive data (Lin, Kim, & Cohen, 2006), and dichotomous adaptive data (Pommerich & Ito, 2008; Pommerich & Segall, 2008). Findings of these studies with regards to the Q_3 and X^2 statistic are reviewed below and summarized in Table 4. The current study draws heavily upon this line of research and extends it to examine the properties of these common LID statistics when applied to polytomous adaptive data.

Table 4. Summary of literature related to properties of the Q_3 and X^2

Author(s)	Administration Condition(s)	Data Type	Measures Included	LID Simulation	Summary of Key Findings
Yen (1984)	CONV	Dichotomous	Q_3	ULD	• Q_3 statistic more negative than expected in null case • Q_3 high power to detect LID
Chen & Thissen (1997)	CONV	Dichotomous	X^2 & Q_3	ULD & SLD	• Q_3 did not follow normal distribution in null case • X^2 less able to detect ULD, equally able to detect SLD, and had lower false-positive rate than Q_3
Kim et al. (2007)	CONV	Dichotomous	Q_3	ULD	• Q_3 had relatively high power and low type-I error rate • Power of the Q_3 increased as test length and LID level increased
Levy, Mislevy, & Sinharay (2009)	CONV	Dichotomous	X^2 & Q_3	ULD	• X^2 did not function properly under the null condition and displayed minimal power • Q_3 demonstrated superiority with more power and lower false positive rate
Lin, Kim, & Cohen (2006)	CONV	Polytomous	X^2 & Q_3	ULD & SLD	• X^2 mirrored theoretical χ^2 distribution in null condition but expected X^2 values were inflated as test length increased • Q_3 mirrored standard normal distribution in null condition and was not affected by test length • Both X^2 and Q_3 powerful enough to detect both ULD and SLD
Pommerich & Ito (2008)	CONV & CAT	Dichotomous	X^2 & Q_3	SLD	• X^2 did not function properly and yielded unrealistically large values in the null condition
Pommerich & Segall (2008)	CAT	Dichotomous	X^2 & Q_3	SLD	• X^2 did not function properly and yielded unrealistically large values in the null condition • Q_3 values in the null condition were close to expected

Dichotomous non-adaptive data. Yen (1984) proposed the Q_3 statistic and investigated its properties, along with an existing LID statistic, the Q_2, when applied to dichotomous non-adaptive data. To examine the null distribution of the statistics, unidimensional data for a 20-item and 40-item test were simulated using the 3PL. Data were then generated under a two-dimensional 3PL model to ascertain the statistics' ability to detect LID. The effect of the two dimensions on the items were manipulated via their discrimination parameters, such that (1) all items were strongly influenced by both underlying traits, (2) only a subset of items was strongly influenced by both traits, and (3) only a subset of items was influenced weakly by the second trait. Under the null condition, the means of the Q_3 statistics were more negative than expected. When data were generated under the multidimensional model, the Q_3 was able to detect LID amongst sets of items that were influenced by both underlying traits, even when the influence of the second trait was weak.

Chen and Thissen (1997) studied the ability of four statistics, including the X^2, to detect ULD and SLD as compared to the Q_3. Given ULD, the number of specific traits and the items' weights on the specific factors relative to their weight on the general factor were manipulated in the simulation study to represent strong and weak multidimensionality. For SLD, the probability with which an individual responds identically to a second item was manipulated in the simulation to represent low, moderate, and strong LID. The number of items was manipulated across all ULD and SLD conditions as well to represent short, medium, and long tests. They concluded that compared to the Q_3, the X^2 was less able to detect ULD, equally able to detect SLD, and has a lower false-positive rate given no LID. Lastly, although they found Q_3 to be an

effective measure for detecting LID, they noted it did not exhibit the assumed $N(0,1)$ distribution under null conditions.

Kim et al. (2007) examined the relative performance of a number of directional and non-directional LID statistics, including the Q_3. These statistics were examined under a null condition in which LII was true and under a condition in which LII was violated. The 3PL testlet model proposed by Wainer, Bradlow, and Du (2000) was used to generate responses with a dependence structure. Test length, LID level, and LID item percentage were factors manipulated in the simulation. Relative performance of the measures was based on the type-I error / false-positive as well as the type-II error / false-negative rate and associated power. Results showed as the test length and LID level increased, the power of the LID statistics was similar. No LID statistic performed the best with regards to all three evaluation criteria, though the Q_3 was identified as one that could be recommended for most of the LID conditions because of its relatively high power and low type-I error rate.

Levy, Mislevy, and Sinharay (2009) examined a collection of LID statistics in the context of posterior predictive model checking (PPMC) under a Bayesian framework. They manipulated factors affecting dimensionality, including the strength of items' dependence on auxiliary dimensions, correlations among the dimensions, the proportion of items exhibiting multidimensionality, and the sample size. The X^2 measure did not function properly under the null condition in this context and displayed minimal power given LID. The Q_3 statistic was among one of the few that demonstrated superiority under both the LII and LID conditions.

Polytomous non-adaptive data. Lin, Kim, and Cohen (2006) examined the ability of four LID statistics, including the X^2 and Q_3, to detect LID in polytomous data. They considered two types of dependency resulting from testlets (i.e., multidimensional data) and from speeded tests. In the multidimensional situation, the number of items, number of dimensions, and the categories of responses were factors in the simulation. In speeded tests, the responses to not-reached items were scored as incorrect, or assigned to the lowest score category, yielding identical responses across unreached items. For this LID condition, the ratio of maximum number of missing items was a factor in the simulation, along with number of items and the categories of responses. Under the LII condition, the distributions of the X^2 values across replications mirrored the expected χ^2 distribution. However, as the number of items increased, the means and standard deviations became slightly larger than the expected values. The mean and standard deviation of the Q_3 was not similarly affected by test length. Under the LID conditions, the authors deemed both the Q_3 and X^2 powerful enough to detect LID among item pairs which (1) shared a specific dimension in addition to the common dimension in the multidimensional condition and (2) were both omitted and given a score in the lowest category in the speeded test condition.

Dichotomous adaptive data. LID statistics were first applied to adaptive data by Pommerich and Segall (2008) in order to evaluate the extent to which LID was present in an operational CAT and whether LID negatively affects score precision. Their simulations of CAT administrations were based on a test of mathematics knowledge which used a 3PL model for item selection and scoring. For LII item pairs, responses were generated according to the 3PL. For LID item pairs, responses were generated using

a variation of the SLD model proposed by Chen and Thissen (1997). The authors investigated the performance of the X^2 and Q_3 statistics in their simulation and found that the X^2 statistic did not function properly when applied to CAT data while the results for the Q_3 statistic looked plausible. Specifically, values of the X^2 statistic appeared unrealistically large even for item pairs with no LID induced.

Given Pommerich and Segall's (2008) findings, Pommerich and Ito (2008) explicitly set out to examine the properties of LID measures when applied to adaptive data. They note that LID statistics may behave differently across adaptive and fixed settings because items are administered across different distributions of examinee ability. Again, the X^2 and Q_3 statistics were investigated in the study. Three types of administration were simulated: a conventional administration of all items in the pool, an administration of 15 randomly-selected items from the pool, and a CAT administration of 15 items. LID was simulated using Pommerich and Segall's (2008) modification of the approach used by Chen and Thissen (1997) to simulate SLD. Only one level of LID – nearly perfect LID – was considered. Supporting Pommerich and Segall's (2008) findings, the average Q_3 values for each LII item pair under each administration condition were close to the value that is expected when no LID exists, and their standard deviations across replications were small. In contrast, the average X^2 values for each LII item pair were close to or below their expected value with no LID for the first two administration conditions, but implausibly large for the CAT administration, with large standard deviations.

When Pommerich and Ito (2008) induced LID among item pairs, the Q_3 statistic was substantially above the expected value given LII across all administration conditions,

showing strong evidence of LID in the responses. The X^2 statistic was again exaggerated for both LID and LII item pairs under CAT administration, even after controlling for the number of times an item pair was administered together (i.e., sample size). These findings led the authors to conclude that the X^2 may not be usable with CAT data. They attributed the differences in the performance of the two statistics to the level at which they function: the X^2 operates at an aggregate level to derive expected and observed frequencies while the Q_3 operates first at an individual level to compare expected and observed performance before these differences are summed over individuals. Furthermore, the aggregate-level calculations for the X^2 assume a normal ability distribution for both observed and expected performance, a condition which is not likely met in CAT settings. In contrast, such distributional assumptions are irrelevant in the calculation of the Q_3.

Chapter 3: Methods

Objective

The objective of this study is to expand the literature on LID detection to evaluate the performance of two widely-used pairwise statistics, Yen's Q_3 statistic and Pearson's Statistic X^2, when applied to polytomous adaptive data. As described in the previous chapter, the goal in practice is to identify LID in a pre-calibration stage (i.e., Step 4 of Table 1) and to "account" for it so that the item parameters in the item bank are based on models without residual LID. Test developers can account for LID when designing the pre-calibration study (e.g., by instructions, item administration, or matrix sampling) and / or estimating the item parameters (e.g., by removing enemy items or fitting a testlet model to the data). However, there is a need to know how to apply LID statistics to polytomous CAT data to see if, in fact, these strategies were effective and there is no evidence of LID in the CAT data (i.e., Step 7 of Table 1).

The use of LID statistics in adaptive settings may also be required to determine whether LID that was not present during pre-calibration emerged due to the mode of administration (or any other difference between pre-calibration and operational CAT administrations, such as motivation). For example, assume that in pre-calibration, two similarly-worded items are separated by a number of other items on a lengthy, fixed form test. Then, in operational administrations, the same two items are selected as a part of a much shorter CAT administration. With a smaller group of items and fewer items in between the LID pair, respondents fail to see the second item as a unique item and simply offered the same response they offered to the first, producing LID. Now there is CAT-specific LID, or LID that was not present during pre-calibration but manifests itself

because of the fact that items are administered adaptively. This scenario is plausible, particularly in survey and health-outcomes contexts where individuals are known to respond differently to subjective items based on context, order, proximity to similar items, etc. (e.g., Couper, Traugott, & Lamias, 2001; Dillman & Smyth, 2007; Smyth, Dillman, Christian, & Stern, 2006).

For these reasons, the study attempts to determine if Yen's Q_3 statistic and Pearson's Statistic X^2 are usable with polytomous CAT data by addressing the following research questions via a simulation study:

1. What impact do the administration condition and level of LID have on the magnitude of the LID statistics in the polytomous adaptive context?
2. Do the statistics perform as expected with polytomous adaptive data given LII, in that they produce results near their respective expected values and false positive rates remain low?
3. Are the statistics powerful enough to detect varying levels of LID in polytomous adaptive data, such that even item pairs exhibiting low levels of LID are frequently identified?
4. Should unique cut-offs for the two statistics be considered across fixed and adaptive settings to flag item pairs as exhibiting LID?

Based on the findings of previous studies focused on the performance of these statistics when applied to polytomous non-adaptive data (Lin, Kim, & Cohen, 2006) and dichotomous adaptive data (Pommerich & Ito, 2008), several hypotheses are posited. With respect to research question 1, the administration condition is expected to minimally

impact the performance of the Q_3 given the difference in test length. The X^2 may be impacted to a greater degree by the administration condition because of the sparseness in the data matrix and variation in sample size across item pairs in adaptive settings. Both the Q_3 and X^2 are expected to produce values larger when LID is induced in the responses than in the null condition. Regarding research question 2, in the null condition, it is anticipated that the Q_3 will yield values near those expected, and maintain a low false positive rate. It is hypothesized that the alternative calculation for expected cell counts described in Equation 7 will improve the performance of the X^2 given adaptive data with no LID, and implausibly large values will not be obtained. With respect to research question 3, the power of both the Q_3 and X^2 is expected to increase as the level of LID increases. Lastly, regarding research question 4, it is hypothesized that a lower cut-off for the Q_3 should be applied in adaptive settings as a result of the shorter test length. A unique cut-off may need to be applied for the X^2 as well given polytomous CAT data due to (1) the variation in within-item pair sample size and (2) the additional "noise" that may be introduced into the calculations when using a rough, empirical trait distribution to calculate expected cell counts as opposed to the theoretical χ^2 distribution.

Simulation Design

Data generation.

Item pool. In order to provide a realistic context for the simulation, the CAT modeled in the study utilized a bank of polytomous items in a domain outside educational measurement. The PROMIS Fatigue bank of 95 items accompanied by five-point response scales, calibrated using the GRM (Samejima, 1969), best met the desired criteria. As described on the PROMIS website, the Fatigue bank was one of several

included in Wave I and Wave II of the item bank testing. Wave I data were collected from approximately 20,000 individuals from the U.S. general population and multiple disease populations. These data were used to calibrate items for each domain, estimate profile scores for various sub-populations, create linking metrics to existing questionnaires, confirm factor structures of the domain, and conduct item and bank analyses. The item calibrations from Wave I were used in the current investigation. The specific items included in the Fatigue bank are reproduced in Appendix A. Their item characteristics are publically-available via the PROMIS Assessment Center (www.assessmentcenter.net).

Testlet structure. Testlets consisting of two items, or one LID item pair, were modeled. Although testlets larger in size could be examined, the current study considers the administration of items in adaptive settings in which individuals are only administered a subset of items in the bank. Without placing constraints on the selection algorithm, the likelihood of all items in a multi-item testlet being administered together would be low. Fixing the testlet size to pairs of items increases the number of administrations in which LID items are administered together. Furthermore, Pommerich and Ito (2008) considered only item pairs in their investigation.

With 95 items in the PROMIS Fatigue bank, there are 4,465 potential pairings of items, ignoring order. To make the scope of the simulation study more manageable, 48 of the possible pairs of items were strategically selected and the LID statistics for these pairs were tracked. Before item pairs were selected, the items were ordered from "easiest" to "hardest" based on the location of their category boundaries along the trait dimension. The items were then separated into three equal groups, such that the first third of the

items located towards the lower end of the scale were defined as "easy" items (E), the middle third were defined as "medium" difficulty items (M), and the last third located towards the upper end of the scale were defined as "hard" items (H). Pairs of items were deliberately selected to produce pairs of items similar in difficulty (i.e., E-E, M-M, H-H) and pairs of items of unmatched difficulty (i.e., E-M, E-H, M-H).

Other LID studies have considered classifications of item pairs based only on their local dependence classification. For example, Pommerich and Ito (2008) simply tracked LID item pairs and non-LID item pairs. Kim et al. (2007) tracked four types of pairs, including two items belonging to the same testlet (LID pairs), two items belonging to different testlets (LID-LID pairs), one item belonging to a testlet and one not belonging to a testlet (LID-LII pairs) and two items that do not belong to testlets (LII-LII pairs).

In this study, items are classified by both local dependency classification and difficulty classification as discussed above. The decision to track LID and LII pairs of different item difficulty combinations in this study was made because of the role item difficulty plays in CAT. The frequency with which an item pair is administered together in a CAT setting, as well as the trait range of the individuals' answering those items, is likely to vary across different combinations of item difficulty. Tracking LID and LII item pairs of different difficulty combinations ensures item pairs with a variety of within-item pair sample sizes are included.

Twenty-four pairs of items with similar, finely specified content were identified as LID pairs (see Appendix B). The two items in a pair essentially represent the same question rephrased in different terms, which is one potential cause of LID in the context

of the PROMIS Fatigue instrument. For example, the stem of the first item in one LID pair reads, “In the past 7 days, how often was it an effort to carry on a conversation because of your fatigue?” while the second reads, “In the past 7 days, how hard was it for you to carry on a conversation because of your fatigue?” In another LID pair, one stem reads, “In the past 7 days, how often did your fatigue limit you at work (include work at home)?” while the other reads, “In the past 7 days, how often were you less effective at work due to your fatigue (include work at home)?” Four pairs of LID items were selected for each of the six difficulty combinations.

Twenty-four pairs of items that were not as similar in content were selected as LII pairs (also displayed in Appendix B). For example, the stem of one of the items in the LII pair reads, “In the past 7 days, to what degree did your fatigue make it difficult to organize your thoughts while doing things at home?” while the stem of the second reads, “In the past 7 days, on how many days was your fatigue worse in the morning?” Four pairs of LII items were selected for each of the six difficulty combinations. Thus, 24 LII item pairs were selected to mirror the 24 LID item pairs in terms of their item characteristics. Furthermore, each item in the bank was assigned to only one of the 48 pairs tracked in the simulation study (with the exception of a single item because there were only 95 items in the bank, and not 48 x 2, or 96, items). With 48 of 95 items in the bank assigned to an LID pair, the LID percentage considered in this study is effectively 50%.

Generating model. Samejima’s (1969) GRM, as previously specified in Equation 1 and Equation 2, was used to generate individuals’ responses to the PROMIS Fatigue items under the assumption of LII. IRTGEN (Whittaker, Fitzpatrick, Williams, & Dodd,

2003) was used to generate the response data. IRTGEN is a collection of SAS macros that can generate known trait scores for simulees according to the random normal or random uniform distribution and item responses for simulees based on a number of dichotomous and polytomous IRT models, including the GRM. To use IRTGEN, a SAS dataset containing item parameters for a given IRT model is input and other characteristics of the data generation are specified, namely the name of the IRT model, number of items, and number of examinees. The output is a SAS dataset containing responses of each respondent to each item, along with his or her known trait score. In the current study, respondents' true θ values were sampled from a $N(0,1)$ distribution. The item parameter estimates obtained in Wave 1 testing were treated as known and used in conjunction with the known trait values to generate the response data under the GRM.

The method proposed by Chen and Thissen (1997) to model SLD was used to introduce a dependency structure to the responses. Given a pair of locally dependent items, an individual responds identically to the second item with a certain probability (π_{LID}) without the underlying processing implied by the IRT model. With probability $1-\pi_{LID}$, the response to the second item is generated using the IRT model, independent of the response to the first item. Expanding SLD Chen and Thissen's (1997) model to the polytomous case with five response options, the model determining the response to the second item in an LID pair is the following:

With probability $1-\pi_{LID}$*, the IRT model:*

$$\textit{Response to Item 2} = \begin{cases} 1, \textit{with } P(X_2 \mid \theta) \\ 2, \textit{with } P(X_2 \mid \theta) \\ 3, \textit{with } P(X_2 \mid \theta) \\ 4, \textit{with } P(X_2 \mid \theta) \\ 5, \textit{with } P(X_2 \mid \theta). \end{cases}$$

With probability π_{LID}*:* (10)

$$\textit{Response to Item 2} = \begin{cases} 1, \textit{if } X_1 = 1 \\ 2, \textit{if } X_1 = 2 \\ 3, \textit{if } X_1 = 3 \\ 4, \textit{if } X_1 = 4 \\ 5, \textit{if } X_1 = 5. \end{cases}$$

To further illustrate how this process works, consider the following example based on two items from the PROMIS Fatigue bank. Assume similarly-worded items FATIMP11 ("In the past 7 days, how often did your fatigue make you more forgetful?") and FATIMP44 ("In the past 7 days, to what degree did your fatigue make you more forgetful?") form an LID pair. FATIMP11 is an item of medium difficulty whereas FATIMP44 is more difficult. As predicted by the IRT model, a respondent with an above-average level of fatigue $(\theta_i = 1)$ would be most likely to score in the 3rd category for FATIMP11 ("Sometimes") and only in the 2nd category for FATIMP44 ("A little bit"). However, assume that when presented with FATIMP44 after selecting Category 3 for FATIMP11, with a certain probability (π_{LID}), the respondent simply offers the same response to FATIMP44, endorsing again the 3rd category instead of the 2nd.

Although alternative approaches could be selected to produce responses with a dependency structure (e.g., using a multidimensional model or random-effects testlet

model as the generating model), Chen and Thissen's (1997) model for surface LID was selected in order to make results most comparable to Pommerich and Ito's (2008) study investigating the properties of the LID statistics given dichotomous adaptive data. Furthermore, the SLD model is well-aligned with a carryover effect, which could be observed in a health-survey context where subjective traits are being measured.

Simulation Factors. Two independent variables were manipulated in the simulation study: administration condition and LID level. The simulation conditions are summarized in Table 5.

Table 5. Simulation conditions

Condition	Administration	LID Level
1	CONV	None
2	CONV	Low
3	CONV	Medium
4	CONV	High
5	CAT	None
6	CAT	Low
7	CAT	Medium
8	CAT	High

Administration conditions. Two types of administration were simulated, representing the key administration conditions considered by Pommerich and Ito (2008): (1) a conventional administration of all 95 items in the pool and (2) an administration of a 20-item instrument, where items are selected adaptively from the pool. A CAT of 20 items matches the maximum test length allowed in a PROMIS Fatigue CAT administration and closely mirrors the CAT length considered by Pommerich and Ito (2008) of 15 items. The first condition, referred to as CONV, offers essentially a fixed-form or full-bank context, much like that under which LID statistics have been studied

previously. Figure 5 depicts the concentration of responses in the data matrix with individuals ordered by trait level from lowest to highest across the top and items ordered by difficulty from easiest to hardest down the side. In the CONV administration, there is a consistent, heavy concentration of responses throughout the data matrix because all individuals respond to all items.

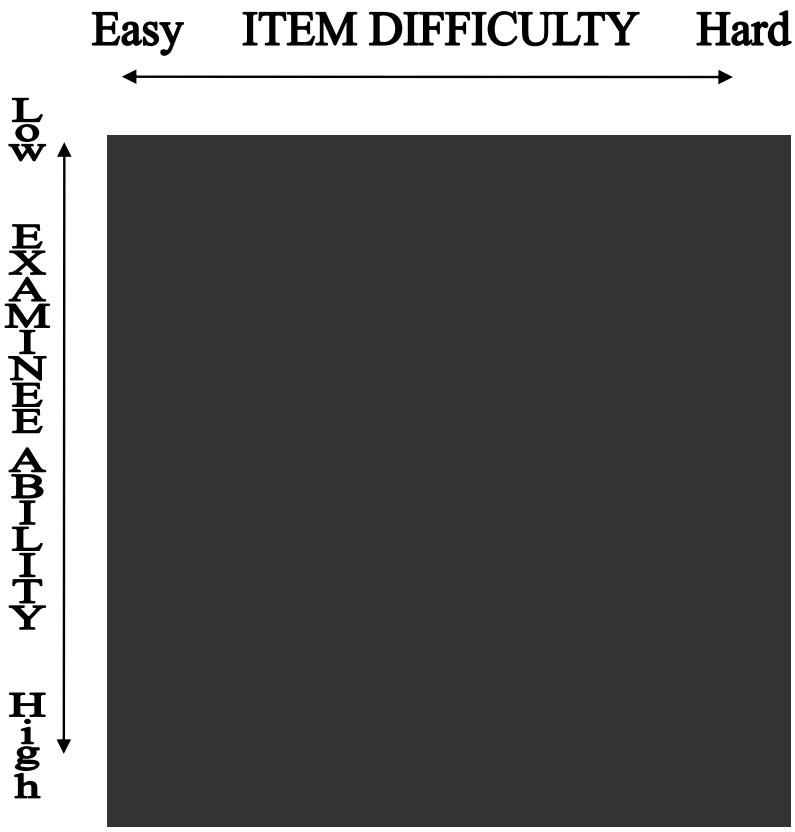


Figure 5. Concentration of responses in CONV administration

The second condition, referred to as CAT, allows an evaluation of LID statistics given a particular pattern of sparseness. Specifically, in the CAT condition, the range in

trait levels that occurs in the item samples was restricted. The concentration of data under the CAT condition is depicted in Figure 6. It sharply contrasts with Figure 5 because the concentration of responses is not consistent throughout the data matrix. The concentration is heavy in certain areas (e.g., the area corresponding to hard items and high-level individuals) while it is quite sparse in others (e.g., the area corresponding to easy items and high-level individuals).

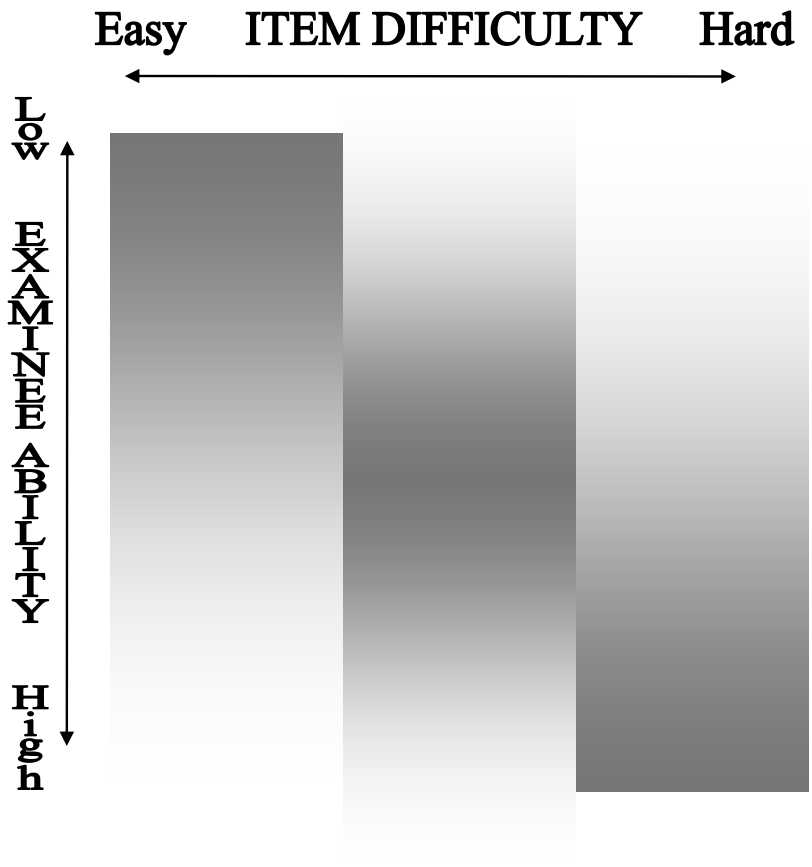


Figure 6. Concentration of responses in CAT administration

SIMPOLYCAT (Chen & Cook, 2009) was used to simulate the CAT administration. SIMPOLYCAT is a SAS macro that simulates CAT applications based on polytomous IRT models, including the GRM. SIMPOLYCAT allows users to evaluate

key features of a CAT, such as initial trait estimates, starting and stopping rules, and item exposure. Users can input item responses from an external file or prompt SIMPOLYCAT to generate data internally based on given item parameters. Users specify features of the CAT, such as rules for item selection, method for updating the trait estimates, and a stopping criterion. Two output files are produced, one SAS dataset containing item-level details of the CAT run, and the other containing summary results for each respondent.

In the current study, the initial trait estimate was set at zero for every respondent. A MII approach was used to select items and EAP with 20 quadrature points and a normal prior was selected as the trait estimation method, which is consistent with the SIMPOLYCAT defaults. Note that some features used in real CAT administrations of the PROMIS Fatigue instrument were altered to meet the purpose of the current study. Most critically, the stopping rule required a 20-item instrument.

Also, although in real applications of the Fatigue CAT, PROMIS uses the MPWI algorithm, a MII algorithm for item selection was utilized in the current research. Both MII and MPWI are available in SIMPOLYCAT. Pilot analyses, however, suggested MII as the more feasible option, with a computation time just under 6 hours per replication as compared to MPWI which failed to complete in a 36-hour test run, which is not uncommon for complex Bayesian estimation routines.

Further investigation suggested that selecting an MII algorithm as opposed to MPWI would have limited consequences for the current research. Choi and Swartz (2009) compared the performance of the MII approach to item selection to two newer selection methods purported to be superior, including the MPWI approach. Their simulation study considered CATs of different lengths (5, 10, and 20 items) under the

GRM. For all three procedures, EAP was used to update the trait estimate. The three approaches performed similarly, especially for medium and long tests (10 and 20 items, respectively), suggesting that the more complex and computing-intensive item selection procedures did not provide practical benefits over the standard MII when used in conjunction with the EAP. Veerkamp and Berger (1997) also found that the MII with EAP estimation produces similar gain to the weighted information functions.

In the current study, 20 quadrature points were used for the EAP trait estimation. Increasing the number of quadrature points typically leads to higher resolution or a reduction in the latent trait estimation error. Although some researchers have advocated the use of as many as 80 quadrature points (e.g., Bock & Mislevy, 1982), Weissman (2002) indicates that between 20 and 30 quadrature points are typically used in EAP estimation. For example, Gorin, Dodd, Fitzpatrick, and Shieh's (2005) study related to trait estimation in polytomous CAT utilized 20 quadrature points. Chen, Hou, and Dodd (1998) specifically considered the impact of number of quadrature points on the performance of the EAP estimation in a CAT based on the PCM. They found that increasing the number of quadrature points from 20 to 80 did not meaningfully increase the accuracy of the EAP estimation.

Thissen and Mislevy (1990, p.113) indicate that "even rough [trait] estimates are sufficient to select appropriately informative items" in CAT administrations and that "computational efficiency should play a greater role than fine points of precision and accuracy in determining the method of provisional proficiency estimation." Thus, MII item selection with EAP estimation based on 20 quadrature points is adequate for the task

at hand – namely, to produce a sufficiently representative CAT data matrix to which the LID statistics can be applied.

LID level. Pommerich and Ito (2008) considered only one level of LID, representing nearly perfect dependency among item pairs ($\pi_{LID} = 1.0$); they recommended future investigations consider different levels of LID. In the current study, three levels of LID were included, representing a low, medium, and high level of LID ($\pi_{LID} = 0.2$, $\pi_{LID} = 0.5,$ and $\pi_{LID} = 0.8,$ respectively); these LID levels are similar to LID conditions utilized by Chen and Thissen (1997). The manipulation of this probability will indicate the ability of the LID statistics to detect varying levels of LID. In addition, a null condition in which no LID is induced was included ($\pi_{LID} = 0.0$). In this case, there is no LID and all responses were generated according to the ordinary GRM.

Sample size. Total sample size was not a manipulated factor in the current study. Pommerich and Ito (2008) fixed the number of times a pair of items was administered at 2,000, requiring them to simulate data for an extremely large sample of individuals (60,000+) to ensure item pairs would be administered at least 2,000 in CAT administrations. However, in the current study, the number of individuals remains fixed at a total of 20,000, regardless of the administration condition. This sample size was selected because it reflects the actual sample size in Wave 1 of the PROMIS project. For both the CONV and CAT administration conditions, response data were generated for a total of 20,000 individuals. A fixed sample size of 20,000 individuals better represents that which occurs in practice; test developers will have resources to include a fixed number of participants in the pre-testing phase of a CAT, and will not be able to

drastically increase the sample size simply to ensure all possible item pairs are administered together with a certain frequency.

Replications. Previous studies investigating the performance of LID statistics have conducted as few as 10 or 50 replications (Levy, Mislevy, & Sinharay, 2009; Pommerich & Ito, 2008) and as many as 1,000 replications per condition (Kim et al., 2007). However, 100 replications were frequently selected by researchers interested in developing an empirical distribution for the LID statistic and 95% critical values under the null condition (Chen & Thissen, 1997; Chen & Wang, 2007; Lin, Kim, & Cohen, 2006); 100 replications also allows for intuitive calculations of power and false positive rates for other conditions using the cut-off obtained under the null condition. For these reasons, 100 replications of each condition were conducted.

Simulation summary. There were four main stages in the simulation study. First, item responses to all 95 PROMIS Fatigue items for 20,000 respondents were generated according to the GRM using IRTGEN. Second, a dependency structure was introduced to the data, such that responses to the second item in an LID pair were recoded to match the response to the first item with a given probability (π_{LID}). Third, the administration condition was simulated such that only responses to the 20 items selected for administration in a SIMPOLYCAT run were retained in the CAT condition; responses to all items were retained for the CONV condition. Lastly, the values of the LID statistics for the tracked item pairs were calculated. Specific components of each step are described in more detail in Appendix C. All simulation components were conducted using SAS® software, version 9.1 (Copyright © 2002-2003 by SAS Institute Inc., Cary, NC, USA).

Evaluation Criteria

Descriptively, the mean and standard deviation of the LID statistic values across the replications were examined for each administration condition. Although the Q_3 cannot be used for hypothesis testing, its expected value when LII holds is approximately -1/(*J*-1), where *J* is test length (Yen, 1993). Thus, the expected value is -1/(95-1) = -0.01 for the 95-item instrument and -1/(20-1) = -0.05 for the 20-item instrument in the null case. The X^2 is approximately distributed as a χ^2 distribution with degrees of freedom equal to $(K\text{-}1)^2$, or 16 given the 5 x 5 contingency tables of item responses. Thus, the critical value for $\alpha = 0.05$ under this distribution is 26.30. When no LID is induced, the LID values for item pairs should be close to the expected values in the null case and the standard deviation should not be large across the replications. When LID is induced, the average LID values for item pairs should be larger than their expected values. Statistics performing poorly under a given condition will not follow this pattern.

To help quantify the effects of the LID level and administration factors on the two statistics of interest, a series of means comparisons were conducted under a general linear modeling framework (GLM) for LID item pairs and LII item pairs, respectively. Along with expected cell means, the significance and size (as measured by eta-squared, η^2, and partial eta-squared, η_p^2) of the main and interaction effects were determined. Effect sizes for the η^2 index were described following Cohen's (1988, p. 283; 1992) conversion guidelines where .01 constitutes a small effect, .06 a medium effect, and .14 a large effect.

For the GLM analyses, the dependent variables were the Q_3 and X^2 values and the covariate was within-item pair sample size. Unlike LID level and administration

condition, within-item pair sample size was not systematically manipulated in the simulation. However, pilot analyses suggested that it had an impact on the magnitude of the LID statistics. Furthermore, sample size served as a proxy for the item pair difficulty combination, as pairs of a similar difficulty level were administered together more often than those of dissimilar difficulty levels. Thus, sample size was included as a covariate to statistically remove its influence on the values of the statistics so that the effects of the manipulated simulation factors on their variation can be more accurately assessed.

The within-subjects factor, LID level, involved four levels: none, low, medium, and high. The between-subjects factor, administration, involved two levels: CONV and CAT. Though the same item pairs were technically measured for both the CONV and CAT conditions, making the administration condition conceptually a second within-subjects factor, it was considered a between-subjects factor for the current analysis. This decision was made because only a subset of the tracked item pairs appeared in the CAT administration condition, resulting in missing data for the LID statistics for the remaining pairs. If administration had been treated as a second within-subjects factor, only those pairs appearing in both the CONV and CAT conditions would have been considered in the analysis, thereby drastically reducing the number of observations. Thus, administration condition was treated as a between-subjects factor under the premise that the item pairs across the CONV and CAT condition were "different but exchangeable." Interaction effects were also included where possible.

Results were first averaged across the 100 replications for each item pair such that each item pair contributed one value in its affiliated cell(s) of the design. That is, data were analyzed at the aggregate level to avoid the dependencies among values of the

statistics that would have arisen had values from all 100 replications been considered separately in the analysis.

Lastly, the empirical sampling distributions that were computed across the 100 replications also helped to determine whether different cut-off values should be applied for different administration conditions and within-item pair sample sizes in practice.

Real Data Application

A real-data component was included in this dissertation to (1) serve as a motivating example for the current research, (2) substantiate choices made in the simulation, and (3) suggest directions for future research. Via a collaboration agreement (see Appendix D), NIH has shared a pre-existing, de-identified dataset containing actual response data obtained in Wave I of the PROMIS project. An application was submitted to the University of Maryland, College Park Institutional Review Board (IRB; www.umresearch.umd.edu/IRB/) requesting a review of this research (see Appendix E). Given that the real data used in the current study are pre-existing and de-identified, the IRB approved the application, noting that it exempt from review (see Appendix F). Furthermore, in the time since the dataset was shared via the collaboration agreement, the de-identified Wave 1 data have been released into the public domain and researchers can request the public use datasets from the PROMIS Biostatistics and Data Management Core.

The de-identified dataset for the Fatigue domain, obtained by the PROMIS Statistical Coordinating Center from the polling firm that collected the data, was utilized. It contains response data for respondents in the full bank sample, with roughly 800 individuals responding to each of the 95 items included in the bank. Although response

data from CAT administrations were not available, post-hoc or "real-data" simulations were conducted using PROMIS data collected in fixed-form settings to mimic adaptive conditions. Essentially, instead of generating artificial response data for 20,000 simulees according to the Fatigue item parameters, the GRM, and randomly-selected trait values, the real responses for the 800 individuals were fed into SIMPOLYCAT. One CONV and one CAT administration was simulated post-hoc for each individual using the same procedures as in the pure simulation component. The data matrices and trait estimates resulting from each administration condition, along with the Fatigue item parameters, were then used to calculate the Q_3 and X^2 statistics for the same 48 item pairs that were tracked in the simulation.

Chapter 4: Results

Null Distribution of the LID Statistics

CONV administration. Table 6 presents descriptive statistical results for the LID statistics for all tracked item pairs under the CONV administration with no LID induced in the responses. Results are summarized over 100 replications with each replication containing 20,000 respondents. Because all individuals were administered all items in the CONV administration condition, the average within-item pair sample size (N) is equal to 20,000 and the standard deviation of the sample size is equal to zero (because all 100 replications included 20,000 individuals).

Descriptive statistics for X^2 under CONV condition. Regarding the LID statistics, Table 6 shows that the average X^2 values generally fell around 23 across both the LID and LII item pairs and that the standard deviations were small. Additionally, across the 48 tracked pairs, the 95th percentile of the empirical distribution fell near 36. These values are somewhat higher than the expected value and 0.05 critical value for a χ^2 distribution with 16 degrees of freedom (16.00 and 26.30, respectively).

Figure 7 shows the histogram of the empirical distribution of the X^2 values for all 48 pairs across 100 replications along with the χ^2 distribution curve with a mean of 23.[3] Again, these results suggest that the X^2 is not well approximated by a χ^2 distribution with 16 degrees of freedom. Instead, the sampling distribution is shifted to the right and is more similar to a χ^2 distribution with 23 degrees of freedom.

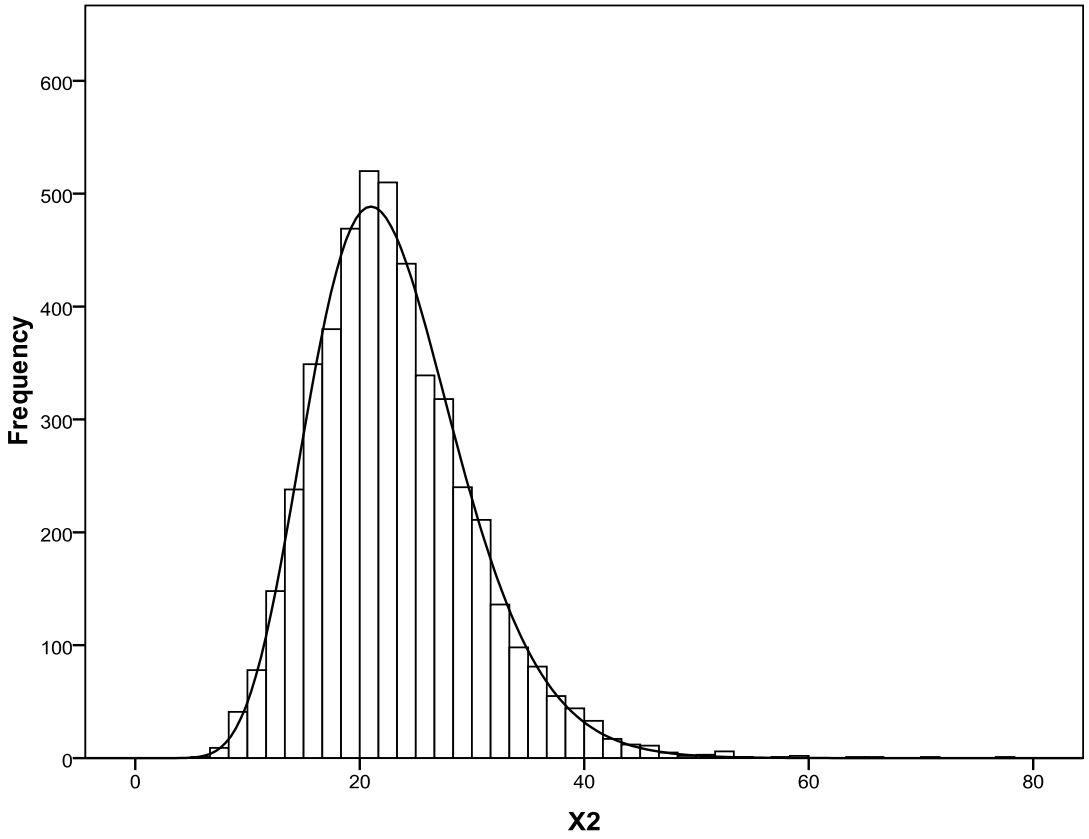


Figure 7. Combined X^2 results for tracked item pairs in the CONV condition, no LID

[3] Because such similar results were observed across the tracked pairs, combined results for all item pairs across the 100 replications are depicted in Figures 4 and 5, as is consistent with Chen and Thissen (1997) and Lin, Kim, and Cohen (2006).

Table 6. Results summary for tracked item pairs in the CONV condition with no LID

Pair	Within-Item Pair N		X^2							Q_3						
					Percentiles							Percentiles				
	Mean	Stdv.	Mean	Stdv.	1	5	50	95	99	Mean	Stdv.	1	5	50	95	99
LID Pairs																
1	20000	0.000	23.482	7.174	8.960	13.195	23.066	35.788	42.636	.021	0.008	.005	.007	.022	.034	.035
2	20000	0.000	22.453	6.548	6.077	13.597	21.260	33.514	45.277	.017	0.008	.003	.004	.017	.030	.039
3	20000	0.000	22.915	6.633	7.468	12.178	22.442	36.583	41.152	.017	0.008	-.001	.001	.018	.031	.037
4	20000	0.000	21.137	6.138	8.117	11.723	21.143	32.481	36.849	.016	0.007	-.003	.004	.017	.028	.035
5	20000	0.000	23.708	7.766	10.322	12.731	21.569	39.931	44.934	.015	0.006	-.005	.004	.015	.026	.029
6	20000	0.000	22.256	6.044	11.015	12.704	21.811	33.663	42.023	.013	0.009	-.010	-.004	.013	.030	.042
7	20000	0.000	23.079	6.009	12.105	12.931	22.807	34.438	39.072	.016	0.008	-.004	.003	.016	.028	.036
8	20000	0.000	21.654	6.697	8.682	11.102	21.426	34.685	39.862	.021	0.008	.000	.007	.022	.035	.039
9	20000	0.000	23.291	7.380	10.503	13.330	22.387	36.642	48.208	.013	0.009	-.013	-.001	.013	.028	.033
10	20000	0.000	24.432	7.570	9.091	13.102	23.650	39.199	45.597	.010	0.009	-.009	-.006	.011	.024	.026
11	20000	0.000	22.619	6.170	9.758	12.268	22.499	32.698	48.178	.021	0.007	.005	.008	.022	.033	.038
12	20000	0.000	22.641	6.276	10.274	13.162	22.573	33.714	40.720	.003	0.008	-.013	-.011	.003	.016	.028
13	20000	0.000	23.253	6.757	9.185	12.223	23.064	33.971	39.195	.014	0.008	-.003	.002	.013	.028	.036
14	20000	0.000	23.777	8.197	7.340	13.276	22.645	43.393	52.372	.019	0.008	-.002	.005	.018	.031	.044
15	20000	0.000	22.008	6.773	7.916	11.981	21.192	36.742	44.826	.011	0.006	-.005	.001	.011	.023	.029
16	20000	0.000	23.579	7.204	8.062	13.323	23.117	39.013	42.348	.024	0.007	.003	.011	.024	.036	.045
17	20000	0.000	22.617	5.469	10.396	12.751	21.974	32.862	37.212	.010	0.007	-.008	-.003	.010	.023	.027
18	20000	0.000	24.573	6.777	12.448	15.493	23.942	37.855	51.966	.004	0.006	-.012	-.006	.004	.015	.018
19	20000	0.000	23.760	10.507	10.680	12.090	20.918	52.316	65.416	.020	0.008	.002	.005	.020	.033	.040
20	20000	0.000	21.621	6.473	7.508	12.606	21.029	33.169	42.721	.019	0.007	-.003	.006	.020	.030	.035
21	20000	0.000	22.379	7.047	10.377	11.632	20.560	37.021	42.685	.014	0.007	-.007	.002	.014	.026	.031
22	20000	0.000	24.211	9.244	9.932	14.443	22.631	35.431	76.864	.019	0.008	-.002	.008	.019	.031	.035
23	20000	0.000	22.216	7.324	7.784	12.935	20.495	38.000	46.584	.023	0.007	.003	.011	.023	.034	.037
24	20000	0.000	23.138	5.961	8.935	13.926	22.811	33.386	38.996	.010	0.008	-.015	-.004	.010	.023	.027

LII Pairs																
25	20000	0.000	25.041	10.020	8.995	13.032	22.426	47.874	64.130	.009	0.007	-.008	-.003	.009	.020	.025
26	20000	0.000	24.469	6.920	9.802	14.789	23.295	39.085	40.187	.009	0.008	-.012	-.003	.009	.024	.030
27	20000	0.000	23.521	7.200	11.096	12.298	22.727	38.825	44.741	.005	0.007	-.012	-.007	.004	.019	.024
28	20000	0.000	23.503	6.746	9.471	13.305	23.756	34.590	45.997	.016	0.007	-.002	.005	.016	.027	.038
29	20000	0.000	23.533	7.462	9.533	12.416	22.428	37.544	46.092	.022	0.007	.004	.010	.022	.034	.038
30	20000	0.000	21.971	6.321	10.139	13.426	20.726	34.113	37.175	.020	0.007	.002	.007	.020	.031	.035
31	20000	0.000	23.955	6.705	8.202	14.416	22.744	36.287	49.209	.011	0.008	-.006	.000	.012	.024	.037
32	20000	0.000	23.031	6.517	8.792	12.468	22.103	35.175	43.607	.018	0.008	.001	.006	.018	.033	.037
33	20000	0.000	22.610	6.562	8.553	11.771	22.661	34.040	40.739	.017	0.008	-.004	.004	.017	.030	.043
34	20000	0.000	22.216	5.454	12.154	14.092	21.657	32.634	39.468	.017	0.007	-.006	.005	.017	.030	.034
35	20000	0.000	22.190	7.064	9.575	11.770	21.147	38.956	40.778	.007	0.008	-.016	-.007	.007	.022	.029
36	20000	0.000	23.937	7.162	9.601	13.484	24.030	37.125	42.304	.016	0.009	-.001	.002	.016	.032	.043
37	20000	0.000	23.115	7.349	7.915	11.725	22.018	38.518	43.954	.009	0.007	-.012	-.002	.009	.020	.030
38	20000	0.000	23.856	7.816	10.408	13.334	22.928	36.509	59.158	.013	0.006	-.005	.003	.013	.023	.033
39	20000	0.000	21.639	6.574	9.840	12.138	20.856	34.949	39.163	.019	0.008	.002	.008	.018	.032	.040
40	20000	0.000	23.159	7.692	10.806	13.143	21.580	38.160	47.163	.017	0.008	-.006	.004	.017	.028	.035
41	20000	0.000	23.294	6.112	11.218	14.421	22.671	35.648	40.534	.012	0.007	-.005	.000	.011	.025	.027
42	20000	0.000	22.651	6.195	11.601	13.806	22.316	35.426	42.526	.013	0.008	-.005	-.001	.013	.026	.036
43	20000	0.000	22.102	6.825	8.469	11.936	22.139	34.763	42.520	.014	0.007	-.008	.001	.014	.026	.029
44	20000	0.000	21.913	7.407	9.029	10.513	21.018	35.203	46.330	.010	0.008	-.010	-.003	.009	.024	.031
45	20000	0.000	22.622	5.554	9.725	13.634	22.131	31.823	36.833	.016	0.007	-.004	.002	.016	.028	.032
46	20000	0.000	23.416	7.623	8.405	12.560	21.569	37.661	43.618	.011	0.008	-.004	-.001	.011	.023	.026
47	20000	0.000	23.190	7.499	8.931	12.352	22.793	36.849	46.297	.024	0.007	.005	.013	.024	.038	.047
48	20000	0.000	22.798	7.235	10.735	11.985	21.392	36.746	51.400	.009	0.008	-.015	-.005	.009	.021	.027

In their study based on polytomous non-adaptive data, Lin, Kim, and Cohen (2006) used the theoretical population distribution (i.e., $N(0,1)$) to produce expected cell counts. For an 80-item test containing items with 5 response categories, the empirical X^2 distribution they obtained under the null condition closely followed a χ^2 distribution with 16 degrees of freedom. Therefore, it is possible that the X^2 inflation observed in the current study can be attributed to the substitution of an approximated trait distribution in the calculation of the statistic. Another possible explanation is that the χ^2 is known to be influenced by sample size (Wang, Fan, & Wilson, 1996). Lin, Kim, and Cohen (2006) only considered a sample size of 1,000 in their simulation, as opposed to the current simulation which utilized a much larger sample of 20,000 individuals.

Descriptive statistics for Q_3 under CONV condition. For the CONV administration with locally independent data, the average Q_3 values displayed in Table 6 were near zero across the tracked pairs with small standard deviations around these values; the mean values were reasonably close to the expected value of -0.01 for a 95-item instrument. The 95th percentile of the empirical distribution generally fell between 0.02 and 0.03, which was substantially lower than the 0.20 cut-off value typically used in practice to screen item pairs for LID. This result is consistent with Chen and Thissen (1997) and Kim et al. (2007), who also found that optimal cut-points obtained from simulated data tended to be about one tenth the magnitude of the recommended value.

Figure 8 shows the histogram of the empirical sampling distribution of the Q_3 values for all 48 pairs across 100 replications, along with a superimposed Gaussian curve. The sampling distribution is bell shaped, approximately $N(0.01, 0.01)$. This distribution is very similar to the $N(-0.01, 0.03)$ distribution observed by Lin, Kim, and Cohen (2006)

given a non-adaptive test containing 80 items with five response categories. Following Yen's (1993) formulation of the expected value of the Q_3 which is influenced by test length, one would expect a mean that is slightly more negative given an 80-item test as opposed to a 95-item test. The similarity of the current findings despite the drastically different sample sizes modeled in the two simulations also supports Pommerich and Ito's (2008) assertion that the Q_3 is not notably affected by sample size.

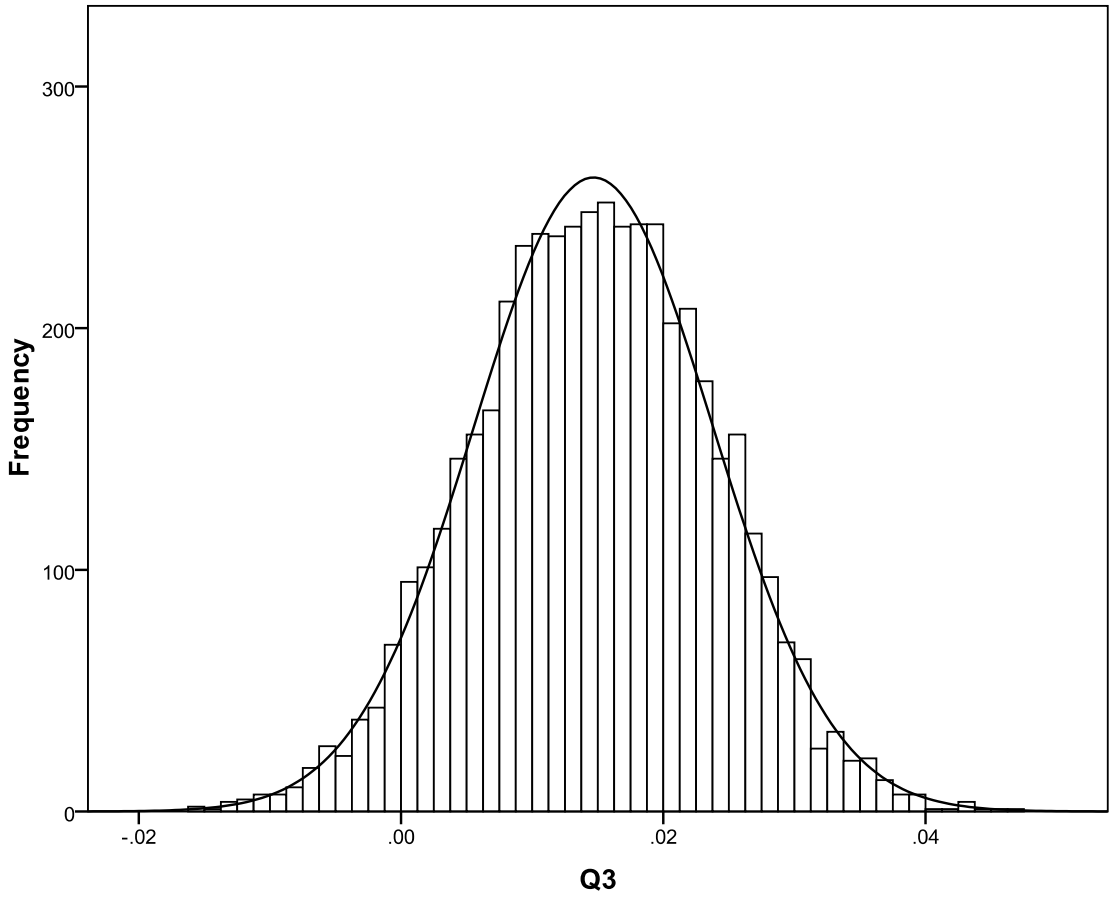


Figure 8. Combined Q_3 results for tracked item pairs in the CONV condition, no LID

CAT administration. Before examining the null distributions resulting from the CAT condition, it is important to note the varying frequency with which items and item pairs were selected for administration.

Sample sizes per item pair. First, the use of items in the Fatigue bank in the simulated CAT administrations was uneven. About a third of the items were selected fewer than 100 times in 20,000 administrations and about a fifth were never selected for administration. The most frequently administered item (FATIMP3) was administered to all 20,000 individuals as it was the most informative item at the respondents' initial trait estimate of zero.

Table 7 shows that, in a given replication, items with higher discrimination parameters were selected for administration more frequently than items with lower discrimination parameters. On average, items with discrimination parameters greater than 3.75 were included in more than half the simulated CAT administrations, whereas items with discrimination parameters less than or equal to 2.50 were only selected for a small fraction of respondents. This finding is not surprising given that the simulation used a MII item selection method. Because information-based selection rules select the most informative items at one's currently estimated trait level, items with large discrimination parameters are more likely to be selected leading to extremely skewed item exposure rates (Chang & Ying, 1999).

Table 7. PROMIS Fatigue item selection in 20,000 simulated CATs, no LID

Item Discrimination	Number of Fatigue Items	Average Number of Administrations
>3.75	20	13720.050
3.51-3.75	15	5396.133
3.01-3.50	26	889.885
2.51-3.00	19	651.842
≤2.50	15	609.000

Consistent with the observed variation in the sample size per item, the within-item pair sample size also varied drastically under the CAT condition. As shown in Table 8, only 21 of the 48 tracked item pairs (11 LID pairs and 10 LII pairs) had a non-zero within-item pair sample size.

Variation in within-item pair sample size was also observed in Pommerich and Ito's (2008) study based on dichotomous adaptive data; the authors noted that in one replication in which 62,000 sessions were simulated, more than a third of the possible item pairs were never administered together and over a quarter were administered together for fewer than 100 examinees.

The drastic variation in the within-item pair sample size seemed driven by two factors: the quality of the items belonging to the pair (as measured by item discrimination) and the similarity of the pair's difficulty. Pairs similar in difficulty level (e.g., E-E) tended to be administered together more often than those of dissimilar difficulty (e.g., E-H). For example, only one of the eight E-H item pairs yielded observations (likely because it contained the most frequently administered item), whereas all eight E-E pairs yielded observations as did half the M-M pairs. This finding is not surprising given the nature of adaptive testing; after the first few items are administered, the sequential administration of subsequent items typically stays within a similar difficulty range for a given individual. Thus, it would not be expected that a large proportion of the CAT administrations included both easy and hard items.

Table 8. Results summary for tracked item pairs with a non-zero sample size in the CAT condition with no LID

			X^2							Q_3						
	Within-Item Pair N				Percentiles							Percentiles				
Pair	Mean	Stdv.	Mean	Stdv.	1	5	50	95	99	Mean	Stdv.	1	5	50	95	99
LID Pairs																
1	2780.490	48.091	21.752	18.160	5.253	8.436	18.058	49.819	145.098	-.038	0.019	-.076	-.069	-.038	-.003	.008
3	99.420	10.023	14.285	16.362	2.893	3.857	9.009	57.180	106.843	-.054	0.106	-.309	-.238	-.067	.113	.223
6	96.210	9.428	36.399	112.826	4.777	6.702	16.636	56.023	831.190	-.052	0.105	-.281	-.226	-.056	.107	.245
10	1.870	1.509	5.902	24.225	0.093	0.115	0.325	20.929	.	.713	0.666	-1.000	-.997	1.000	1.000	.
14	7018.460	60.143	47.593	13.803	26.769	28.523	45.887	74.967	99.117	-.018	0.012	-.043	-.040	-.020	.003	.009
15	3264.590	49.692	28.705	17.064	10.030	11.565	24.792	64.461	101.130	-.050	0.019	-.100	-.079	-.050	-.015	.004
16	12927.460	64.364	35.192	10.722	16.859	20.879	33.615	50.579	96.683	-.029	0.009	-.049	-.043	-.030	-.014	-.009
18	1702.050	37.317	23.697	19.568	5.248	11.098	19.712	65.253	165.200	-.049	0.026	-.111	-.094	-.048	-.011	.015
19	12492.920	60.304	71.506	18.671	38.828	44.380	68.486	105.584	148.639	-.029	0.009	-.052	-.044	-.030	-.014	-.008
23	8216.660	77.399	34.298	9.856	16.774	17.690	33.904	53.440	55.679	-.036	0.012	-.066	-.053	-.037	-.018	-.007
24	21.500	4.602	18.057	33.905	2.421	2.874	9.700	70.582	294.551	-.048	0.228	-.527	-.414	-.059	.348	.448
LII Pairs																
25	2424.140	46.613	27.147	33.145	3.561	9.989	22.484	50.892	327.409	-.076	0.023	-.126	-.117	-.075	-.035	-.020
26	2610.910	49.939	20.830	17.041	4.963	8.564	15.024	46.836	144.950	-.064	0.022	-.131	-.098	-.065	-.028	-.008
27	935.940	27.916	15.312	12.913	3.579	4.134	12.891	35.240	78.305	-.038	0.033	-.100	-.089	-.042	.027	.042
28	4508.940	53.941	23.875	16.070	8.029	10.276	21.056	40.931	153.752	-.047	0.017	-.089	-.076	-.046	-.020	-.010
29	86.410	8.145	20.308	49.396	3.019	3.235	10.905	34.858	328.923	-.079	0.108	-.340	-.263	-.079	.081	.197
30	2339.540	43.553	30.428	18.966	12.146	15.325	25.053	56.599	144.181	-.039	0.017	-.087	-.068	-.038	-.007	.003
36	39.500	6.422	17.445	39.602	2.166	2.998	10.320	39.089	386.181	-.052	0.167	-.408	-.314	-.059	.249	.366
38	6.300	2.572	8.911	26.940	0.425	0.899	3.359	16.486	204.179	-.054	0.559	-1.000	-.999	-.196	1.000	.
39	37.450	6.199	18.320	94.297	1.590	2.370	6.393	29.885	936.889	-.037	0.178	-.331	-.295	-.027	.308	.485
47	12386.480	63.193	84.020	15.757	53.419	59.591	82.233	111.346	125.786	-.032	0.008	-.050	-.045	-.032	-.020	-.007

Respondents' trait distributions for item pairs. Figure 9 depicts the estimated trait distributions for individuals responding to a pair of easy items (Pair 18) and a pair of hard items (Pair 6) in one replication of 20,000 simulated CAT administrations with no LID, along with a superimposed standard normal distribution. The 1,753 individuals who responded to the E-E item pair had trait estimates ranging from about -3 to -1 on the θ scale, with a mean of -1.80. In contrast, the 89 individuals responding to the H-H item pair had trait estimates ranging from about 2 to 4, with a mean of 2.80. Both of these estimated trait distributions sharply contrast with the $N(0,1)$ population trait distribution which is typically assumed in calculations of the LID statistics (Chen & Thissen, 1997; Lin, Kim, & Cohen, 2006).

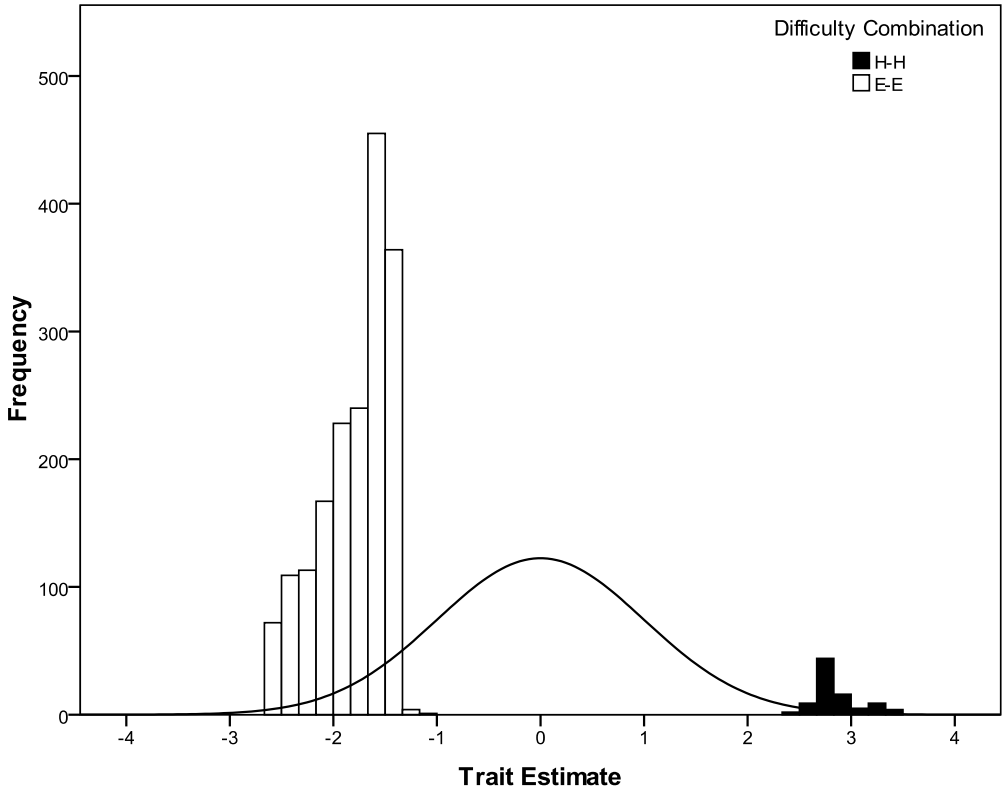


Figure 9. Estimated trait distributions for those responding to an E-E and H-H pair

In non-adaptive settings, it may be adequate to assume a standard normal population trait distribution in calculations because individuals across the entire trait range are responding to all items. In adaptive settings, however, the assumed distribution should reflect a restricted trait range. As shown in Figure 9, responses to a given item pair were only provided by individuals representing a particular subset of the population. Furthermore, the particular population subset represented will differ across item pairs depending on the difficulty level(s) of the items.

Descriptive statistics for X^2 under CAT condition. Table 8 also presents descriptive results for the LID statistics for all tracked item pairs under the CAT administration with no LID induced in the responses. Again, results are summarized over 100 replications with each replication containing 20,000 respondents. Given LII, across the item pairs that had a non-zero sample size in the CAT administration, the average X^2 values fell around 29 and the 95th percentiles of the empirical distributions fell near 54. Not only was the average X^2 value somewhat larger in the CAT condition as compared to the CONV condition (29 vs. 23, respectively), the standard deviations in the CAT condition were notably larger than those observed in the CONV condition. Although there does appear to be a slight inflation of the X^2 values in the CAT administration condition, the "implausibly high" values observed by Pommerich and Ito (2008) were not observed in this analysis.

The fact that unrealistically large X^2 values were not observed in the current study is likely due to the use of a restricted trait range in the calculation of the statistic, which was based on only the subset of individuals responding to the particular item pair. Since the X^2 statistic summarizes differences between expected and observed cell counts, if

expected cell counts are derived based on an inaccurate assumption (i.e., the theoretical population distribution), X^2 values will be large.

The results in Table 8 also suggest an association between within-item pair sample size and the magnitude of the X^2. That is, the means of the sampling distributions for item pairs with a larger sample size, relatively speaking, tended to be greater than the means for item pairs with a moderate or low sample size. Figure 10 presents a scatterplot of within-item pair sample size and the obtained X^2 values (first averaged across 100 replications of the CAT condition) for the 21 pairs with non-zero sample sizes. Indeed, the X^2 values were highly correlated with sample size (r = .82).

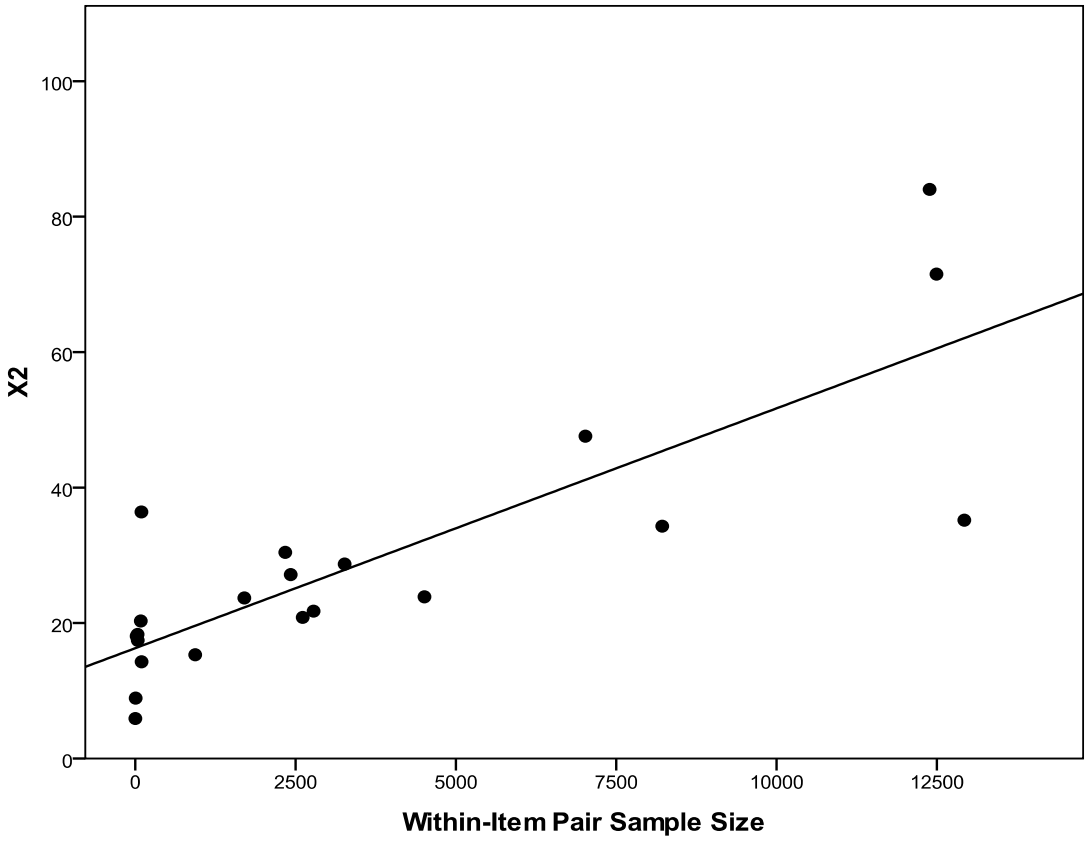


Figure 10. Scatterplot of sample size and X^2 values for pairs in the CAT condition

Figures 11 through 13 depict the empirical X^2 sampling distributions associated with three item pairs selected to represent small, moderate, and large underlying sample sizes. Pair 3 is an example of an item pair with a smaller sample size having been administered to about 100 out of 20,000 individuals on average; it yielded an average X^2 value of 14. In contrast, Pair 19 is an example of a pair with a large sample size having been administered to about 12,000 of 20,000 individuals; it yielded an average X^2 value of almost 72. Lastly, Pair 28 represents pairs with a moderate sample size having been administered to about 4,500 of 20,000 individuals; it yielded an average X^2 value of 24, which is closer to the average values observed in the CONV condition. These results may again be reflective of the sample-size dependencies of the χ^2 (Wang, Fan, & Wilson, 1996).

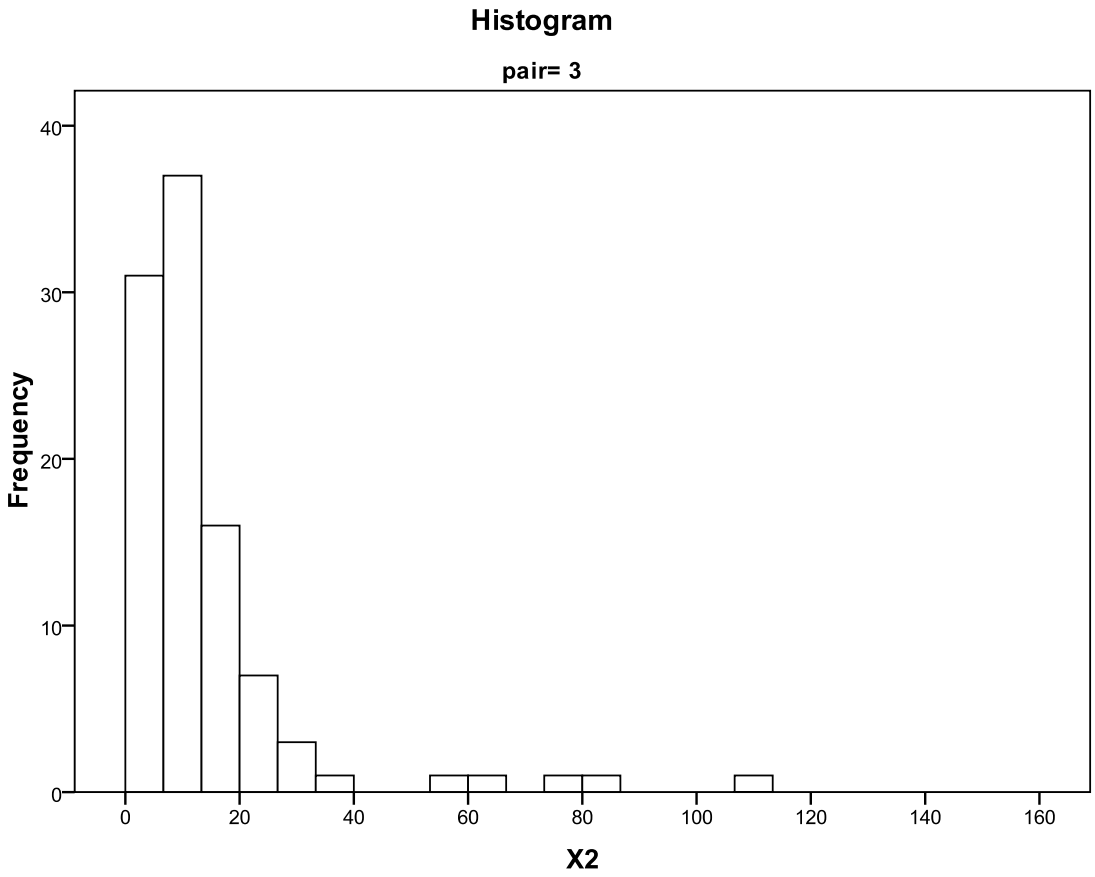


Figure 11. X^2 histogram for pair with N_{ave} = 99 in CAT condition, no LID

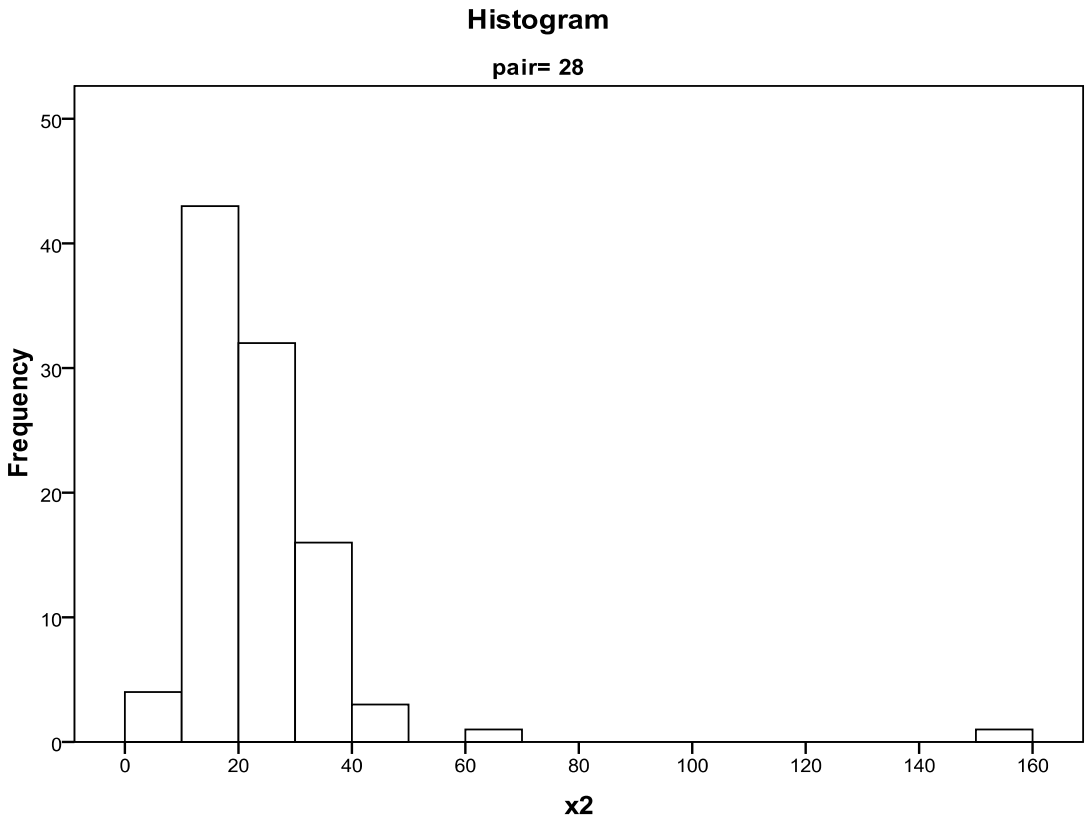


Figure 12. X^2 histogram for pair with N_{ave} = 4,508 in CAT condition, no LID

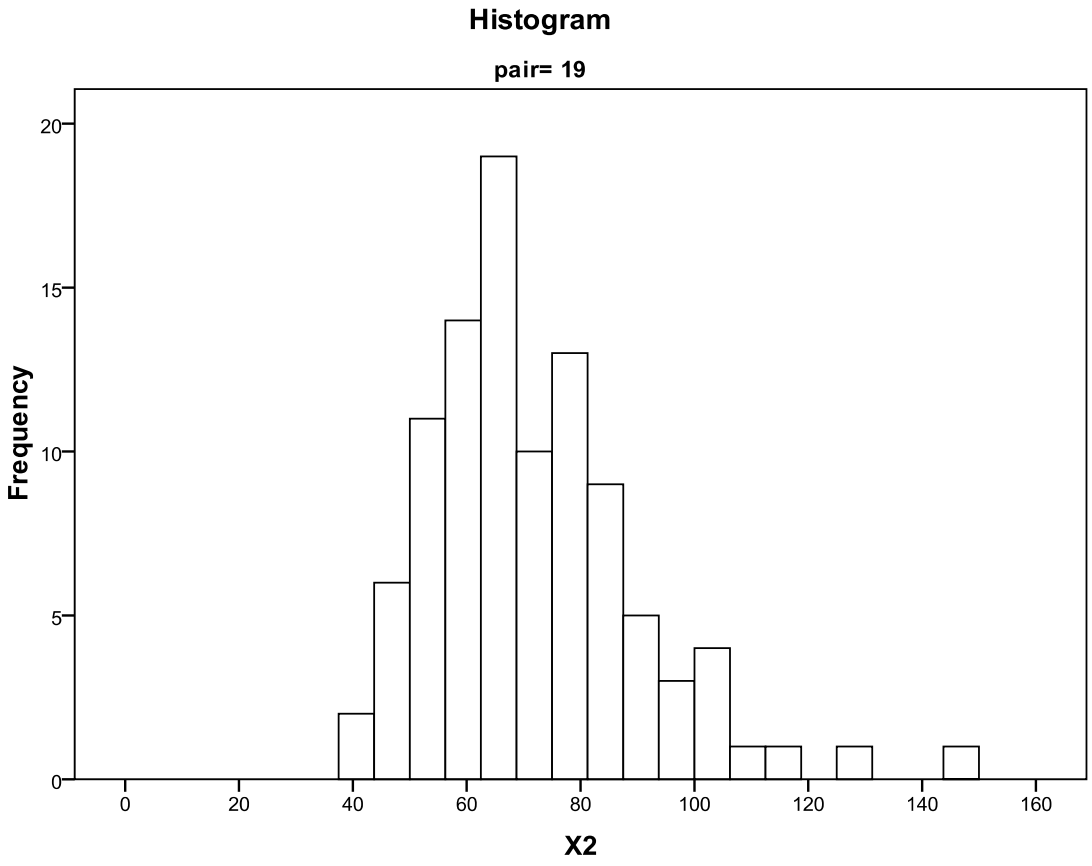


Figure 13. X^2 histogram for pair with N_{ave} = 12,493 in CAT condition, no LID

Descriptive statistics for Q_3 under CAT condition. Regarding the Q_3 statistic, for the CAT administration with locally independent data, the average values displayed in Table 8 were near zero and slightly negative across the pairs with non-zero sample sizes. These results are reasonably close to the expected value of -0.05 for a 20-item instrument. Furthermore, the results are consistent with Pommerich and Ito's (2008) based on dichotomous adaptive data; they obtained average Q_3 values ranging from -0.08 to -0.05 for LII pairs using dichotomous data from a 15-item CAT. For many of the item pairs in the current study, the 95th percentile of the empirical distribution generally fell around

0.03, which was again substantially lower than the 0.20 cut-off value typically applied in practice.

The results presented in Table 8 also suggest that the Q_3 is somewhat influenced by the within-item pair sample size in CAT, though not in the same way as the X^2. As seen in Table 8, although the means of the Q_3 sampling distributions remained relatively consistent across the tracked item pairs (with the exception of Pair 10), larger standard deviations were observed for pairs with smaller within-item pair sample sizes. Figure 14 presents a scatterplot of within-item pair sample size and the obtained Q_3 values (averaged across 100 replications of the CAT condition) for the 21 pairs with non-zero sample sizes. There is an obvious outlier in Figure 14: Pair 10. Pair 10 had the smallest within-item pair sample size of the tracked pairs having been administered together on average in only 2 of 20,000 simulated CAT administrations. As a result, the Q_3 values for this pair bounced between -1 and 1 across the 100 replications. Excluding Pair 10 from the analysis to eliminate an outlier effect, the Q_3 values are moderately correlated with sample size (r = .61).

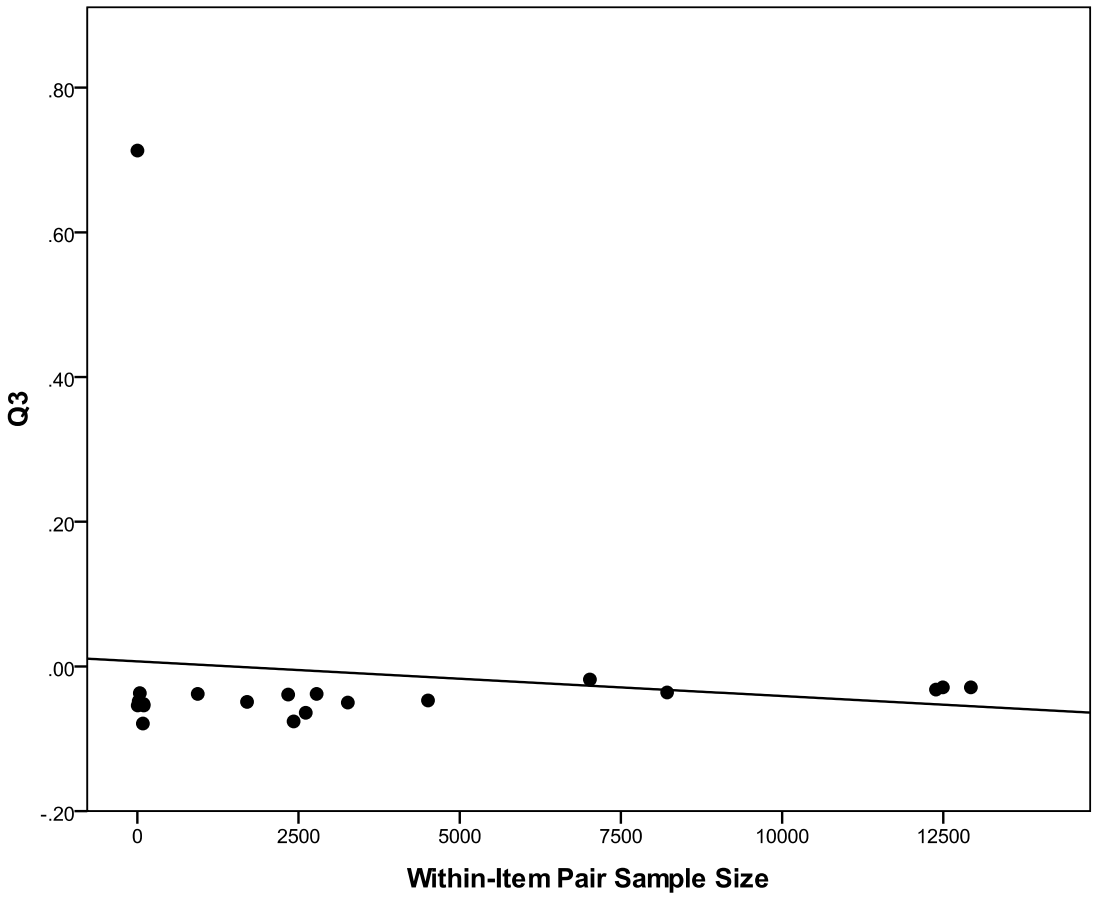


Figure 14. Scatterplot of sample size and Q_3 values for pairs in the CAT condition

Pair 24 is another example of an item pair with a relatively small within-item pair sample size having been administered, on average, to roughly 20 of 20,000 individuals in CAT; it yielded an average Q_3 value of -.05. Pair 26, however, represents a pair with a moderate within-item pair sample size of about 2,600; it yielded an average Q_3 of -.06. Figures 15 and 16 depict the empirical Q_3 sampling distribution associated with Pairs 24 and 26, respectively. Although the means across the distributions were similar, the shape for Pair 26 in Figure 16 more closely resembled a normal distribution. Furthermore, the distribution associated with Pair 24 in Figure 15 was more spread out as indicated by the larger standard deviation (0.23 vs. 0.02 for Pair 26) with observed values ranging from

nearly -.50 to .50 across the 100 replications. In other words, within-item pair sample size impacts the shape and spread of the empirical Q_3 sampling distribution more so than its location / mean value.

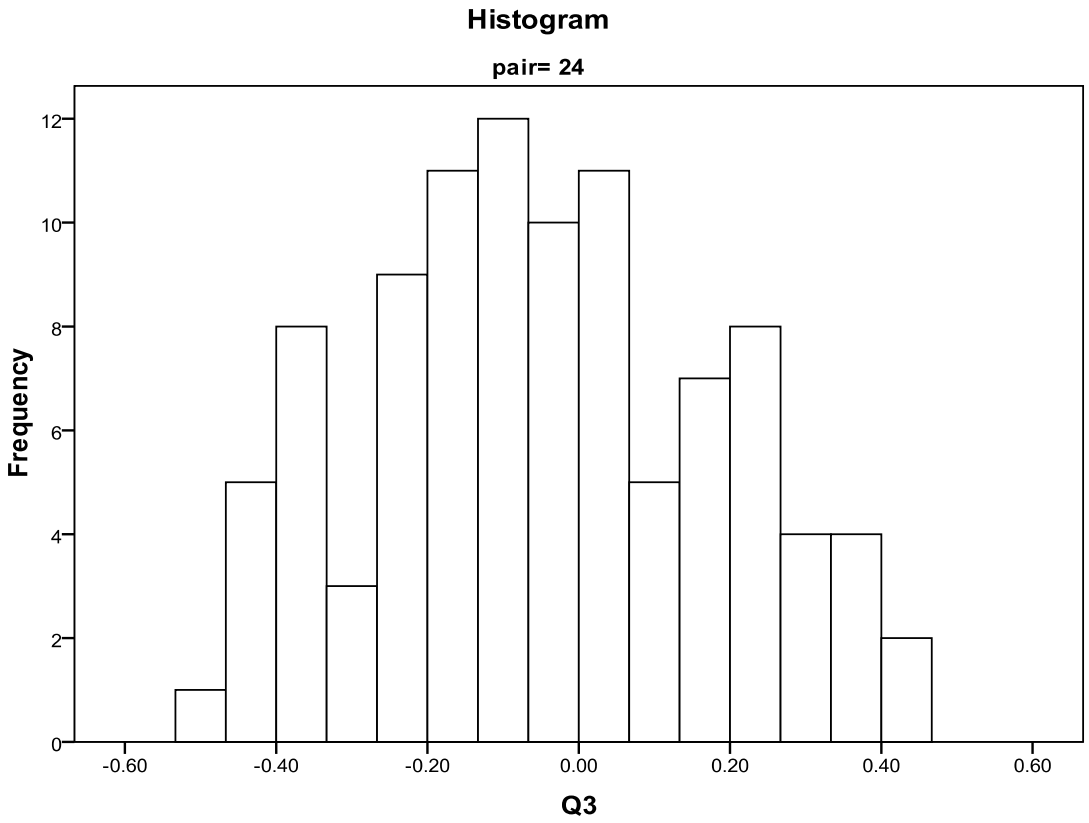


Figure 15. Q_3 histogram for pair with N_{ave} = 22 in CAT condition, no LID

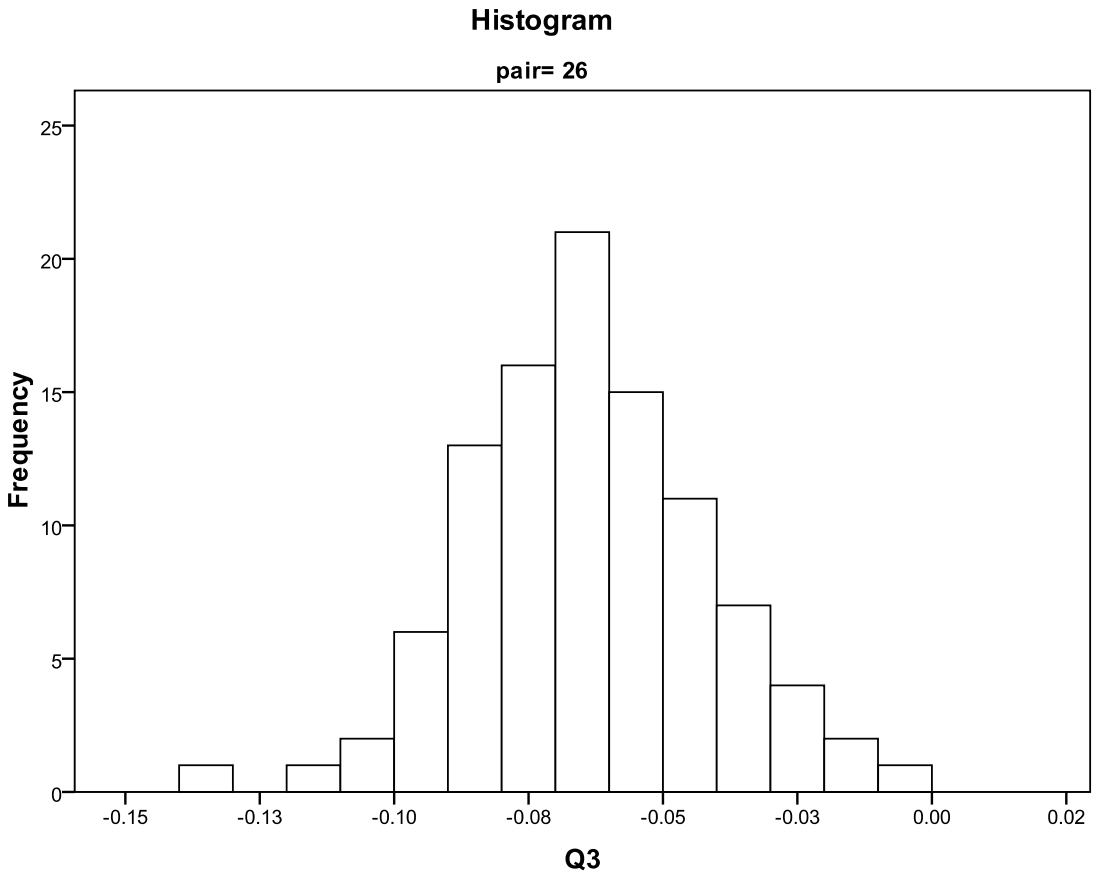


Figure 16. Q_3 histogram for pair with N_{ave} = 2,611 in CAT condition, no LID

Pommerich and Ito (2008) observed similar trends in their simulation with dichotomous data. For pairs with a sample size of less than 100, the minimum and maximum Q_3 statistics were -1 and 1, respectively, and the standard deviation of the Q_3 across the pairs was larger than it was for groups of items with larger sample sizes. The smallest range of minimum and maximum values was observed for pairs with sample sizes greater than or equal to 2,000.

LID Conditions

Descriptive results. In the following, results are presented by administration condition in three subsections, (1) sample size per item pair, (2) descriptive statistics for the X^2 statistic, and (3) descriptive statistics for the Q_3 statistic.

Sample size per item pair for CONV condition. Table 9 shows the results for all tracked item pairs under the CONV administration with varying levels of LID induced in the response, summarized over 100 replications of each condition. Again, in the CONV condition, all individuals were administered all items so the average within-item pair sample size for every item pair was equal to 20,000 and the standard deviation of the sample size was equal to zero.

Descriptive statistics for X^2 in CONV condition. Regarding the LID statistics, Table 9 shows that when a low level of dependency was induced in the responses, the average X^2 values fell between 600 and 3,000 across the 24 LID pairs; these average values were substantially higher than those around 23 observed in the null condition when no LID was induced in the responses. The average values for the LII pairs, however, fell around 26, just slightly inflated as compared to their values in the null condition. When a medium level of dependency was induced in the responses, the average X^2 values associated with the LID pairs jumped to about 6,000, while the average values for the LII pairs fell around 40. Lastly, when a high level of LID was induced, average X^2 values for the LID pairs ranged from 9,000 all the way to almost 50,000; X^2 values for the LII pairs generally fell around 60. For all LID conditions, the standard deviations of the values of the X^2 statistic for the LII pairs remained small while the standard deviations for the LID pairs increased with the level of LID.

Table 9. Results summary for tracked item pairs in the CONV condition with varying LID

Pair	Q_3						X^2					
	Low LID		Med LID		High LID		Low LID		Med LID		High LID	
	Mean	Stdv.	Mean	Stdv.	Mean	Stdv.	Mean	Stdv.	Mean	Stdv.	Mean	Stdv.
							LID Pairs					
1	.224	0.009	.521	0.008	.807	0.006	784.930	51.925	4847.747	114.755	12495.106	149.118
2	.232	0.009	.535	0.008	.818	0.005	788.333	60.328	4951.188	129.580	12802.346	170.200
3	.218	0.008	.503	0.009	.789	0.006	721.626	48.040	4499.471	141.761	11629.818	147.391
4	.221	0.010	.516	0.010	.807	0.006	774.272	55.749	4821.662	144.296	12468.085	157.538
5	.194	0.009	.453	0.008	.733	0.006	861.135	61.462	5350.729	122.957	13810.136	170.734
6	.206	0.010	.491	0.010	.783	0.007	595.833	49.445	3647.567	114.224	9415.539	187.980
7	.202	0.008	.491	0.008	.789	0.006	789.862	54.668	4933.113	123.699	12699.769	168.631
8	.219	0.010	.516	0.008	.802	0.006	749.195	58.009	4720.692	133.293	12205.579	194.966
9	.231	0.009	.529	0.009	.809	0.005	722.711	53.121	4516.404	137.552	11707.097	198.560
10	.219	0.009	.517	0.009	.803	0.006	939.643	60.313	5819.743	162.046	15031.717	216.687
11	.204	0.009	.486	0.009	.786	0.006	696.535	50.534	4317.008	135.167	11173.708	153.457
12	.144	0.008	.359	0.008	.634	0.007	3110.073	135.873	19201.458	363.551	48409.761	449.998
13	.204	0.008	.496	0.008	.791	0.006	891.995	60.909	5586.141	105.503	14384.656	169.745
14	.207	0.008	.491	0.009	.790	0.006	669.935	46.143	4124.218	127.207	10640.169	127.240
15	.205	0.008	.498	0.008	.798	0.007	1047.674	63.367	6514.154	152.680	16693.040	170.979
16	.208	0.009	.484	0.009	.773	0.006	691.565	44.345	4287.581	116.633	10978.031	128.066
17	.171	0.009	.433	0.009	.736	0.006	953.954	61.536	6000.458	146.979	15381.825	189.355
18	.211	0.008	.515	0.008	.808	0.005	1196.329	72.739	7407.950	157.137	18975.314	206.394
19	.199	0.009	.479	0.008	.779	0.006	612.244	48.161	3743.559	105.551	9582.506	120.883
20	.220	0.009	.514	0.008	.803	0.006	676.527	51.333	4236.156	114.627	10964.564	140.778
21	.168	0.009	.396	0.007	.650	0.006	1248.825	74.790	7818.833	151.836	19938.668	233.472
22	.215	0.008	.465	0.008	.732	0.006	1237.374	74.639	7744.518	165.725	19725.886	255.681
23	.207	0.009	.486	0.010	.778	0.006	667.890	48.078	4131.780	113.966	10684.281	135.459
24	.179	0.008	.412	0.009	.668	0.006	1067.339	61.205	6657.027	161.619	17069.697	181.508

						LII Pairs						
25	.010	0.007	.011	0.008	.011	0.008	27.913	7.446	42.837	7.675	69.781	9.921
26	.012	0.007	.012	0.007	.014	0.007	28.017	7.586	44.066	8.865	68.953	10.188
27	.007	0.007	.007	0.008	.007	0.008	26.359	6.876	41.551	8.696	66.670	10.547
28	.017	0.008	.018	0.007	.021	0.008	27.274	7.807	45.421	6.899	70.713	9.196
29	.022	0.007	.026	0.007	.028	0.008	26.732	6.425	41.062	6.937	64.030	10.794
30	.021	0.007	.024	0.008	.027	0.007	26.404	7.264	40.165	7.616	62.633	8.917
31	.011	0.008	.010	0.008	.012	0.008	26.784	6.654	38.450	7.512	57.706	9.485
32	.022	0.009	.023	0.008	.025	0.009	26.525	7.112	37.579	7.178	60.202	8.576
33	.018	0.009	.019	0.007	.020	0.008	24.794	7.411	34.100	7.155	54.201	9.555
34	.018	0.008	.020	0.007	.021	0.008	25.306	7.297	34.658	7.554	52.587	9.168
35	.009	0.007	.008	0.008	.010	0.008	26.165	6.809	33.249	8.076	47.279	8.456
36	.019	0.008	.020	0.008	.020	0.008	26.606	7.217	35.594	6.498	57.047	9.318
37	.012	0.007	.012	0.008	.014	0.007	26.589	8.020	43.113	9.108	65.231	9.312
38	.014	0.006	.016	0.007	.018	0.008	26.193	5.970	42.490	7.750	65.429	9.082
39	.019	0.007	.021	0.008	.023	0.008	26.833	6.770	41.509	7.375	64.095	9.686
40	.017	0.007	.020	0.008	.021	0.007	28.043	6.872	39.575	7.628	63.979	9.815
41	.012	0.008	.013	0.009	.014	0.009	25.004	7.136	35.306	8.457	53.684	9.935
42	.014	0.007	.013	0.008	.017	0.007	25.984	7.030	38.177	8.489	54.514	9.112
43	.013	0.008	.015	0.008	.015	0.007	24.726	6.587	37.110	7.922	55.284	9.042
44	.010	0.009	.012	0.008	.012	0.008	26.200	7.680	35.865	8.752	52.589	10.014
45	.018	0.007	.021	0.008	.020	0.007	27.910	7.783	41.843	8.366	62.007	9.252
46	.012	0.007	.015	0.008	.015	0.007	25.770	7.023	38.273	7.920	59.335	9.375
47	.027	0.008	.029	0.008	.030	0.008	25.970	6.328	40.456	7.428	63.901	9.279
48	.009	0.008	.010	0.007	.010	0.007	26.201	6.595	39.501	9.439	58.644	8.777

Descriptive statistics for Q_3 in CONV condition. Table 9 also shows that the average Q_3 values across the LID pairs were near .20 given a low level of LID, .50 given a medium level of LID, and .80 given a high level of LID; essentially, the Q_3 values seemed to reflect the π_{LID} values that were used to induce dependencies in the responses within an item pair. On the other hand, the average Q_3 values across the LII pairs remained near zero, regardless of the level of LID induced in the responses to other items on the instrument. For all pairs in all LID conditions of the CONV administration, the standard deviations of the Q_3 values across the 100 replications remained small.

It is difficult to compare the descriptive results from the CONV administration to those of previous studies with the exception of Pommerich and Ito (2008). In most papers, means and standard deviations of the LID statistics were only presented for the null conditions and not for the LID conditions; instead, only power and false positive rates were presented (Chen & Thissen, 1997; Kim et al., 2007; Lin, Kim, & Cohen, 2006). For the LID pairs tracked in Pommerich and Ito's (2008) study that used dichotomous data and a fixed $\pi_{LID} = 1$, X^2 values ranged from 100 to 800 and Q_3 values ranged from 0.39 to 0.95. Because the generating model was the 3PL, which, unlike the GRM, includes a guessing parameter, it is not actually possible to simulate perfect LID. This potential for guessing could be the reason that the Q_3 values did not consistently reflect the π_{LID} value as they did in the current study.

Sample size per item pair for CAT condition. Table 10 shows the results for all tracked item pairs with a non-zero sample size under the CAT administration with varying levels of LID induced in the responses. Although the within-item pair sample size for a given pair was not identical across the three LID conditions, it remained quite

consistent with the sample sizes presented in Table 8, and thus, is not presented again in Table 10.

Descriptive statistics for X^2 in CAT condition. For the X^2, the influence of sample size on the statistic is again apparent in the CAT conditions with LID. When the LID level was low, ignoring the pair with a near-zero sample size (Pair 10), the average X^2 value for LID pairs ranged from about 100 to 4,800; with a medium level of LID, they ranged from about 400 to 8,000, and from 1,000 to 26,000 with a high level of LID. Furthermore, the standard deviation of the X^2 values across the 100 replications was large, and also appeared to increase with the level of LID. There was also variation in the average X^2 values across the LII item pairs, though the range was far less extreme than for the LID pairs. Given a low level of LID, the average X^2 values fell between 10 and 90 for the LII pairs; with a medium level of LID, they ranged from 6 to 109, and from 11 to 140 with high LID. Although X^2 values as large as 26,000 were also observed for LID pairs in Pommerich and Ito's (2008) study based on dichotomous adaptive data, the trustworthiness of their values as identifying LID is questionable given that similarly large values were observed for LII pairs is well. The alternate approach specified in Equation 7 for calculating the X^2 with an empirical distribution of person parameters does a considerably better job of yielding values that distinguish the LID pairs from the LII pairs than the traditional approach specified in Equation 5.

Table 10. Results summary for tracked item pairs with a non-zero sample size in the CAT condition with varying LID

Pair	Q_3						X^2					
	Low LID		Med LID		High LID		Low LID		Med LID		High LID	
	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
LID Pairs												
1	.178	0.023	.493	0.020	.798	0.015	4804.833	18772.790	8088.162	20498.379	21165.992	43267.309
3	.158	0.127	.450	0.101	.754	0.081	119.662	842.850	467.408	1977.620	26013.156	172879.562
6	.145	0.127	.452	0.117	.770	0.094	194.532	1076.156	962.494	2184.048	999.697	3113.818
10	.753	0.642	.785	0.606	.922	0.370	2.799	11.223	3.638	15.869	1.054	4.896
14	.166	0.014	.455	0.013	.769	0.010	315.134	37.921	1708.120	76.518	4319.484	109.848
15	.142	0.020	.441	0.021	.766	0.017	635.400	700.099	2949.724	1644.018	8926.229	5463.026
16	.154	0.010	.443	0.011	.753	0.007	586.260	46.668	3740.549	126.767	9602.805	145.479
18	.151	0.033	.468	0.028	.781	0.023	985.156	1857.929	3871.251	4438.124	9876.556	13032.308
19	.149	0.010	.434	0.009	.753	0.008	447.810	42.009	2817.288	95.509	7382.000	100.610
23	.137	0.012	.426	0.012	.749	0.011	380.840	39.032	2353.222	97.432	6098.478	154.493
24	.172	0.300	.492	0.283	.785	0.175	168.703	457.251	5221.869	31024.737	4492.885	17654.716
LII Pairs												
25	-.071	0.020	-.063	0.021	-.058	0.023	26.386	19.078	47.468	73.686	64.796	15.417
26	-.063	0.022	-.053	0.021	-.049	0.022	20.369	12.965	37.116	44.516	59.476	15.287
27	-.028	0.035	-.031	0.030	-.026	0.034	19.135	21.834	17.391	12.502	40.885	153.876
28	-.045	0.018	-.034	0.019	-.031	0.014	34.355	32.285	57.565	31.778	109.005	33.729
29	-.089	0.114	-.088	0.108	-.071	0.101	15.975	38.754	13.931	8.435	16.396	12.552
30	-.036	0.021	-.032	0.021	-.025	0.020	32.527	18.117	38.700	24.797	58.968	69.854
36	-.037	0.172	-.015	0.171	-.034	0.154	18.402	37.031	19.003	45.455	19.362	53.260
38	-.106	0.520	-.118	0.507	-.015	0.497	9.967	38.192	6.295	7.384	10.957	30.872
39	-.022	0.152	-.042	0.150	-.035	0.161	15.961	88.520	22.149	141.200	10.716	18.149
47	-.029	0.009	-.025	0.010	-.022	0.009	90.604	17.374	109.191	18.710	139.189	15.944

Descriptive statistics for Q_3 in CAT condition. Also shown in Table 10, the average Q_3 values across the LID pairs in the CAT administration mirrored those observed in the CONV administration: near .20 given low LID, .50 given medium LID, and .80 given high LID. Again, the average Q_3 values across the LII pairs remained near zero and slightly negative as would be expected with a 20-item test. For both LID and LII pairs, the standard deviations of the Q_3 values across the 100 replications for the LID conditions of the CAT administration remained small. Although Pommerich and Ito (2008) only considered one level of near-perfect LID, they also observed similar patterns. In the CAT administration, the Q_3 values for LII pairs were near zero and slightly more negative than those observed in the CONV administration.

Power and type-I error rates. Given that the null distributions did not adequately reflect the theoretical distributions, in particular for the X^2 statistic in the CAT condition, empirical cut-off values were used instead to determine the power and false positive rates of the LID statistics. Specifically, the 95th percentile for each item pair in the null condition was used as a cut-off value in the LID conditions.[4] Separate cut-off values were considered for the CONV and CAT administrations. For example, the 95th percentile of the X^2 distribution for Pair 19, an LID pair, was 52.32 in the CONV administration and 105.58 in the CAT administration. To be flagged for exhibiting LID at a .05 type-I error rate, the X^2 value for Pair 19 would need to be greater than 52.32 in the CONV conditions but greater than 105.58 in the CAT conditions. To this end, power was calculated as the number of replications out of 100 in which an LID pair was

[4] Even though the Q_3 is a non-directional LID index, the 95th percentile for the Q_3 was used because only positive LID was simulated in the current study.

appropriately flagged as displaying some degree of local dependence. In contrast, the result was considered a false positive if an LII pair was inappropriately flagged as exhibiting LID.

Table 11 presents the power rates of the LID statistics for flagging LID pairs as exhibiting LID under both the CONV and CAT administration conditions. Looking across the 24 LID item pairs, the power of the Q_3 and X^2 was similar. In the CONV administration, both the Q_3 and X^2 were able to identify the LID pairs in all 100 replications for the low, medium, and high LID conditions.

Although not directly comparable to results from previous studies, either because the study considered dichotomous data, the method of LID simulation differed, or both, these findings are reasonably consistent with the literature. For example, in their work with dichotomous data, Kim et al. (2007) found that the Q_3 was appropriately able to identify LID pairs around 70% of the time when the LID level was low and there was a small proportion of LID items. It was able to identify LID pairs close to 100% of the time when the LID level was high regardless of the proportion of items exhibiting LID on the test. Chen and Thissen (1997) also found that nearly 100% of the Fisher's z-transformed Q_3 statistics were greater than the empirical cut-off for LID pairs given even a low level of LID. Finally, Lin, Kim, and Cohen (2006) found that LID pairs were identified nearly 100% of the time using the z-transformed Q_3 statistic even with weak multidimensionality present.

Table 11. Power rates of the LID statistics to flag LID pairs as exhibiting LID

LID Pairs	Q_3						X^2					
	CONV			CAT			CONV			CAT		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
1	100	100	100	100	100	100	100	100	100	100	100	100
2	100	100	100	--	--	--	100	100	100	--	--	--
3	100	100	100	66	100	100	100	100	100	14	43	97
4	100	100	100	--	--	--	100	100	100	--	--	--
5	100	100	100	--	--	--	100	100	100	--	--	--
6	100	100	100	58	99	100	100	100	100	23	81	100
7	100	100	100	--	--	--	100	100	100	--	--	--
8	100	100	100	--	--	--	100	100	100	--	--	--
9	100	100	100	--	--	--	100	100	100	--	--	--
10	100	100	100	7	4	7	100	100	100	5	4	2
11	100	100	100	--	--	--	100	100	100	--	--	--
12	100	100	100	--	--	--	100	100	100	--	--	--
13	100	100	100	--	--	--	100	100	100	--	--	--
14	100	100	100	100	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100	100	100	100	100
16	100	100	100	100	100	100	100	100	100	100	100	100
17	100	100	100	--	--	--	100	100	100	--	--	--
18	100	100	100	100	100	100	100	100	100	100	100	100
19	100	100	100	100	100	100	100	100	100	100	100	100
20	100	100	100	--	--	--	100	100	100	--	--	--
21	100	100	100	--	--	--	100	100	100	--	--	--
22	100	100	100	--	--	--	100	100	100	--	--	--
23	100	100	100	100	100	100	100	100	100	100	100	100
24	100	100	100	30	65	98	100	100	100	22	38	61
Mean	100.000	100.000	100.000	78.273	88.000	91.364	100.000	100.000	100.000	69.455	78.727	87.273

The power of the LID statistics under the CAT administration condition was also high, particularly for medium and high levels of LID. Instances in which the power was low, or the LID statistics were not able to identify the pairs as exhibiting LID, were generally observed for pairs with low within-item pair sample sizes. For example, Pair 24 was administered roughly 20 times out of 20,000 CAT administrations; given a low level of LID, the Q_3 flagged this pair in only 30 of the 100 replications and the X^2 flagged it in just 22. However, even for pairs with smaller sample sizes, the power of both LID statistics increased as the level of LID increased, as would be expected.

The false positive rates of the LID statistics are displayed in Table 12, indicating the number of times out of 100 replications each LII pair was inaccurately flagged as exhibiting LID. The false positive rates for the Q_3 were comparably low across the CONV and CAT administration conditions. Under both methods of administration, the false positive rate of the Q_3 across the pairs fell around 6% given low LID, 10% given medium LID, and 13% given high LID. Although these false positive rates remained relatively low, it should be noted that the type-I error rate was twice as high as the nominal level for LII pairs when other pairs exhibited a moderate level of dependence, and nearly three times as high when other pairs exhibited a high degree of dependence.

These findings are similar to Kim et al. (2007) in that the false positive rate of the Q_3 typically remained below 10% in their study but was as high as 20-30% when there was a high level of LID among a large proportion of other item pairs on the test. Lin, Kim, and Cohen (2006) also observed false positive rates between 20 and 40% for the z-transformed Q_3 statistic when there was a strong degree of multidimensionality in the test.

Table 12. False positive rates of the LID statistics where LII pairs are flagged as LID

	Q_3						X^2					
	CONV			CAT			CONV			CAT		
LII Pairs	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
25	6	14	13	4	7	11	1	27	99	8	18	83
26	4	3	7	5	10	17	7	71	100	5	18	81
27	2	7	6	8	4	7	2	57	100	10	10	12
28	8	8	21	6	27	25	17	94	100	18	71	100
29	5	12	26	5	4	6	7	68	100	3	3	5
30	11	18	31	8	12	18	13	77	100	5	10	27
31	7	2	8	--	--	--	7	61	99	--	--	--
32	12	12	19	--	--	--	10	64	100	--	--	--
33	7	5	9	--	--	--	11	48	97	--	--	--
34	8	9	14	--	--	--	15	56	97	--	--	--
35	3	7	7	--	--	--	3	21	84	--	--	--
36	5	3	6	5	6	2	4	43	100	9	10	9
37	8	16	15	1	--	--	6	67	100	--	--	--
38	4	16	23	1	2	4	5	76	100	6	8	8
39	4	8	13	1	0	2	14	82	100	3	4	6
40	7	15	18	--	--	--	9	52	100	--	--	--
41	4	14	7	--	--	--	6	51	100	--	--	--
42	7	5	12	--	--	--	8	59	100	--	--	--
43	2	7	7	--	--	--	6	58	100	--	--	--
44	5	6	3	--	--	--	13	50	96	--	--	--
45	6	24	18	--	--	--	30	89	100	--	--	--
46	6	16	15	--	--	--	4	53	100	--	--	--
47	9	12	16	18	33	41	7	66	100	10	46	98
48	8	6	8	--	--	--	7	58	100	--	--	--
Mean	6.167	10.208	13.417	5.636	10.500	13.300	8.833	60.333	98.833	7.700	19.800	42.900

Unlike the Q_3, the false positive rate of the X^2 notably increased as the level of LID increased. In the CONV administration, given low LID, the average false positive rate across the pairs was less than 10%. However, it jumped to 60% given a medium level of LID and near 100% given a high level of LID. A similar pattern was observed for the CAT condition, though the false positive rates were not quite as high, likely due to the smaller within-item pair sample sizes associated with CAT. Thus, the X^2 statistic almost always inappropriately identified LII pairs as LID when there were other pairs of items on the instrument that were heavily influenced by LID. Put differently, the statistic seemed to flag all items as aberrant when some items were showing high levels of LID while others did not.

The false positive rates observed in the current study were somewhat larger than those observed by Chen and Thissen (1997) and Lin, Kim, and Cohen (2006). For dichotomous non-adaptive data, Chen and Thissen (1997) saw type-I error rates just slightly larger than the nominal 5% level, particularly when the number of test items was small. Lin, Kim, and Cohen (2006) saw false positive rates as low as 2% but as great as 70% with strong multidimensionality.

Trend Analyses

Visual inspection. Figure 17 and Figure 18 serve to visually summarize the descriptive patterns observed for the LID statistics across combinations of the two administration methods (CONV and CAT) and four LID levels (none, low, medium, and high). They visually display the mean for the LID statistic and 95% confidence intervals, where results for each pair were first averaged across the 100 replications and then averaged across the LII and LID pairs.

The top left quadrant of Figure 17 shows that the mean of the Q_3 statistic was near zero and the confidence intervals were small for LII pairs in the CONV administration. In the top right quadrant, one can see that the mean in the CAT condition was slightly lower than in the CONV condition, which would be expected given a shorter test. Additionally, the confidence intervals were slightly wider, reflecting both the smaller number of LII pairs appearing in CAT and the greater variability in the average Q_3 statistics across pairs.

The bottom left quadrant shows that, for LID pairs in the CONV administration, the mean Q_3 increased in a linear or slightly quadratic fashion as the level of LID increased from LII to high levels of LID. In the bottom right quadrant, a similar trend can be observed for the average Q_3 values associated with LID pairs in the CAT administration, though the confidence intervals surrounding the mean were notably larger. In particular, the width of the confidence intervals surrounding the average Q_3 values in the lower right quadrant was heavily influenced by Pair 10, the LID pair with the smallest within-item pair sample size. As previously noted, with such a small sample size, the Q_3 values for this pair bounced between -1 and 1 across the 100 replications and averaged around .70 regardless of the LID level. This pair was an obvious outlier in the no and low LID conditions, increasing the width of the confidence intervals associated with these conditions. However, this outlier effect was reduced as the level of LID increased (e.g., the medium and high LID conditions), because the Q_3 statistics associated with all pairs fell in the .50 to .80 range. As a result, the confidence intervals surrounding the average Q_3 value in the medium and high LID condition are relatively smaller.

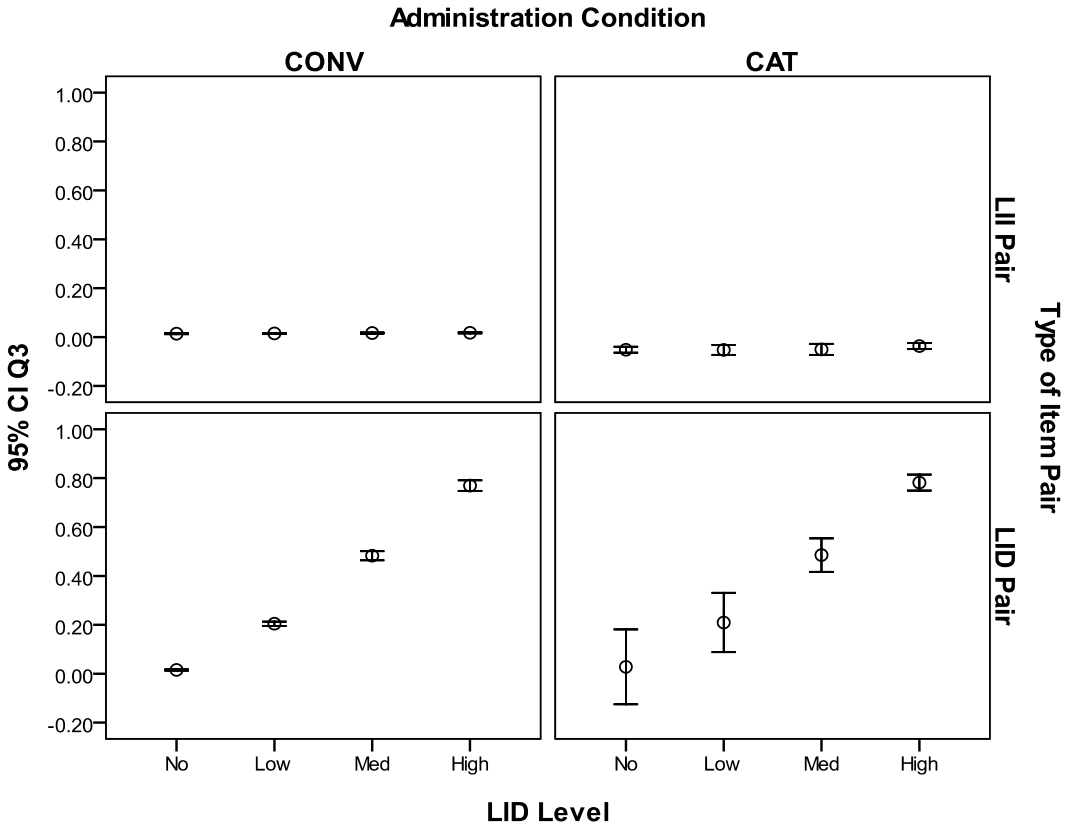


Figure 17. Mean and 95% confidence interval for the Q_3 statistic across conditions

The top left and right quadrants of Figure 18 depict X^2 results for LII items under the CONV and CAT administration, respectively. The scale makes the quadrants appear quite similar in terms of the mean X^2 and width of the confidence intervals, though the results presented in Tables 9 and 10 revealed a slight inflation of the average X^2 across LII pairs as the level of LID induced in other pairs increased. The bottom left quadrant shows that, for LID pairs in the CONV administration, the mean X^2 value increased in a curvilinear fashion as the level of LID increased. In the bottom right quadrant, a similar trend is observed for the average X^2 values associated with LID pairs in the CAT administration, though the confidence intervals surrounding the mean were much wider

than in the CONV condition. Again, this is likely due to the small number of LII item pairs appearing in CAT administrations and the variability in X^2 values related to within-item pair sample size. Thus, the more meaningful comparison apparent in Figure 18 is the comparison of LII pairs against LID pairs; even though there is a slight inflation of the mean for LII item pairs as the LID level increases, they are still easily distinguished from the LID pairs.

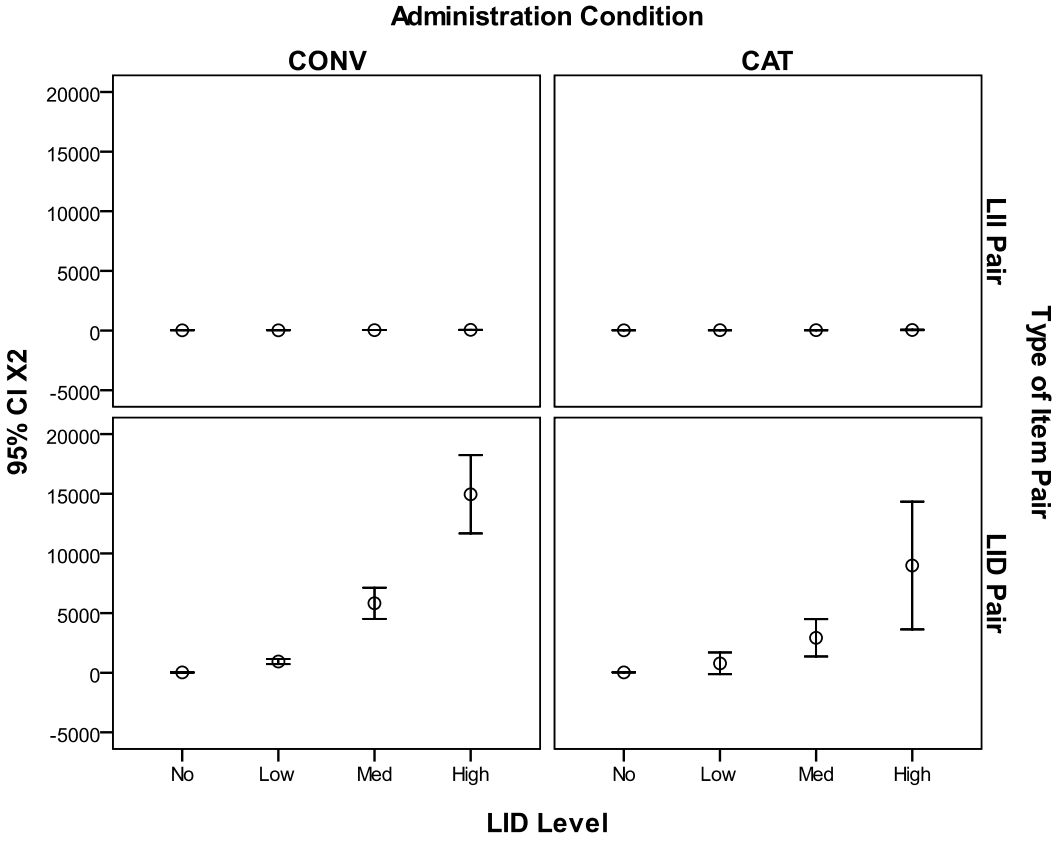


Figure 18. Mean and 95% confidence interval for the X^2 statistic across conditions

General linear modeling analyses. Unadjusted group means, adjusted group means evaluated at the mean within-item pair sample size, and their associated standard

deviations for the Q_3 and X^2 are presented in Tables 13 and 14, respectively. Note that because of its outlier effect, Pair 10 has been excluded from the statistical analyses.

LID item pairs. The first series of tests focused on only the 24 LID item pairs. Results for the Q_3 statistic are presented first, followed by results for the X^2 statistic.

Q_3 results. GLM assumptions were examined using the Q_3 statistic as the dependent variable. Within-item pair sample size was included as a control variable in the statistical analysis and an interaction term between LID level and within-item pair sample size was incorporated in the model to account for any potential violation of the assumption of homogeneity of regression slopes related to the covariate. Mauchly's test (1940) indicated that the sphericity condition did not hold for the within-subjects factor at all levels of the between subjects factor ($p < .05$). In other words, the variance of the difference scores between any two columns in the design matrix was not constant for all pairs of columns. To control for inflation in the type-I error rate associated with this violation, the degrees of freedom and *p*-values associated with the critical F-values in Table 15 have been adjusted accordingly using the Huynh-Feldt (1976) correction.

As shown in Table 15, the GLM indicated a non-significant main effect for administration condition on the Q_3 statistic, $F(1, 31) = 0.22$, $p > .05$, and a non-significant effect for within-item pair sample size, $F(1, 31) = 2.68$, $p > .05$. There was, however, a large, significant main effect for LID level, $F(1.22, 38.57) = 421.75$, $p < .05$, $\eta^2 = .92$. Orthogonal polynomial contrasts among means were examined, revealing highly significant linear, quadratic, and cubic effects for the LID levels. The interaction between LID level and mode of administration was not significant, $F(1.22, 38.57) = 0.89$, $p > .05$,

and neither was the interaction between LID level and within-item pair sample size, $F(1.22, 38.57) = 2.14$, $p > .05$.

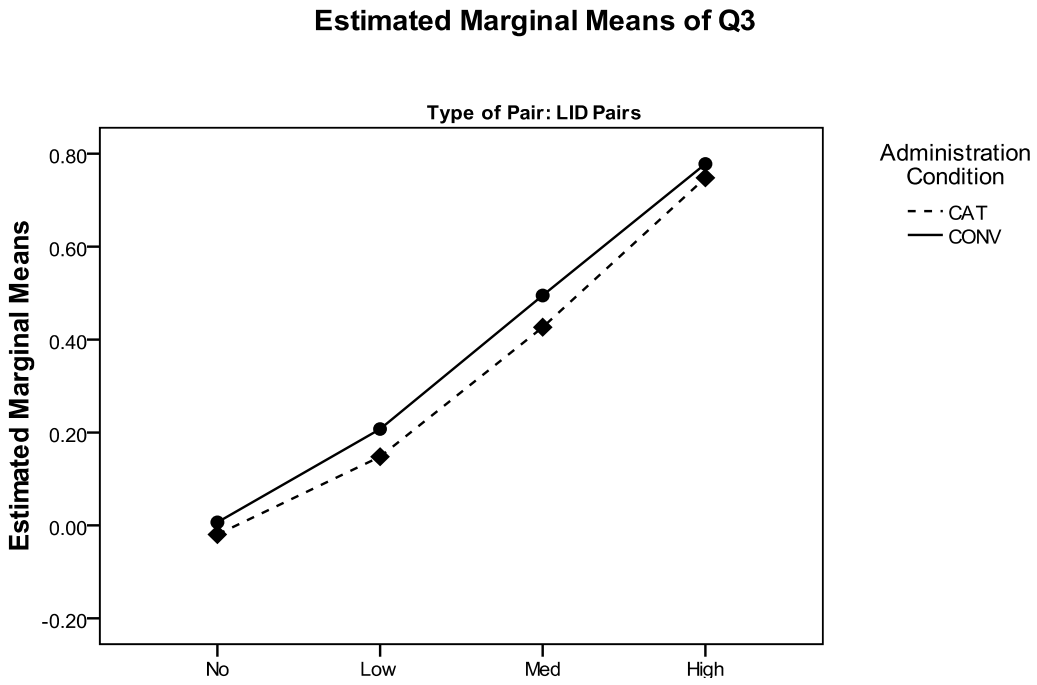


Figure 19. Estimated marginal means of the Q_3 for LID item pairs

Figure 19 depicts the GLM results graphically. It shows that the Q_3 marginal means across the CAT and CONV administration conditions are nearly identical, and increase as the level of LID increases, essentially following a linear trend.

Table 13. Unadjusted and adjusted Q_3 group means

LID Level	Unadjusted Q_3						Adjusted Q_3					
	CONV			CAT			CONV			CAT		
	N	Mean	Stdv	N	Mean	Stdv	N	Mean	Stdv	N	Mean	Stdv
LID Pairs												
None	24	.015	0.005	10	-.040	0.012	24	.007	0.010	10	-.020	0.016
Low	24	.204	0.021	10	.155	0.013	24	.208	0.034	10	.148	0.047
Medium	24	.483	0.044	10	.455	0.023	24	.495	0.069	10	.427	0.098
High	24	.770	0.052	10	.768	0.016	24	.778	0.078	10	.748	0.111
LII Pairs												
None	24	.014	0.005	10	-.052	0.017	24	.005	0.024	10	-.031	0.035
Low	24	.015	0.005	10	-.053	0.028	24	.003	0.039	10	-.022	0.054
Medium	24	.016	0.006	10	-.050	0.032	24	.000	0.039	10	-.011	0.060
High	24	.018	0.006	10	-.037	0.018	24	.012	0.024	10	-.022	0.038

Note. The adjusted Q_3 values are evaluated at within-item pair sample size = 15,547.659 for LID pairs and 14,873.950 for LII pairs.

Table 14. Unadjusted and adjusted X^2 group means

LID Level	Unadjusted X^2						Adjusted X^2					
	CONV			CAT			CONV			CAT		
	N	Mean	Stdv	N	Mean	Stdv	N	Mean	Stdv	N	Mean	Stdv
LID Pairs												
None	24	22.950	0.904	10	33.148	16.728	24	11.864	10.866	10	59.754	15.195
Low	24	937.325	501.594	10	863.833	1409.011	24	1070.707	1542.992	10	543.716	2157.473
Medium	24	5828.298	3098.247	10	3218.009	2219.783	24	5888.152	5187.216	10	3074.359	7252.968
High	24	14952.804	7762.406	10	9887.728	7802.979	24	16498.979	13913.239	10	6176.908	19454.048
LII Pairs												
None	24	23.072	0.847	10	26.660	21.042	24	-4.205	9.102	10	92.125	13.399
Low	24	26.429	0.947	10	28.368	23.154	24	-4.515	6.917	10	102.633	10.179
Medium	24	39.248	3.340	10	36.881	30.094	24	-1.023	10.944	10	133.531	16.109
High	24	60.437	6.172	10	52.975	43.404	24	5.195	24.789	10	185.556	36.486

Note. The adjusted X^2 values are evaluated at within-item pair sample size = 15,547.659 for LID pairs and 14,873.950 for LII pairs.

Table 15. GLM summary table for LID item pairs

	Q_3							X^2						
Source	SS	df	MS	F	*p*	η^2	η_p^2	SS	df	MS	F	*p*	η^2	η_p^2
Between Subjects														
Intercept	0.713	1	0.713	265.216	.000			175800000.000	1	175800000.000	5.810	.022		
Administration	0.001	1	0.001	0.224	.640	.011	.007	8349316.555	1	8349316.555	0.276	.603	.008	.009
Sample Size	0.007	1	0.007	2.679	.112	.077	.080	39420000.000	1	39420000.000	1.303	.262	.040	.040
Error	0.083	31	0.003					938000000.000	31	30260000.000				
Within Subjects														
LID Level	0.537	1.244	0.432	421.749	.000	.924	.932	238800000.000	1.192	200300000.000	5.886	.016	.152	.160
Linear	0.535	1	0.535	465.596	.000			196300000.000	1	196300000.000	6.075	.019		
Quadratic	0.001	1	0.001	14.359	.001			39590000.000	1	39590000.000	5.633	.024		
Cubic	0.001	1	0.001	32.961	.000			2872982.108	1	2872982.108	2.349	.136		
LID Level x Sample Size	0.003	1.244	0.002	2.144	.147	.005	.065	18620000.000	1.192	15620000.000	0.459	.536	.012	.015
Linear	0.002	1	0.002	1.751	.195			11820000.000	1	11820000.000	0.366	.550		
Quadratic	0.001	1	0.001	6.781	.014			5032805.337	1	5032805.337	0.716	.404		
Cubic	0.000	1	0.000	2.026	.165			1766924.733	1	1766924.733	1.444	.239		
LID Level x Administration	0.001	1.244	0.001	0.885	.375	.002	.028	58180000.000	1.192	48790000.000	1.434	.244	.037	.044
Linear	0.000	1	0.000	0.014	.907			47440000.000	1	47440000.000	1.468	.235		
Quadratic	0.001	1	0.001	11.130	.002			10220000.000	1	10220000.000	1.455	.237		
Cubic	0.000	1	0.000	0.771	.387			523853.869	1	523853.869	0.428	.518		
Error	0.040	38.566	0.001					1258000000.000	36.965	34020000.000				
Linear	0.036	31	0.001					1002000000.000	31	32320000.000				
Quadratic	0.003	31	0.000					217800000.000	31	7027254.952				
Cubic	0.001	31	0.000					37920000.000	31	1223249.149				

Note: Huynh-Feldt adjusted degrees of freedom and *p*-values are presented in the table for tests of main and interaction effects.

X^2 results. GLM assumptions were next examined using the X^2 statistic as the dependent variable. Again, Mauchly's test (1940) revealed a significant departure from sphericity ($p < .05$) so the degrees of freedom and *p*-values presented in Table 15 have been adjusted accordingly using the Huynh-Feldt (1976) correction. Also, an interaction term between LID level and within-item pair sample size was again included in the model to relax model assumptions associated with the covariate.

As shown in Table 15, the GLM did not detect a significant main effect for administration condition on the X^2, $F(1, 31) = 0.28$, $p > .05$. These results indicate that the marginal mean for the CONV condition was not different than that of the CAT condition after controlling for within-item pair sample size. The effect associated with within-item pair sample size was also non-significant, $F(1, 31) = 1.30$, $p > .05$. However, LID level did have a large, significant main effect, $F(1.19, 36.97) = 5.89$, $p < .05$, $\eta^2 = .15$. Orthogonal polynomial contrasts were examined, revealing significant linear and quadratic trends among the means. After controlling for within-item pair sample size, the interaction between LID level and mode of administration was also non-significant, $F(1.19, 36.97) = 1.43$, $p > .05$. Additionally, the interaction between LID level and within-item pair sample size was non-significant, $F(1.19, 36.97) = 0.46$, $p > .05$.

Figure 20 depicts the GLM results graphically, illustrating the quadratic trend in X^2 means associated with the level of LID. It shows that the marginal means across the CAT and CONV administration conditions are nearly identical given no LID or a low level of LID. However, as the level of LID increased, the means associated with the CONV administration appeared somewhat greater than their counterparts in the CAT administration, though not to a statistically significant degree.

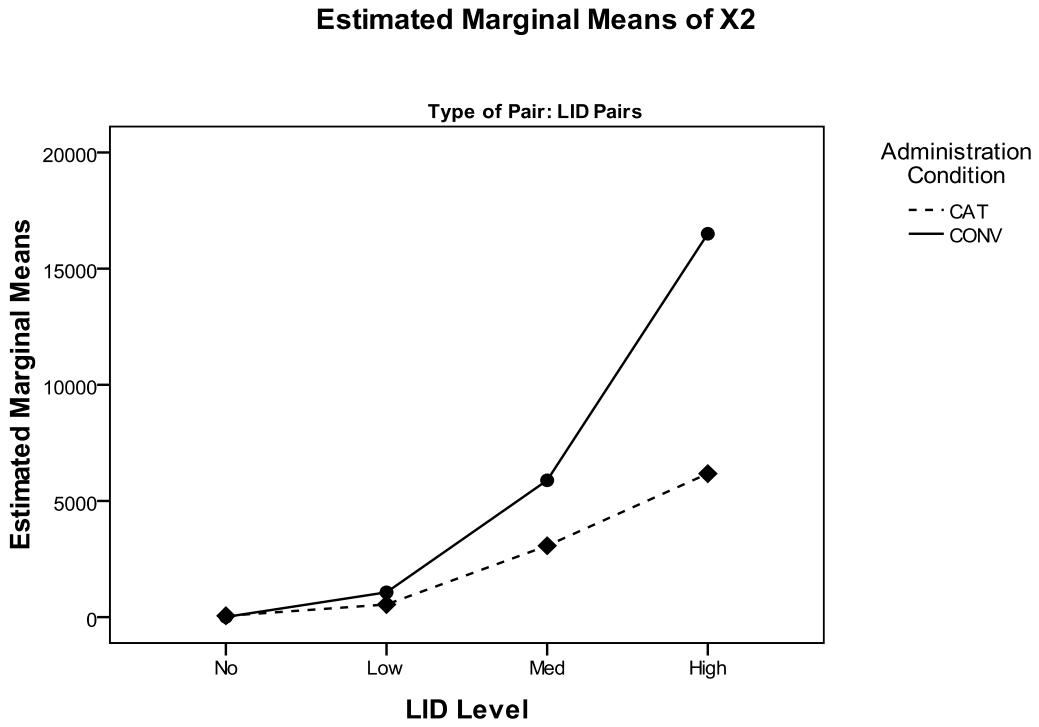


Figure 20. Estimated marginal means of the X^2 for LID item pairs

LII item pairs. The next series of tests focused on only the 24 LII item pairs. Again, results for the Q_3 statistic are presented first followed by results for the X^2 statistic.

Q_3 results. GLM assumptions were examined using the Q_3 statistic as the dependent variable. Mauchly's test (1940) revealed a significant departure from sphericity ($p < .05$) so the degrees of freedom and p-values presented in Table 16 have been adjusted accordingly using the Huynh-Feldt (1976) correction. Also, an interaction term between LID level and within-item pair sample size was again included in the model.

Table 16. GLM summary table for LII item pairs

	Q_3							X^2						
Source	SS	df	MS	F	p	η^2	η_p^2	SS	df	MS	F	p	η^2	η_p^2
Between Subjects														
Intercept	0.006	1	0.006	12.578	.001			8769.171	1	8769.171	207.934	.000		
Administration	0.002	1	0.002	4.678	.038	.111	.131	28375.287	1	28375.287	672.834	.000	.506	.956
Sample Size	0.001	1	0.001	2.296	.140	.056	.069	26358.728	1	26358.728	625.017	.000	.470	.953
Error	0.015	31	0.000					1307.357	31	42.173				
Within Subjects														
LID Level	0.000	1.336	0.000	2.155	.144	.000	.065	488.018	1.250	390.364	4.782	.027	.065	.134
Linear	0.000	1	0.000	4.410	.044			469.573	1	469.573	5.551	.025		
Quadratic	0.000	1	0.000	1.892	.179			18.445	1	18.445	1.520	.227		
Cubic	0.000	1	0.000	2.828	.103			0.001	1	0.001	0.000	.991		
LID Level x Sample Size	0.000	1.336	0.000	1.301	.272	.000	.040	2240.143	1.250	1791.880	21.953	.000	.297	.415
Linear	0.000	1	0.000	0.201	.657			2086.715	1	2086.715	24.667	.000		
Quadratic	0.000	1	0.000	1.295	.264			153.427	1	153.427	12.646	.001		
Cubic	0.000	1	0.000	2.307	.139			0.000	1	0.000	0.000	.997		
LID Level x Administration	0.000	1.336	0.000	0.684	.454	.000	.022	1652.624	1.250	1321.927	16.195	.000	.219	.343
Linear	0.000	1	0.000	0.642	.429			1532.483	1	1532.483	18.116	.000		
Quadratic	0.000	1	0.000	0.606	.442			120.077	1	120.077	9.897	.004		
Cubic	0.000	1	0.000	1.490	.231			0.064	1	0.064	0.012	.913		
Error	0.007	41.417	0.000					3163.347	38.755	81.624				
Linear	0.000	31	0.000					2622.418	31	84.594				
Quadratic	0.006	31	0.000					376.115	31	12.133				
Cubic	0.001	31	0.000					164.815	31	5.317				

Note: Huynh-Feldt adjusted degrees of freedom and *p*-values are presented in the table for tests of main and interaction effects.

As shown in Table 16, the GLM showed a moderate, significant main effect for administration condition, $F(1, 31) = 4.68$, $p < .05$, $\eta^2 = .11$. These results indicate that the marginal mean for the CONV condition was greater than that of the CAT condition, which reflects the impact of test length on the expected value of the Q_3. However, there was no significant effect associated with within-item pair sample size, $F(1, 31) = 2.30$, $p > .05$. After controlling for sample size, there was no evidence of a main effect for LID level, $F(1.34, 41.42) = 2.16$, $p > .05$. Lastly, the interaction between LID level and mode of administration was non-significant, $F(1.334, 41.42) = 0.68$, $p > .05$, and neither was the interaction between LID level and within-item pair sample size, $F(1.34,41.42) = 1.30$, $p > .05$

Figure 21 depicts the GLM results graphically. It show that the Q_3 marginal means associated with the CONV administration consistently fell above those associated with the CAT condition. It also shows that there was no clear trend in the marginal means as the level of LID increased.

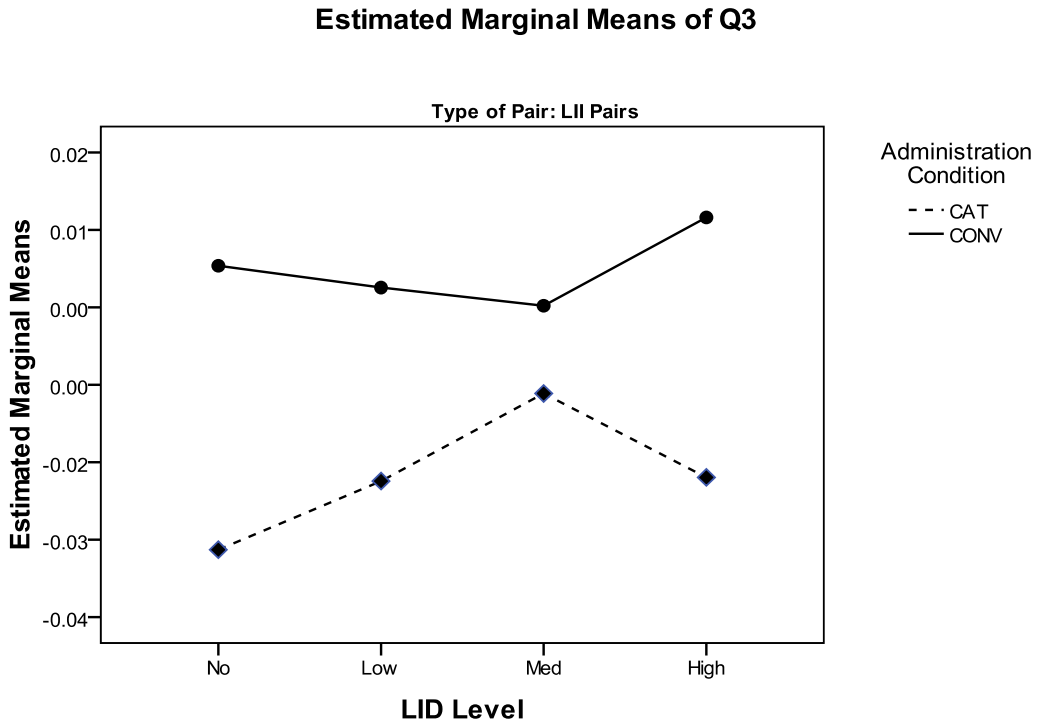


Figure 21. Estimated marginal means of the Q_3 for LII item pairs

X^2 results. GLM assumptions were examined using the X^2 statistic as the dependent variable. Once again, Mauchly's test (1940) revealed a significant departure from sphericity ($p < .05$) so the degrees of freedom and p-values presented in Table 16 have been adjusted accordingly using the Huynh-Feldt (1976) correction. Also, an interaction term between LID level and within-item pair sample size was again included in the model.

As shown in Table 16, there was evidence of a large, statistically significant effect associated with sample size, $F(1, 31) = 625.02$, $p < .05$, $\eta^2 = .47$. Additionally, the main effect of administration on the X^2 was large and significant, $F(1, 31) = 672.83$, $p < .05$,

$\eta^2 = .51$. This finding indicates that, after controlling for within-item pair sample size, the marginal means associated with the CAT condition are greater than those in the CONV condition. The GLM results also showed evidence of a medium, significant main effect associated with LID level, $F(1.25, 38.76) = 4.78$, $p > .05$, $\eta^2 = .07$. Orthogonal polynomial contrasts were examined, revealing significant linear trend among the means. There was also evidence of a large effect for the interaction between LID level and mode of administration, $F(1.25, 38.76) = 16.20$, $p < .05$, $\eta^2 = .22$. Orthogonal polynomial contrasts revealed significant linear and quadratic trends among the means. Finally, there was a large, significant effect associated with the interaction between LID level and sample size, $F(1.25, 38.76) = 21.95$, $p < .05$, $\eta^2 = .30$; a trend analysis showed evidence of linear and quadratic effects.

Figure 22 depicts the GLM results graphically. It shows that the X^2 marginal means associated with CAT administrations were consistently higher than those in the CONV administrations. Figure 22 also reveals the increase in the average X^2 values across both the CONV and CAT conditions as the level of LID increased.

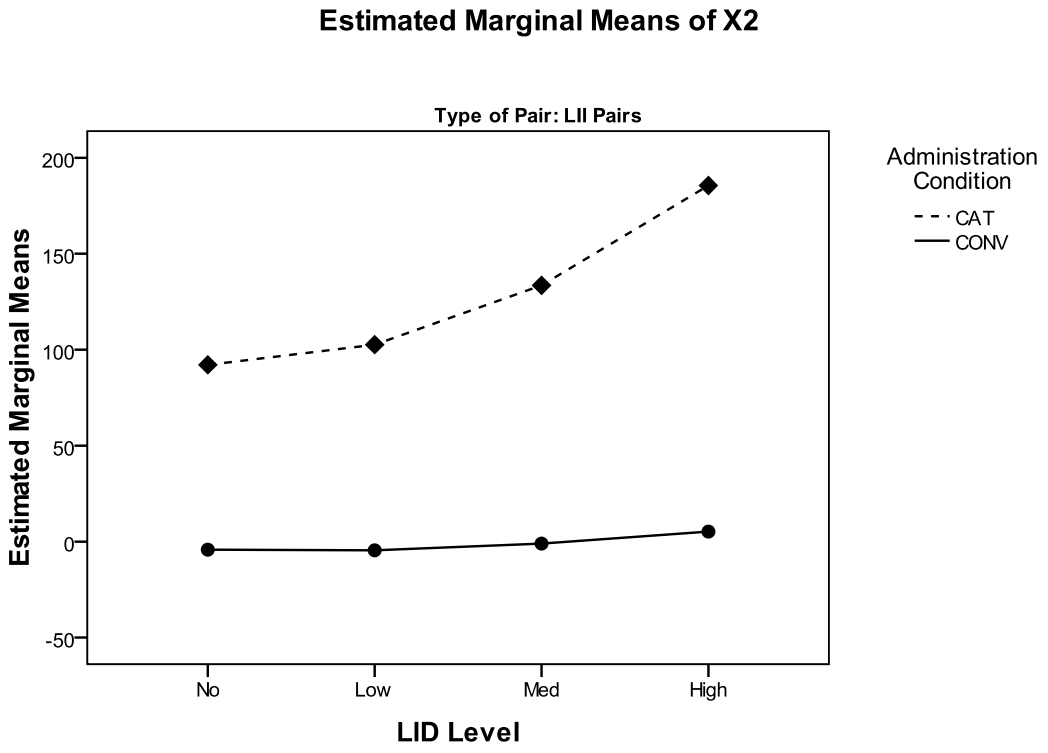


Figure 22. Estimated marginal means of the X^2 for LID item pairs

Statistical summary. Table 17 serves to summarize the patterns identified by the GLM analyses by presenting the effect sizes associated with key factors.

Taken together, the results suggest that the Q_3 outperforms the X^2 for several reasons. First, Table 17 shows that the Q_3 is less influenced by within-item pair sample size than the X^2. For LII item pairs, sample size and the interaction between sample size and LID level were significant sources of variation in the associated X^2 values; there was no evidence of such effects, however, with regards to the Q_3. Second, as indicated by the affiliated large and moderate effects, the performance of the X^2 was notably impacted by administration condition and any LID that may be present in other responses. In contrast,

the Q_3 performed more consistently for LII pairs. Lastly, although the effect associated with LID for both statistics was large, the Q_3 was better able than the X^2 to reveal the degree to which LID is present. For LID pairs, LID level explained only 15% of the total variation in the X^2 values but 92% of the total variation in the Q_3 values.

Table 17. Summary of effects

Factor	Q_3	X^2
LID Pairs		
Administration	NS	NS
Sample Size	NS	NS
LID Level	Large	Large
LID Level x Administration	NS	NS
LID Level x Sample Size	NS	NS
LII Pairs		
Administration	Medium	Large
Sample Size	NS	Large
LID Level	NS	Medium
LID Level x Administration	NS	Large
LID Level x Sample Size	NS	Large

Impact of LID on Trait Estimation

A small follow-up analysis was conducted to determine the impact of various levels of LID exhibited by a sub-set of item pairs on trait estimation. For this analysis, true θ values were sampled from a standard normal distribution for 20,000 individuals and responses to the 95 PROMIS items were generated following the GRM, just as was done in each replication of the primary simulation study. However, using these same 20,000 individuals - as opposed to a new sample - one replication of all eight conditions of the simulation was run. Then, for each condition, respondents' trait estimates were correlated with their true values. The absolute difference between their trait estimates and true trait values was also examined; all results are presented in Table 18.

Table 18. Impact of LID on trait estimation

LID Level	N	$\lvert\hat{\theta}_i - \theta_i\rvert$				$r(\hat{\theta}_i, \theta_i)$
		Min	Max	Mean	Stdv	
			CONV			
No	20000	0.000	1.492	0.107	0.079	.991
Low	20000	0.000	1.492	0.109	0.081	.991
Medium	20000	0.000	1.492	0.112	0.085	.990
High	20000	0.000	1.492	0.115	0.090	.989
			CAT			
No	20000	0.000	1.545	0.115	0.092	.989
Low	20000	0.000	1.545	0.119	0.096	.988
Medium	20000	0.000	1.545	0.124	0.101	.987
High	20000	0.000	1.545	0.131	0.107	.986

Table 18 shows that, for the CONV administration with no LID, the average deviation between individuals' true scores and trait estimates was relatively small (0.11) and the correlation between the true values and estimates was extremely high (r = .99). As the level of LID increased, the average deviation increased and the correlation decreased. The same pattern was observed for the CAT administration; as the level of LID increased, the average absolute difference between respondents' trait estimates and their true values increased and the correlation between the true and estimated values decreased. These results indicate that more "noise" is introduced in the calculation of the LID statistics as the level of LID increases because the LID that is present in responses to certain item pairs begins to compromise the quality of the trait estimation.

Comparing across administration conditions for the same level of LID, the average deviation was always smaller in the CONV administration condition than the CAT administration condition. For example, given high LID, the mean value was 0.12 in the CONV condition and 0.13 in the CAT condition. A similar trend was observed with

regards to the correlation between the estimated and true values; the correlations were stronger in the CONV condition than in the CAT condition. In other words, the trait estimation was slightly more accurate in the CONV condition than in the CAT condition. However, the improvement was not substantial, likely a result of the long CATs modeled in the simulation (i.e., 20 items).

In general, the largest deviations between respondents' true and estimated trait values were observed for individuals with true θ values at the extreme ends of the scale. For example, even with no LID, the mean absolute difference for the 449 individuals with a true θ value of -2 or below was 0.23 in the CONV condition and 0.25 in the CAT condition. In contrast, the 2,725 individuals with true θ values between 1 and 2 had an average absolute deviation of only 0.09 in the CONV condition and 0.10 in the CAT condition.

These results are not surprising given that the maximum item information function for most of the PROMIS Fatigue items falls in this range; the bank has fewer items that are maximally informative in the lower range of the θ scale, making it more challenging to accurately estimate the trait levels of individuals with little fatigue.

Choi and Swartz (2009) acknowledged similar coverage problems in the bank of depression items upon which their CAT simulation study was based. The researchers noted that it is more challenging to generate items targeting the lower range for the construct of depression. In turn, the item bank lacked informative items at the extreme θ levels and the scale information function was shifted markedly towards the moderate to severe levels of depression in relation to the trait distribution. This limited coverage

inevitably leads to poor measurement of individuals at extreme trait levels (i.e., floor and ceiling effects).

Real Data Analysis

The results of the real data analysis are presented in Table 19. Under the CONV administration, the within-item pair sample size was near - but not equal to - the total sample size of 803 given occasional missing responses. The same 48 pairs of items were tracked in the real data analysis as in the simulation component of the study. The average X^2 value across the first 24 item pairs was larger than the average value across the second 24 item pairs (84.84 vs. 47.12, respectively). In other words, the item pairs with similar stems appeared to exhibit more evidence of LID on average than the item pairs with dissimilar stems. The Q_3 values further supported this finding, with the average Q_3 value across the first 24 item pairs equal to 0.32 and 0.02 across the second 24 pairs.

The CAT administration results displayed in Table 19 demonstrate the difficulties associated with identifying LID in real-world adaptive settings. The within-item pair sample sizes for the real-data analysis mirrored those observed in the CAT simulations in terms of their relative magnitude; many of the 48 item pairs were never administered together across the 803 individuals, and sample sizes were generally small - less than 100 - for even those pairs that were administered together. Given that the properties of the LID statistics in the null condition appeared unstable with small sample sizes, the interpretability of the X^2 and Q_3 values is questionable. However, for the pairs with reasonable within-item pair sample sizes (i.e., approximately 100 or more), the LID level appeared low; most Q_3 values fell below 0.20 and most X^2 values fell below 100.

Table 19. Results for tracked item pairs, real data analysis

Item Pair	CONV			CAT			Item Pair	CONV			CAT		
	N	X^2	Q_3	N	X^2	Q_3		N	X^2	Q_3	N	X^2	Q_3
	Similar Stems							Dissimilar Stems					
1	761	108.227	.371	0	--	--	25	736	114.366	.016	84	809.495	-.154
2	761	38.508	.280	515	29.926	.175	26	738	37.942	.116	90	3.087	-.114
3	738	52.174	.279	0	--	--	27	769	58.559	-.052	38	2.190	-.194
4	784	81.436	.261	0	--	--	28	784	37.878	.122	202	17.774	-.047
5	763	20.529	.175	297	33.827	.095	29	758	70.292	-.034	5	11.491	-.805
6	770	220.978	.640	130	63.639	.651	30	781	28.040	.040	101	10.509	.010
7	779	133.649	.377	6	6.749	1.000	31	736	34.416	.067	0	--	--
8	764	24.358	.060	2	0.443	-1.000	32	768	52.613	.088	1	0.235	--
9	773	77.867	.363	510	61.907	.206	33	762	57.616	-.003	0	--	--
10	782	106.238	.413	0	--	--	34	780	17.490	.027	0	--	--
11	739	44.312	.150	280	41.144	.102	35	769	42.292	.072	0	--	--
12	758	91.588	.368	0	--	--	36	783	44.946	-.021	0	--	--
13	762	89.815	.437	0	--	--	37	773	23.962	.041	2	0.059	1.000
14	758	103.897	.307	2	1.838	1.000	38	772	90.747	-.122	0	--	--
15	775	79.864	.347	8	9.665	.364	39	771	56.292	-.001	5	6.822	.001
16	739	33.515	.178	0	--	--	40	784	29.208	.034	4	7.811	.017
17	761	61.237	.439	1	0.655	--	41	736	26.973	.002	0	--	--
18	776	135.055	.461	6	16.145	.973	42	738	23.497	-.047	0	--	--
19	771	178.753	.460	65	31.918	.337	43	768	47.025	-.054	0	--	--
20	739	34.486	.252	0	--	--	44	738	50.810	-.041	0	--	--
21	761	66.505	.159	117	209.325	.093	45	772	31.263	.003	2	6.167	1.000
22	766	118.756	.512	1	0.015	--	46	785	24.219	.010	0	--	--
23	781	87.247	.337	0	--	--	47	736	48.476	.041	446	46.989	-.030
24	763	47.117	.144	0	--	--	48	777	81.852	.067	1	0.016	--
Mean	763.500	84.838	.324	80.833	36.228	.333	Mean	763.083	47.116	.015	40.875	70.973	.062

Chapter 5: Discussion

The current study extended the LID detection literature by examining the properties of popular LID statistics when applied to polytomous adaptive data. Specifically, the study evaluated the performance of Yen's Q_3 and Pearson's X^2 via a simulation study in which the administration condition (CONV vs. CAT) and LID level (none, low, medium, and high) were manipulated. The study's design was driven by and results were supplemented with real data from the PROMIS Fatigue Instrument, the item bank modeled in the simulations.

This chapter begins with a summary of key findings from the study. A discussion of its limitations and suggestions for future research follows. Before concluding, the chapter then addresses implications for practice generally and for the PROMIS network specifically.

Summary of Key Findings

In their study focused on dichotomous adaptive settings, Pommerich and Ito (2008) concluded that the Q_3 appears usable with CAT data and that the X^2 does not. Results of the current study also support the use of the Q_3 with polytomous CAT data, as long as the within-item pair sample size is reasonable. Under the CONV administration in which all individuals respond to all items, the Q_3 values were near zero given no LID; values near zero were also obtained under the CAT administration, though the values in the null condition tended to be higher when the within-item pair sample size was less than 100. These findings suggest that the inferences drawn from the Q_3 regarding the presence of LID may be compromised when the within-item pair sample is small.

When LID was induced, the Q_3 values in the CONV condition surpass the expected values; Q_3 values for LID pairs hover around 0.20 given low LID, 0.50 given medium LID, and 0.80 given high LID. Similar results are observed for LID pairs under the CAT administration conditions. These results suggest the Q_3 is powerful enough to detect even low levels of LID in polytomous CAT data for reasonably large within-item pair sample sizes.

Pommerich and Ito (2008) concluded that the X^2 was unable to perform as desired in CAT and should not be used with CAT data unless modified to use a restricted-range ability distribution in computing expected cell counts. The trait distributions across item pairs of varying difficulty combinations depicted in Figure 9 indeed demonstrate the inappropriateness of integrating over the unrestricted, common ability distribution in X^2 calculations with CAT data. Instead, the current study used an alternative approach to approximate the trait distribution based on the trait estimates of only those individuals responding to the pair of items. Thus, the shape / location of the distribution was not assumed, nor was it assumed to be consistent across item pairs.

In the current study, under the CONV administration condition, X^2 values generally fell around 23, above the theoretical expected value of 16 under LII. This minimal inflation of the X^2 given LII may be due to the fact that the approach used to calculate the X^2 was based on individuals' trait estimates and the trait distribution was approximated empirically using those individuals responding to the pair of items. In other words, basing the calculation on these estimates introduces a small amount of "noise" into the X^2 values, particularly when the within-item pair sample size is small as will often be the case in CAT. Although the X^2 values hover slightly above the critical value

given LII, the X^2 still appears able to distinguish LID item pairs from LII item pairs. When even a low level of LID was present, the X^2 values fell drastically above the critical value – values of more than 1000 – in the CONV administrations.

Although the alternative approach for calculating the X^2 used in the current study is not without its limitations, results suggest it outperforms the traditional approach used by Pommerich and Ito (2008) and LID software programs such as LDIP and IRTFIT given adaptive data. Under the CAT administration condition, no X^2 values were observed for LII pairs that even remotely approach the implausibly large values observed by Pommerich and Ito (2008). Again, the observed X^2 values tended to surpass the χ^2 critical value associated with the theoretical .05 type-I error rate, particularly when the item pair was administered to a large sample of individuals. Ultimately, the X^2 appeared able to distinguish between LID and LII item pairs in the CAT condition as well; the X^2 values fell substantially above the critical value for LID pairs even when the LID level was low.

It is also worth noting that the additional inflation of the LID statistics associated with LII item pairs in LID conditions as compared to the LII condition is not necessarily inaccurate or unexpected. For both the Q_3 and X^2, the calculation of expected performance is conditional on trait level, or, in practice, trait level estimates given than true levels are generally unknown. The LID that is induced in responses to LID pairs will impact the estimation of these trait levels. The trait estimates will be poorer when LID is present, introducing "noise" into the calculation of expected performance for all items, even LII pairs. This "noise" will have even more of an impact on trait estimates when instruments are short (e.g., a 20-item CAT as compared to a 95-item fixed-form).

Limitations and Future Directions

Several limitations of the current study restrict the generalizability of the results and prompt suggestions for future research.

LID probabilities. In the current study, a constant level of LID was simulated across all pairs using a fixed π_{LID} value in each condition, which implies that every item pair is equally affected. Using a fixed π_{LID} value with the SLD model is the precedent in the LID detection literature (Chen & Thissen, 1997; Pommerich & Ito, 2008; Pommerich & Segall, 2008). One reason this is done is that the mean value of a given LID statistic across all LID pairs, which is a common evaluation criterion, is more interpretable when a fixed π_{LID} value is utilized. Furthermore, it allows for an investigation of the performance of the LID statistics under the range of extremes (i.e., all pairs are affected to a low degree to all pairs are affected to a high degree).

However, the degree to which a fixed π_{LID} value realistically represents real-world scenarios is questionable (Pommerich & Segall, 2008). The real data analysis in the current study suggests that different item pairs are affected to varying degrees given that the Q_3 values ranged from -0.12 to 0.64 across the 48 tracked item pairs and averaged 0.17 and that the X^2 values ranged from 17.49 to 220.98 and averaged 65.98. This suggests that a mixed π_{LID} condition where some item pairs are more likely to produce identical responses than others may be more representative of what occurs in reality than the fixed π_{LID} conditions.

A mixed π_{LID} condition or a variety of mixed conditions was not considered in the current study because it would not be possible to tease out the impact of the mixed π_{LID}

unless the mixed probabilities were systematically varied; such an additional manipulation would have significantly increased the number of conditions and altered the scope of the current study. Another challenge associated with a mixed π_{LID} condition for the current design is the sample size. A total of 48 item pairs were tracked, a subset of which were never administered together in CAT (yielding no associated Q_3 or X^2 values). If the item pairs had been further divided based on the mixed probabilities used to simulate LID, certain cells of the design would have had few or no observations.

Including systematically varied mixed π_{LID} conditions seems a worthy extension of the research, however. Such conditions could reveal, for example, the expected Q_3 and X^2 values for a pair that exhibits a low level of LID when other pairs in the instrument are affected to a high degree, inducing more "noise" into the θ estimates than if all pairs were affected to a low degree.

Additionally, the Q_3 values observed in the current study seemingly reflect the π_{LID} values used to induce LID in the item responses. However, this study only provides a "snapshot" look at this relationship for select item pairs and four LID levels. Future research could consider additional levels of LID to investigate the nature of the relation between π_{LID} and the magnitude of the Q_3. Although not a focus of the current research, it may be possible to derive the expected value of the Q_3 given different levels of LID instead of, or in addition to, deriving them empirically. If, for example, the relationship between the two was discovered to be linear or indeed isomorphic, then the Q_3 statistic could be used to gauge the level of LID present in responses.

Response categories. Because it was based on the PROMIS Fatigue bank, the current study considered only items accompanied by five-point response scales calibrated under the GRM. Future research may consider polytomous items with an alternative number of response options.

Lin, Kim, and Cohen (2006), for example, considered polytomous items with three, four, and five categories in their simulation based on polytomous non-adaptive data. As would be expected given the theoretical χ^2 distribution posited, the mean and standard deviation of the X^2 distribution were influenced by the number of response categories; as the number of categories increased, so did the mean and standard deviation of the empirical distribution under the null condition. This implies that different the critical values should be considered to flag item pairs for LID depending on the number of response options. Note that in Lin, Kim, and Cohen's (2006) study, the Q_3 was not similarly affected by the number of response categories, and was instead affected by test length.

The current study has provided information about the empirical distribution of the X^2 when five-category items are considered. Based on the results of previous research, and limitations of the current research, it is unwise to generalize these empirical cut-offs to items with a different number of response options. Future research could systematically vary the number of response options accompanying items to investigate its influence on the LID statistics. It may also be useful to track item pairs in which the items have a different number of response options. This could help determine, for example, what cut-offs should be considered when one item utilizes a three-point scale while the other utilizes a seven-point scale.

Model for LID. In the current research, a dependency structure was induced in the response data using the SLD model. As previously discussed, this model is appropriate for modeling certain types of dependency, primarily those associated with item order (e.g., carryover effect, practice and fatigue effects, speededness). Thus, the findings of the present study can only be generalized to such circumstances. Future studies could consider alternative models to generate LID. For example, a multidimensional model or random-effects testlet model could be used to mimic the dependencies commonly encountered in educational settings when a set of items is based on a common stimulus. Results could help determine the stability of the empirical Q_3 and X^2 distributions, expected values, and cut-offs given alternative generating models.

CAT length. The current study considered two administration conditions: a full-bank administration of 95 items and a fixed-length CAT administration of 20 items. The CAT modeled in the study would be considered long (Choi & Swartz, 2009) and represents the maximum number of items allowed in an adaptive administration of the PROMIS Fatigue instrument. In health-outcomes settings where instruments are administered to the very ill, young children, and the elderly, respondent burden is of significant concern and short CATs are desirable (Bjorner et al, 2007). Real administrations of the Fatigue CAT, for example, typically include as few as five or six items (Gershon et al, 2009).

Long, 20-item CATs were modeled in the simulation to increase the within-item pair sample size for the tracked item pairs without drastically increasing the number of CAT administrations, as was the case in Pommerich and Ito's (2008) study. In other words, longer CATs allowed the tracked item pairs to be administered together in more

of the administrations. Had short CATs of just five items been considered, only a small subset of the most informative items in the bank would appear consistently in administrations, and very few item pairs would have a non-zero sample size. Even with a 20-item CAT, approximately half the tracked item pairs were never administered together in 20,000 administrations, yielding no associated Q_3 and X^2 statistics.

Future research could investigate the performance of the Q_3 and X^2 statistics in the context of variable-length CATs. In other words, the stopping rules in simulated CAT administrations could be altered such that individuals receive varying numbers of items. Instead of fixing the minimum and maximum test length – as was done in the current study – perhaps the parameters could be set so that the CAT is terminated after a predetermined level of precision is met. Variable-length CAT would likely yield a more drastic, uneven level of sparseness in the data matrix given that individuals with typical response patterns would likely respond to a very small number of items and those with atypical patterns would respond to several more.

Future research may also consider the impact of LID on the quality of the trait estimation in shorter CATs and / or variable-length CATs. In the current study, even when some item pairs exhibited a high level of dependency, individuals' trait estimates were highly correlated with their true θ values. This is likely due to the fact that each respondent provided a sufficient number of responses that were not affected by LID, leading to reasonably accurate trait estimation. It is hypothesized that the impact of LID on trait estimation would be more variable in the context of short CATs. With just a few items, many administrations will not contain any item pairs affected by LID, while other

administrations could be drastically impacted when just one or two pairs are included given the small number of total items.

Sample size. In the current simulation, within-item pair sample size was not a systematically manipulated factor. For the CONV condition, the sample size was fixed at 20,000. In the CAT condition, within-item pair sample size could not be set a priori, and was determined only after the simulation of 20,000 adaptive administrations. Because the sample size was fixed, the current research was unable to investigate the impact of sample size on the magnitude of LID statistics given a full-bank administration. However, results suggest that that the sample size can impact the magnitude of the LID statistics in adaptive settings where within-item pair sample size varies drastically. When the sample size was small, the empirical distribution of the Q_3 in the null condition was no longer bell shaped but practically uniform. For the X^2, the mean of the empirical distribution increased as the sample size increased, even though no LID was induced in responses.

Almost none of the LID detection studies reviewed manipulated sample size, and instead manipulated factors such as test length and level of dependency. For example, Chen and Thissen (1997) and Lin, Kim, and Cohen (2006) used 1,000 examinees in all conditions of their simulation, Pommerich and Ito (2008) and Pommerich and Segall (2008) fixed the within-item pair sample size at 2,000, and Kim et al. (2007) fixed it at 3,000. Only Levy, Mislevy, and Sinharay (2009) manipulated sample size, considering a small sample of 250, a medium sample of 750, and a large sample of 2,500 individuals. Their results indicate that the distributions and type-I error rates of the X^2 and Q_3 are similar across the different sample sizes in fixed form settings, but that a sample size of

several hundred to one thousand subjects is needed for the LID measures to have reasonable power to detect LID in the form of multidimensionality; even then, the X^2 performed poorly. However, the disparity in the measures' performance across the CONV and CAT conditions in the current study leads one to question the generalizability of Levy, Mislevy, and Sinharay's (2009) findings regarding sample size to adaptive settings.

Future research should formally evaluate the minimum sample size needed to yield stable Q_3 results in CAT settings via simulation. This research topic was also recommended by Pommerich and Ito (2008). Such a study could also formally evaluate the nature of the relationship between the average X^2 values and within-item pair sample size in a CAT context.

Additional discrepancy measures. The current study considered only two of the most widely studied and applied LID statistics, namely the Q_3 and X^2. Results suggested that the Q_3, a correlational measure which makes no distributional assumptions, outperformed the X^2, a count-based measure which makes distributional assumptions to determine expected cell frequencies. However, a number of other discrepancy measures have been proposed (e.g., Kim et al., 2007), and could be incorporated in future research. Given the poor performance of the X^2 in this study and others considering adaptive data (Pommerich & Ito, 2008; Pommerich & Segall, 2008), preference could be given to alternative correlation measures such as the model-based covariance (*MBC*; Reckase, 1997) or residual item covariance (Fu, Bolt, & Li, 2005; McDonald & Mok, 1995). Like, the Q_3, the *MBC* also performed well in Levy, Mislevy, and Sinharhay's (2009) study

based on dichotomous non-adaptive data, and could prove to be useable in adaptive settings as well.

Alternative sparseness patterns. In the CAT condition of the current study, data matrices were generated to follow the specific pattern of sparseness associated with adaptive testing. That is, conditionally missing data were simulated such that only high-level individuals responded to hard items and vice versa, reflecting the restricted trait range represented in CAT responses. However, alternative sparseness patterns that are not necessarily conditional on trait level are also likely to be observed in real-world settings. For example, incomplete block or matrix sampling designs are popular in large-scale surveys designed to cover a number of content areas (e.g., Frey, Hartig, & Rupp, 2008; Hombo, 2003; Rock & Nelson, 1992). In this case, the time and effort needed to complete all items in the pool is impractical and beyond what can be reasonably expected of individuals. Instead, the pool is divided into blocks of items that are then rotated across different test booklets, with each booklet constructed to look like a miniature version of the full instrument with respect to content area sampling. Even though each individual is presented with relatively few items, the content representation is maintained across all individuals. Future research could extend the current investigation to evaluate the performance of the LID statistics when applied to datasets with alternative sparseness patterns. It is possible that the performance of the X^2 may be adequate, or at least improved, with sparseness patterns where the (few) responses that are available are not restricted to a particular trait range.

LID in pre-calibration. In the current study, individuals' trait estimates were re-estimated from the simulated response dataset but item parameters were not; the item

parameters accompanying the PROMIS Fatigue bank were treated as fixed and known throughout the CAT administration, trait estimation, and calculation of LID statistics. This approach is consistent with the scenario in which test developers work to ensure LID is not present in the pre-calibration data and new LID manifests in CAT due to the adaptive administration. Thus, the study serves to address whether the LID statistics can detect LID that was not present during pre-calibration but occurs during adaptive administrations.

In some circumstances, however, LID may be present in the pre-calibration data and ignored by test developers in item parameter estimation; this LID may carry over to adaptive administrations of the instrument as well. In this scenario, the focus of the research would be whether LID that is present but unaccounted for in pre-calibration can be detected in adaptive administrations of the instrument. To model this distinct state of nature, the data generation and calculation of the LID statistics utilized in the current study would need to be altered. That is, LID should be induced in the pre-calibration data and item parameters should be re-estimated from this data matrix with LID. These new item parameters that are not free from LID would be applied in the CAT administration, trait estimation, and calculation of the LID statistics.

Pommerich and Segall (2008), in fact, considered combinations of the two scenarios in their research with dichotomous CAT, looking at the impact of LID on CAT score precision where (1) CAT item parameters were or were not influenced by LID and (2) CAT responses were or were not influenced by LID. The results of their study indicated that LID in CAT item parameters had a very minimal effect on the precision of examinee CAT scores while LID in examinees' CAT item responses had a fairly

substantial effect. This finding supported the focus in the current research on the scenario in which CAT item parameters are not influenced by LID but CAT responses are. However, examining the alternative scenario in which LID is present and unaccounted for in pre-calibration (using item statistics that are re-estimated from a dataset with LID induced) would be a worthy topic for future research to supplement results.

Practical Implications

Applying the LID statistics. One goal of the current study was to provide recommendations for practitioners regarding the application of the LID statistics given polytomous response data, including appropriate cut-off values for flagging item pairs in conventional and adaptive settings. Results indicate that the traditionally suggested cut-off values of ±0.20 for the Q_3 and 26.30 for the X^2 given two items accompanied by five-point scales are inadequate; a strict application of these values would result in low power for the Q_3 and an extremely high false-positive rate for the X^2. Unfortunately, using the results to recommend alternative cut-offs may not be so simple and straightforward.

Based on the simulation results, using a universal cut-off for flagging item pairs seems inappropriate. Both the Q_3 and X^2 are impacted to some degree by the mode of administration. For LII item pairs, the expected value of the Q_3 was smaller for CAT administrations than CONV administrations, reflecting the influence of test length. In contrast, the expected value of the X^2 was larger for CAT administrations than CONV administrations, reflecting both "noise" in the calculation of expected cell counts and variation in the within-item pair sample size. The disparity in results across conditions suggests that unique cut-offs should be used in applications of the LID statistics based on administration mode.

To combat this issue, some researchers have recommended that practitioners run tailored simulations based on their instrument and administration conditions to calculate study-specific empirical cut-off values (e.g., B. Zumbo, personal communication, September 1st, 2010; Chen & Wang, 2007; R. Mislevy, personal communication, December 25th, 2010). This was essentially the approach utilized in the current research, in that the 95th percentile of the statistics' null distribution for each item pair in a given administration condition was used to flag the pair in the affiliated LID conditions. However, the false positive rates associated with these cut-offs were unacceptable, particularly when a high level of LID was induced in responses to other items on the instrument. Thus, the empirical cut-offs derived in the null condition when no items are affected by LID may actually be of limited practical use.

As one alternative, perhaps researchers could consider the empirical cut-offs obtained for LII item pairs when a given level of LID is induced in other responses throughout the instrument. The degree of LID that is modeled could reflect the maximally-acceptable level of LID or a realistically expected level of LID based on previous studies in that domain. For example, instead of using the 95th percentile of the empirical distribution in the null condition, the 95th percentile for LII pairs from the high LID condition in the tailored simulation could be utilized to flag item pairs in that application. In the current study, across the tracked LII pairs, the average Q_3 95th percentile in the high LID condition was 0.03 for the CONV administration and 0.15 for the CAT administration. The average X^2 95th percentile across LII pairs in the high LID condition was 77.67 for the CONV administration and 81.55 in the CAT administration.

With the exception of the Q_3 in the CONV administration, these values are notably higher than the associated empirical cut-offs in the no LID condition.

Another alternative would be to apply the statistics liberally instead of literally to flag items for LID. That is, the results of hypothesis tests could be ignored and the magnitude of the LID statistics could be examined from a descriptive standpoint alone using statistics effectively as effect-size measures. The current study demonstrated that both the Q_3 and X^2 were able to distinguish LID pairs from LII pairs, even though the values associated with LII pairs were slightly inflated as a result of the dependencies present elsewhere in the instrument. Take, for example, Pair 25, an LII pair. The X^2 values associated with this pair crept upwards of 60 in the high LID conditions, leading one to reject the null hypothesis in error. However, compared to the X^2 values associated with LID pairs that were in the thousands, the slightly inflated value of 60 is of little practical concern. Researchers can use the LID statistics in a descriptive manner to focus instead on those other item pairs of obvious concern.

Implications for PROMIS. The analysis of actual PROMIS data demonstrated the inherent challenges in identifying LID in real-world CAT settings. Despite the fact that responses were available for about 800 individuals, the real-data simulations of CAT administrations left very few item pairs with enough observations to draw strong conclusions about the presence of LID, if there were indeed any observations for the item pair at all. Regardless of which LID statistics are used and how they are calculated, within-item pair sample size may be the primary challenge to overcome in LID investigations with CAT data.

The varying frequency with which items were selected for administration in CAT – observed both in the simulation and real-data components of the current study – raises questions about the inclusion of rarely-used items with low discrimination in the PROMIS Fatigue bank. If there is a desire to retain all items, alpha-stratified adaptive testing could be used to address the large number of underexposed items in the pool (Chang & Ying, 1999; van der Linden & Chang, 2003). Alpha stratification forces the CAT algorithm to select items with lower discrimination parameters earlier in the administration when errors in the trait estimate are large, reserving the more discriminating items for later in the administration when the trait estimates converge. Although this method of adaptive testing is typically used in high-stakes educational testing to control item exposure for security reasons, researchers in PROMIS could consider its use to ensure more uniform exposure rates of items. More even use of items in the bank would also improve consistency in the sample size across item pairs, permitting more useful LID investigations with CAT data.

Another key finding of the study with regards to the real-data analysis comes from the full-bank administration. The full-bank analysis of the PROMIS data reveals that the item pairs with similar stems produced larger LID statistics, on average, than pairs which were not as similar in content. Some health-outcomes scientists (e.g., Thissen, Reeve, Bjorner, & Chang, 2007) have posited that problems that require accommodations for LID in CAT settings may be rarer in health-outcomes settings than educational settings. The results of the real-data analysis demonstrate that LID may indeed be a valid concern for polytomous data in non-educational settings.

Conclusion

In conclusion, when the goal is to investigate the presence of LID in CAT data using pairwise statistics, researchers should examine the level at which the statistics function and the trait distribution assumed in their calculation. LID statistics that function at the aggregate level to derive expected performance, particularly those assuming a common trait distribution, may not be useable with CAT data. Preference should be given to LID statistics and / or calculation approaches in which distributional variations across item pairs are irrelevant.

Additionally, the role of within-item pair sample size in CAT applications should be carefully considered in applications of the LID statistics. It is difficult for the statistics to accurately and reliably identify LID among item pairs when the sample size is small. However, even if such pairs are affected by LID, the ability to detect it may be of limited practical concern. That is, if items in LID pairs are almost never administered together, the LID will have little or no impact on a broad scale in operational administrations. Instead, the ability to detect LID among item that are administered together frequently is of greater importance; it is critical to identify and account for dependency issues among these item pairs that will be included often in operational administrations.

Acknowledgements

The author wishes to thank a number of individuals who offered their advice and assistance throughout various stages of this research in addition to her co-advisors, Drs. Jeffrey R. Harring and André A. Rupp, and other committee members: her father, Dr. Robert Mislevy, for his invaluable feedback and encouragement; Dr. Mary Pommerich for early discussions regarding the detection of LID in adaptive data and simulation

approaches; NIH and PROMIS outcomes scientists Drs. Bryce Reeve and Jin-Shei Lai for facilitating the collaboration agreement and access to PROMIS data; and, Drs. Heather Buzick, Daniel “Chip” Denman, and Clement Stone, for assistance and advice regarding the SAS code used in the study.

Appendix A: PROMIS Fatigue Items

PROMIS Item Bank v. 1.0 – Fatigue

Fatigue – Calibrated Items

Please respond to each item by marking one box per row.

	In the past 7 days...	**Never**	**Rarely**	**Sometimes**	**Often**	**Always**
FATEXP02	How often did you feel run-down?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP05	How often did you experience extreme exhaustion?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP08	How often did you feel tired even when you hadn't done anything?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP07	How often did you feel your fatigue was beyond your control?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP16	How often were you sluggish?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP18	How often did you run out of energy?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP19	How often were you physically drained?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP20	How often did you feel tired?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Never	Rarely	Sometimes	Often	Always
FATEXP22	How often were you bothered by your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP24	How often did you have enough energy to enjoy the things you do for fun?.....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
FATEXP26	How often were you too tired to enjoy life?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP28	How often were you too tired to feel happy?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP29	How often did you feel totally drained?....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP31	How often were you energetic?.....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
FATEXP48	How often did you find yourself getting tired easily?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP49	How often did you think about your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP54	How often did you have physical energy?.....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1

In the past 7 days....

		Never	Rarely	Sometimes	Often	Always
FATIMP03	How often did you have to push yourself to get things done because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP04	How often did your fatigue interfere with your social activities?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP05	How often were you less effective at work due to your fatigue (include work at home)?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP06	How often did your fatigue make you feel slowed down in your thinking?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP08	How often were you too tired to watch television?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP09	How often did your fatigue make it difficult to plan activities ahead of time?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP10	How often did your fatigue make it difficult to start anything new?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP11	How often did your fatigue make you more forgetful?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Never	Rarely	Sometimes	Often	Always
FATIMP13	How often were you too tired to do errands?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP14	How often did your fatigue make it difficult to organize your thoughts when doing things at work (include work at home)?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP15	How often did your fatigue interfere with your ability to engage in recreational activities?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP16	How often did you have trouble finishing things because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP17	How often did your fatigue make it difficult to make decisions?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP18	How often did you have to limit your social activities because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Never	Rarely	Sometimes	Often	Always
FATIMP19	How often were you too tired to do your household chores?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP20	How often did your fatigue make you feel less alert?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP21	How often were you too tired to take a bath or shower?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP22	How often did your fatigue make it difficult to organize your thoughts when doing things at home?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP24	How often did you have trouble starting things because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP25	How often was it an effort to carry on a conversation because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP26	How often were you too tired to socialize with your family?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP29	How often were you too tired to leave the house?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Never	Rarely	Sometimes	Often	Always
FATIMP30	How often were you too tired to think clearly?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP33	How often did your fatigue limit you at work (include work at home)?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP40	How often did you have enough energy to exercise strenuously? ……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
FATIMP42	How often were you less effective at home due to your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP53	How often were you too tired to take a short walk?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP55	How often did you have to force yourself to get up and do things because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP56	How often were you too tired to socialize with your friends?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Not at all	A little bit	Somewhat	Quite a bit	Very much
AN1	I feel listless ("washed out")……....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN2	I feel tired…….....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN3	I have trouble starting things because I am tired……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN4	I have trouble finishing things because I am tired…….....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN5	I have energy……..	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN7	I am able to do my usual activities…..……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN8	I need to sleep during the day……..……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN12	I am too tired to eat……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN14	I need help doing my usual activities……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
AN15	I am frustrated by being too tired to do the things I want to do……..….	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1

In the past 7 days...

		Not at all	A little bit	Somewhat	Quite a bit	Very much
AN16	I have to limit my social activity because I am tired……..	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
FATEXP12	To what degree did you feel tired even when you hadn't done anything?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP13	How bushed were you on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP21	How fatigued were you when your fatigue was at its worst?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP34	How tired did you feel on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP35	How much were you bothered by your fatigue on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP36	How exhausted were you on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP38	How fatigued were you on the day you felt most fatigued?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Not at all	A little bit	Somewhat	Quite a bit	Very much
FATEXP40	How fatigued were you on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP41	How run-down did you feel on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP42	How much mental energy did you have on average?.....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
FATEXP43	How physically drained were you on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP44	How energetic were you on average?.....	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
FATEXP45	How sluggish were you on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP50	How fatigued were you on the day you felt least fatigued?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP51	How easily did you find yourself getting tired on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP52	How wiped out were you on average?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Not at all	A little bit	Somewhat	Quite a bit	Very much
FATIMP01	To what degree did you have to push yourself to get things done because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP02	To what degree did your fatigue make you feel slowed down in your thinking?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP27	To what degree did you have trouble starting things because of your fatigue?...	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP28	How hard was it for you to carry on a conversation because of your fatigue?....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP34	To what degree did you have to limit your social activities because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP35	To what degree did your fatigue make it difficult to organize your thoughts when doing things at home?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP36	To what degree did your fatigue make it difficult to start anything new?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Not at all	A little bit	Somewhat	Quite a bit	Very much
FATIMP37	Due to your fatigue were you less effective at work (include work at home)?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP38	To what degree did your fatigue make it difficult to make decisions?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP43	To what degree did your fatigue make it difficult to organize your thoughts when doing things at work (include work at home)?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP44	To what degree did your fatigue make you more forgetful?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP45	To what degree did your fatigue interfere with your ability to engage in recreational activities?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP47	To what degree did you have to force yourself to get up and do things because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP48	To what degree did your fatigue interfere with your social activities?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

In the past 7 days...

		Not at all	A little bit	Somewhat	Quite a bit	Very much
FATIMP49	To what degree did your fatigue interfere with your physical functioning?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP50	Did fatigue make you less effective at home?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP51	To what degree did you have trouble finishing things because of your fatigue?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP52	To what degree did your fatigue make you feel less alert?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
HI7	I feel fatigued……..……..	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
HI12	I feel weak all over ……..……	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1
	In the past 7 days…	**None**	**1 day**	**2-3 days**	**4-5 days**	**6-7 days**
FATEXP46	On how many days was your fatigue worse in the morning?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
		None	**Mild**	**Moderate**	**Severe**	**Very severe**
FATEXP56	What was the level of your fatigue on most days?.....	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Appendix B: Tracked Item Pairs

Difficulty Combo	Pair #	Variable Name	Item Stems
E		FATEXP19	In the past 7 days... How often were you physically drained?
E	5	FATEXP43	In the past 7 days... How physically drained were you on average?
E		FATEXP21	In the past 7 days... How fatigued were you when your fatigue was at its worst?
E	6	FATEXP38	In the past 7 days... How fatigued were you on the day you felt most fatigued?
E		FATEXP31	In the past 7 days... How often were you energetic?
E	19	FATEXP44	In the past 7 days... How energetic were you on average?
E		FATEXP34	In the past 7 days... How tired did you feel on average?
E	9	FATEXP40	In the past 7 days... How fatigued were you on average?
M		FATIMP1	In the past 7 days... To what degree did you have to push yourself to get things done because of your fatigue?
M	21	FATIMP55	In the past 7 days... How often did you have to force yourself to get up and do things because of your fatigue?
M		FATIMP4	In the past 7 days... How often did your fatigue interfere with your social activities?
M	10	FATIMP56	In the past 7 days... How often were you too tired to socialize with your friends?
M		FATIMP33	In the past 7 days... How often did your fatigue limit you at work (include work at home)?
M	7	FATIMP5	In the past 7 days... How often were you less effective at work due to your fatigue (include work at home)?
M		An3	During the past 7 days: I have trouble <u>starting</u> things because I am tired
M	11	FATIMP10	In the past 7 days... How often did your fatigue make it difficult to start anything new?
H		FATIMP14	In the past 7 days... How often did your fatigue make it difficult to organize your thoughts when doing things at work (include work at home)?
H	12	FATIMP43	In the past 7 days... To what degree did your fatigue make it difficult to organize your thoughts when doing things at work (include work at home)?
H		FATIMP28	In the past 7 days... How hard was it for you to carry on a conversation because of your fatigue?
H	17	FATIMP25	In the past 7 days... How often was it an effort to carry on a conversation because of your fatigue?

H		FATIMP34	In the past 7 days... To what degree did you have to limit your social activities because of your fatigue?
H	18	FATIMP48	In the past 7 days... To what degree did your fatigue interfere with your social activities?
H		FATIMP38	In the past 7 days... To what degree did your fatigue make it difficult to make decisions?
H	13	FATIMP17	In the past 7 days... How often did your fatigue make it difficult to make decisions?
E		FATIMP40	In the past 7 days... How often did you have enough energy to exercise strenuously?
M	14	FATIMP53	In the past 7 days... How often were you too tired to take a short walk?
E		FATEXP6	In the past 7 days... How often did you feel tired even when you hadn't done anything?
M	1	FATEXP12	In the past 7 days... To what degree did you feel tired even when you hadn't done anything?
E		FATEXP22	In the past 7 days... How often were you bothered by your fatigue?
M	2	FATEXP35	In the past 7 days... How much were you bothered by your fatigue on average?
E		FATEXP36	In the past 7 days... How exhausted were you on average?
M	15	FATEXP52	In the past 7 days... How wiped out were you on average?
M		FATIMP11	In the past 7 days... How often did your fatigue make you more forgetful?
H	22	FATIMP44	In the past 7 days... To what degree did your fatigue make you more forgetful?
M		An4	During the past 7 days: I have trouble finishing things because I am tired
H	3	FATIMP51	In the past 7 days... To what degree did you have trouble finishing things because of your fatigue?
M		FATIMP13	In the past 7 days... How often were you too tired to do errands?
H	4	FATIMP29	In the past 7 days... How often were you too tired to leave the house?
M		An7	During the past 7 days: I am able to do my usual activities
H	16	An14	During the past 7 days: I need help doing my usual activities
E		FATIMP20	In the past 7 days... How often did your fatigue make you feel less alert?
H	23	FATIMP30	In the past 7 days... How often were you too tired to think clearly?
E		FATEXP24	In the past 7 days... How often did you have enough energy to enjoy the things you do for fun?
H	24	FATEXP28	In the past 7 days... How often were you too tired to feel happy?
E		FATEXP54	In the past 7 days... How often did you have physical energy?
H	8	FATIMP49	In the past 7 days... To what degree did your fatigue interfere with your physical functioning?
E		HI7	During the past 7 days: I feel fatigued
H	20	AN1	During the past 7 days: I feel listless ("washed out")
E		FATEXP20	In the past 7 days... How often did you feel tired?
E	25	An5	During the past 7 days: I have energy
E		An2	During the past 7 days: I feel tired
E	26	FATEXP16	In the past 7 days... How often were you sluggish?

E		FATEXP42	In the past 7 days... How much mental energy did you have on average?
E	27	FATEXP18	In the past 7 days... How often did you run out of energy?
E		FATEXP48	In the past 7 days... How often did you find yourself getting tired easily?
E	28	FATEXP2	In the past 7 days... How often did you feel run-down?
M		FATEXP41	In the past 7 days... How run-down did you feel on average?
M	29	FATIMP19	In the past 7 days... How often were you too tired to do your household chores?
M		FATIMP24	In the past 7 days... How often did you have trouble starting things because of your fatigue?
M	30	FATIMP15	In the past 7 days... How often did your fatigue interfere with your ability to engage in recreational activities?
M		An8	During the past 7 days: I need to sleep during the day
M	31	FATIMP16	In the past 7 days... How often did you have trouble finishing things because of your fatigue?
M		FATIMP52	In the past 7 days... To what degree did your fatigue make you feel less alert?
M	32	FATIMP50	In the past 7 days... Did fatigue make you less effective at home?
H		FATIMP2	In the past 7 days... To what degree did your fatigue make you feel slowed down in your thinking?
H	33	FATIMP9	In the past 7 days... How often did your fatigue make it difficult to plan activities ahead of time?
H		FATIMP26	In the past 7 days... How often were you too tired to socialize with your family?
H	34	FATEXP6	In the past 7 days... How often did you feel tired even when you hadn't done anything?
H		FATIMP35	In the past 7 days... To what degree did your fatigue make it difficult to organize your thoughts when doing things at home?
H	35	FATEXP46	In the past 7 days... On how many days was your fatigue worse in the morning?
H		FATIMP18	In the past 7 days... How often did you have to limit your social activities because of your fatigue?
H	36	FATIMP22	In the past 7 days... How often did your fatigue make it difficult to organize your thoughts when doing things at home?
E		FATEXP51	In the past 7 days... How easily did you find yourself getting tired on average?
M	37	FATEXP50	In the past 7 days... How fatigued were you on the day you felt least fatigued?
E		FATEXP56	In the past 7 days... What was the level of your fatigue on most days?
M	38	FATIMP6	In the past 7 days... How often did your fatigue make you feel slowed down in your thinking?
E		FATEXP13	In the past 7 days... How bushed were you on average?
M	39	FATIMP27	In the past 7 days... To what degree did you have trouble starting things because of your fatigue?
E		FATIMP42	In the past 7 days... How often were you less effective at home due to your fatigue?
M	40	FATEXP26	In the past 7 days... How often were you too tired to enjoy life?
M		FATEXP50	In the past 7 days... How fatigued were you on the day you felt least fatigued?
H	41	An16	During the past 7 days: I have to limit my social activity because I am tired

M		FATIMP6	In the past 7 days... How often did your fatigue make you feel slowed down in your thinking?
H	42	HI12	During the past 7 days: I feel weak all over
M		FATIMP27	In the past 7 days... To what degree did you have trouble starting things because of your fatigue?
H	43	FATIMP21	In the past 7 days... How often were you too tired to take a bath or shower?
M		FATEXP26	In the past 7 days... How often were you too tired to enjoy life?
H	44	An12	During the past 7 days: I am too tired to eat
E		FATEXP45	In the past 7 days... How sluggish were you on average?
H	45	FATIMP45	In the past 7 days... To what degree did your fatigue interfere with your ability to engage in recreational activities?
E		FATEXP49	In the past 7 days... How often did you think about your fatigue?
H	46	FATEXP5	In the past 7 days... How often did you experience extreme exhaustion?
E		FATIMP3	In the past 7 days... How often did you have to push yourself to get things done because of your fatigue?
H	47	An15	During the past 7 days: I am frustrated by being too tired to do the things I want to do
E		FATEXP29	In the past 7 days... How often did you feel totally drained?
H	48	FATIMP8	In the past 7 days... How often were you too tired to watch television?

Appendix C: Simulation Steps

Detailed Description of Simulation Study Steps

1) Generate data using IRTGEN
   a) Input PROMIS Fatigue item parameters and indicate model (GRM), number of items (95), and number of examinees (20,000)
   b) Randomly assign each examinee a known theta value from a N(0,1) distribution
   c) For a given examinee, use the GRM, theta value and item parameters for an item to compute probability of examinee responding in each response category
      i) Sum probabilities to calculate cumulative subtotals for each response category
      ii) Select random number from U(0,1) to introduce random error into response
         (1) Compare random number to cumulative probabilities for certain response category
         (2) Assign response category where random number is at or below the cumulative probability
   d) Repeat for every examinee for each item
2) Introduce a dependency structure using new code
   a) In the IRTGEN output, for a given examinee and LID item pair, select a random number from U(0,1)
      i) If random number is at or below π_{LID}, replace response to 2nd item in LID pair with response to 1st item in LID pair
      ii) If random number is greater than π_{LID}, leave original response
   b) Repeat for every examinee for each LID item pair
3) Simulate administration condition using new code and SIMPOLYCAT
   a) For CONV
      i) Simulate a full-bank administration of all 95 items for all examinees using SIMPOLYCAT
      ii) Obtain EAP full-bank trait estimate from SIMPOLYCAT
   b) For CAT
      i) Simulate a CAT administration of 20 items for all examinees using SIMPOLYCAT
         (1) Specify CAT features including MII item selection, EAP trait estimation, and a 20-item administration
      ii) Obtain EAP trait estimate for 20-item CAT from SIMPOLYCAT
      iii) Impute missing values for items not administered in SIMPOLYCAT run
4) Calculate values of LID indices following code modified from Clement Stone/new code
   a) For each replication, save the Q3 and X2 statistics and within-item pair sample size for each of the 48 tracked item pairs

Appendix D: Collaboration Agreement

UNIVERSITY OF MARYLAND

Mitchell Building, Room 1101
College Park, Maryland 20742-1115
301.405.5590 TEL 301.314.9443 FAX

OFFICE OF INSTITUTIONAL RESEARCH, PLANNING AND ASSESSMENT

Jessica Mislevy
Doctoral Candidate
Measurement, Statistics and Evaluation
1101 Mitchell Building
College Park, MD 20742

July 1, 2009

David Cella, Ph.D.
PROMIS Statistical Coordinating Center
Northwestern University Feinberg School of Medicine
9th Floor Rubloff
750 N. Lake Shore Drive
Chicago, IL 60611

RE: Collaboration with PROMIS network

Dear Dr. Cella:

I am expressing my interest in collaborating with the NIH PROMIS initiative and its network of investigators for my dissertation research in the Department of Measurement, Statistics, and Evaluation at the University of Maryland, College Park. This letter is to confirm our understanding of the terms of the collaboration agreement.

Background

A rapidly expanding arena for item response theory (IRT) is in attitudinal and health-outcomes survey applications, often with polytomous items. In particular, there is interest in computer adaptive testing (CAT). Meeting model assumptions is necessary to realize the benefits of IRT in this setting, however. Although initial investigations of local item dependence (LID) have been studied both for polytomous items in fixed-form settings and for dichotomous items in CAT settings, there have been no publications applying LID detection methodology to polytomous items in CAT despite its centrally to these applications. The proposed research will investigate the extension of widely used methods of LID detection, such as Yen's (1984) Q_3 index, in the polytomous adaptive context, using a simulation study and illustrating its use with a real data set of this type.

In order make the simulation study as realistic as possible, and to illustrate the practical implications of this investigation, a real-data component is proposed. Desirable characteristics for the CAT modeled in the simulation study include an item bank of approximately 60 items accompanied by 5-point response scales, calibrated using the graded response model (GRM), in a domain outside educational measurement. The PROMIS Fatigue bank best meets these criteria with a total of 95 five-category items calibrated with the GRM. Basing the simulation component on the Fatigue items and CAT administration will help ensure the simulated response data is as similar as possible to real-world applications. Furthermore, supplementing the simulation component with a real-data analysis of PROMIS responses will offer a solid demonstration of the study's importance in practical applications.

Hypotheses (if applicable)

It is relatively unknown whether LID statistics can reasonably be applied to CAT data because they must operate on a sparse data matrix and responses to each item come from a sample with a restricted range of ability. Previous studies utilizing dichotomous adaptive data (Pommerich & Ito, 2008; Pommerich & Segal, 2008) have found the X^2 statistic that operates at an aggregate level to derive expected and observed performance does not perform reasonably when applied to CAT data, whereas the Q_3 statistic that operates first at the individual level to compare expected and observed performance and then summarizes over examinees may be useable with CAT data. Although specific research hypotheses have not yet been developed at this time, similar findings are expected given a polytomous adaptive context.

Methods/Operational details

In the proposed research, PROMIS banks and materials will be used in several ways. First, the Fatigue bank will be utilized as the model for the simulation component. Real items and their associated parameters will be considered, and the true CAT administration conditions (e.g., item selection rules, scoring) will be mimicked. Second, actual response data obtained in Waves 1 & 2 of the PROMIS project will be analyzed to demonstrate the functionality of LID statistics in a practical application. If response data from CAT administrations are not available, the post-hoc or "real-data" simulations will be conducted using PROMIS data collected in fixed-form settings to mimic adaptive conditions.

Anticipated Mutual Benefits *(to PROMIS and Mislevy's dissertation research)*

This proposed research has clear practical importance for the burgeoning interest of IRT in survey research with CAT. If LID statistics function as expected in the polytomous, adaptive context, they can be used to gauge the effectiveness of strategies used to prevent LID in adaptive administrations, detect LID among items that were not necessarily expected to exhibit LID, and indicate the need to base the CAT on an IRT model that incorporates such dependencies. As an added benefit to the PROMIS project specifically, the proposed study will consider the magnitudes of LID statistics under the null condition (i.e., local item independence), offering guidelines for appropriate cut-off values to flag item pairs for LID in future applications.

I agree to existing PROMIS intellectual property limitations and open collaboration. Any PROMIS Data/Materials will be treated as confidential until it is released into the public domain, and I will take steps to preserve the confidentiality of shared information consistent with prevailing best practices. I understand that I will own the data collected in the research described above but agree to follow NIH Grants Policies concerning the sharing of research data using PROMIS materials. As outlined by the NIH, I will make available to the public the results of this collaboration and any accompanying data that were supported by the NIH. I will consult the following NIH source for guidance on sharing data: Sharing of Research Data as defined in the "NIH Statement on Sharing of Research Data" February, 2003 and referenced online at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

It is understood that the PROMIS Data/Materials are experimental in nature, and provided "as is" and without any warranty, express or implied. The PROMIS network/institutions make no representation that the use of PROMIS Data/Materials supplied by them will not infringe on any patent or other proprietary rights. The PROMIS network/institutions shall not be liable for any claims, losses, or damages resulting from the use of PROMIS Data/Materials or for any loss, claim, damage, or liability of any nature which may arise in connection with use, handling, or storage of the PROMIS Data/Materials, or the research to be conducted with the PROMIS Data/Materials. I am and will remain in compliance with all applicable federal, state, and local laws and regulations, and any lack of compliance shall be at my own risk.

I agree to provide the PROMIS network with a minimum data set of standard demographic information and I will adhere to standard file transfer specifications. I agree to report the version of PROMIS items used in our work in accordance with the PROMIS naming convention. I also

agree not to change the wording of any aspect of a PROMIS item, including the item context, stem and response options. If a PROMIS item is changed, it will not be represented as a PROMIS item in any communication, including publication.

I agree that any costs of the PROMIS network extending support to my project beyond what can be offered under its existing funding will be borne by me. I also agree to cite PROMIS in relevant published or presented work, to provide the PROMIS investigators and network with access to our work in progress, and I agree to work collaboratively in furthering mutual objectives.

The appropriate representative from Dr. Jin-Shei Lai's research team has read and concurred with the terms of this letter.

Sincerely,
Jessica Mislevy
Doctoral Candidate

References

Pommerich M., & Ito, K. (2008, March). *An examination of the properties of local dependence measures when applied to adaptive data.* Presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY, March 25-27.

Pommerich, M., & Segall, D. O. (2008). Local dependence in an operational CAT: Diagnosis and Implications. *Journal of Educational Measurement, 45* (3), 201-223.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8,* 125-145.

Appendix E: IRB Application

UNIVERSITY OF MARYLAND, COLLEGE PARK
Institutional Review Board
Initial Application for Research Involving Human Subjects

Name of Principal Investigator (PI) or Project Faculty Advisor *(NOT a student or fellow)* Jeffrey Harring **Tel. No** 301.405.3630

Name of Co-Investigator (Co-PI) Andre Rupp **Tel. No** 301.405.3623

E-Mail Address of PI harring@umd.edu **E-Mail Address of Co-PI** ruppandr@umd.edu

Name and address of contact to receive approval documents
Jessica Mislevy
1101 Mitchell Building
College Park, MD 20742

Name of Student Investigator Jessica Mislevy **Tel. No.** 301.405.5590

E-Mail Address of Student Investigator jmislevy@umd.edu

Check here if this is a student master's thesis ☐ or a dissertation research project X

Department or Unit Administering the Project Measurement, Statistics and Evaluation

Project Title Detecting Local Item Dependence in Polytomous Adaptive Data

Funding Agency:

ORAA Proposal ID Number:

Names of any additional Federal agencies providing funds or other support for this research project:

Target Population: The study population will include (Check all that apply):

- ☐ pregnant women
- ☐ minors/children
- ☐ human fetuses
- ☐ neonates
- ☐ prisoners
- ☐ students
- ☐ individuals with mental disabilities
- ☐ individuals with physical disabilities

Exempt or Nonexempt (Optional): You may recommend your research for exemption or nonexemption by checking the appropriate box below. For exempt recommendation, list the numbers for the exempt category(s) that apply. Refer to pages 6-7 of this document.

☐ **Exempt----List Exemption Category(s)** ***Or*** **X Non-Exempt**

If exempt, briefly describe the reason(s) for exemption.

8/27/09 **Date** **Signature of Principal Investigator or Faculty Advisor**

08/27/09 **Date** **Signature of Co-Principal Investigator**

8/27/09 **Date** **Signature of Student Investigator**

Date **REQUIRED** Departmental Signature
Name George Macready George Macready **Title** Dept. IRB Rep.
(Please also print name of person signing above)

(PLEASE NOTE: The Departmental signature block should not be signed by the investigator or the student investigator's advisor.)

For Internal Use Only (to be completed by the IRB Office) Application #:

Application Materials

Title

Detecting Local Item Dependence in Polytomous Adaptive Data

1. Abstract

A rapidly expanding arena for item response theory (IRT) is in attitudinal and health-outcomes survey applications, often with polytomous items. In particular, there is interest in computer adaptive testing (CAT). Meeting key model assumptions is necessary to realize the benefits of IRT in this setting, however. The current research will investigate the properties of statistics used to detect local item dependence (LID) in the polytomous adaptive context, using a simulation study and illustrating its use with a real data set of this type.

To make the simulation study as realistic as possible and illustrate the practical implications of the investigation, a real-data component is included. The National Institutes of Health (NIH) and health-outcomes scientists at institutions across the country have formed a cooperative network to develop the Patient-Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org). The PROMIS Fatigue bank and response data collected in Wave I testing will be used in the current study. The Fatigue CAT will be modeled in the simulation component and actual response data will be analyzed to demonstrate the functionality of LID statistics in a practical application.

Confidentiality of subjects will be maintained through the use of a pre-existing, de-identified dataset.

2. Subject Selection

a. The current study will utilize pre-existing data collected in Wave I of PROMIS item bank testing. Data were collected in 2006 and 2007 from the U.S. general population and multiple disease populations primarily by the polling firm YouGovPolimetrix (www.polimetrix.com). Subjects were selected from a panel of over one million respondents maintained by YouGovPolimetrix. Individuals in the panel regularly participate in YouGovPolimetrix Internet surveys and have provided YouGovPolimetrix with their names, physical addresses, email addresses, and other information. Subjects were recruited by a variety of methods, including e-random digit dialing, invitations via web newsletters, and Internet poll-based recruitment. A small number of subjects were also recruited from primary research sites associated with PROMIS network sites.

b. Subjects were selected to meet specified targets in terms of gender (50% female), age (20% in each of 5 age groups: 18-29, 30-44, 45-59, 60-74, 75+), race/ethnicity (10% black and Hispanic), and education (10% less than high school graduate). Persons who self-reported currently being diagnosed with a given condition were included in the clinical sampled associated with that condition (e.g., chronic pain sample).

c. The selection of subjects was made on this basis so that responses to could be collected to the candidate items from the targeted PROMIS domains for both 1) the general U.S. population and 2) specific disease subpopulations. The subset of the PROMIS general population sample, or the scale setting sub-sample, was primarily used to establish U.S. population norms.

d. Overall, the PROMIS Wave I sample included 21,133 participants. The item calibration sample for the Fatigue bank specifically included 14,931 cases in total.

3. Procedures

Subjects were recruited by YouGovPolimetrix online via e-random digit dialing, invitations via web newsletters, and Internet poll-based recruitment or on-site at PROMIS research sites. YouGovPolimetrix sample data were collected using their website on a secure server, while data from the research sites were collected using the PROMIS Assessment System.

Given the large number of candidate items from the targeted PROMIS domains, it was not possible for each participant to respond to every item in the pool. Instead, two data collection designs – full bank and block administration – were used. In the full bank administration, individuals were administered full banks of items for a subset of the PROMIS domains. In the block administration, individuals were administered a subset, or block, of items from each domain. These administration approaches limited the number of items administered to any respondent to roughly 150, and the administration time to less than 40 minutes.

In addition to the candidate items, participants were asked to answer about 20 auxiliary items consisting of global health rating items and sociodemographic variables such as age, income, gender, and race/ethnicity. They were also asked a series of questions about the presence and degree of limitations related to chronic medical conditions.

A copy of the Fatigue instrument is attached. The PROMIS Statistical Coordinating Center received de-identified datasets from YouGovPolimetrix. The de-identified dataset for the Fatigue domain will be utilized in the current study. It contains response data for respondents in the full bank sample, with roughly 800 individuals responding to each of the 95 items included in the bank.

4. Risks and Benefits

The current study poses no known risks to the subjects, as it utilizes a pre-existing data source. The data are de-identified and subjects will not be re-contacted. There are also no known benefits to subjects. Though the research is not designed to benefit the subjects directly, the results will benefit the health-outcomes assessment community generally, helping to improve the methodology used to collect and analyze health-outcomes data.

5. Confidentiality

The data will be treated as confidential until released into the public domain, and steps will be taken to preserve the confidentiality of shared information consistent with prevailing best practices. The confidentiality and privacy of subjects will be maintained through the sole use of a de-identified dataset. The dataset is free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects. Throughout the duration of the research, the de-identified dataset will be stored on the personal computer in the private residence of the student investigator, and will not be accessible by individuals other than the research team, including the PI, Co-PI, and student investigator. The PROMIS network anticipates releasing Wave I response data into the public domain in the fall of 2009, prior to the conclusion of the current research study. Once in the public domain, no additional steps will be necessary to protect the confidentiality of subjects.

6. Information and Consent Forms

Subjects' participation in the original data collection effort was voluntary. The panel from which the original sample was drawn is an opt-in panel. YouGovPolimetrix provides anyone interested with the opportunity to participate in pools on a variety of topics. Information and consent forms will not be utilized for the current study, as it utilizes this pre-existing data source. Subjects will not be re-contacted and no new information will be requested.

7. Conflict of Interest

No conflict of interest.

8. HIPAA Compliance

Not applicable.

9. Research outside the United States

Not applicable.

10. Research Involving Prisoners

Not applicable.

Appendix F: IRB Approval

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

2100 Lee Building
College Park, Maryland 20742-5125
301.405.2412 TEL 301.314.1475 FAX
irb@deans.umd.edu
www.umresearch.umd.edu/IRB

October 30, 2009

MEMORANDUM

Application Approval Notification

To: Dr. Jeff Harring
Andre Rupp
Jessica Mislevy
Measurement, Statistics & Evaluation

From: Joseph M. Smith, MA, CIM

IRB Manager

University of Maryland, College Park

Re: **IRB Application Number:** 09-0712

Project Title: "Detecting Local Item Dependence in Polytomous Adaptive Data"

Approval Date: October 30, 2009

Expiration Date: October 30, 2012

Type of Application: Initial

Type of Research: Exempt

Type of Review for Application: Exempt

The University of Maryland, College Park Institutional Review Board (IRB) approved your IRB application. The research was approved in accordance with the University IRB policies and procedures and 45 CFR 46, the Federal Policy for the Protection of Human Subjects. Please include the above-cited IRB application number in any future

communications with our office regarding this research.

Recruitment/Consent: For research requiring written informed consent, the IRB-approved and stamped informed consent document is enclosed. The expiration date for IRB approval has been stamped on the informed consent document. Please keep copies of the consent forms used for this research for three years after the completion of the research.

Continuing Review: If you intend to continue to collect data from human subjects or to analyze private, identifiable data collected from human subjects, after the expiration date for this approval (indicated above), you must submit a renewal application to the IRB Office at least 45 days before the approval expiration date. If IRB approval of your project expires, all human subject research activities including the enrollment of new subjects, data collection, and analysis of identifiable private information must stop until the renewal application is approved by the IRB.

Modifications: Any changes to the approved protocol must be approved by the IRB before the change is implemented, except when a change is necessary to eliminate apparent immediate hazards to the subjects. If you would like to modify the approved protocol, please submit an addendum request to the IRB Office. The instructions for submitting a request are posted on the IRB web site at : http://www.umresearch.umd.edu/IRB/addendum app.htm

Unanticipated Problems Involving Risks: You must promptly report any unanticipated problems involving risks to subjects or others to the IRB Manager at 301-405-0678 or jsmith@umresearch.umd.edu.

Student Researchers: Unless otherwise requested, this IRB approval document was sent to the Principal Investigator (PI). The PI should pass on the approval document or a copy to the student researchers. This IRB approval document may be a requirement for student researchers applying for graduation. The IRB may not be able to provide copies of the approval documents if several years have passed since the date of the original approval.

Additional Information: Please contact the IRB Office at 301-405-4212 if you have any IRB-related questions or concerns or email at irb@umd.edu.

References

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18,* 256-275.

Ackerman, T., & Spray, J. (1987). *A general model for item dependency* (RR-87-9). Iowa City, IA: ACT.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16,* 95-108.

Bjorner, J. B., Smith, K. J., Stone, C., & Sun, X. (2007). IRTFIT: A macro for item fit and local dependence tests under IRT models [Computer software]. Lincoln, RI: Quality Metric Health Surveys. Available from http://outcomes.cancer.gov/areas/measurement/irt_model_fit.html.

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

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems.* Unpublished doctoral dissertation, University of Texas at Austin.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., Matthias, R., & on behalf of the PROMIS cooperative group. (2007). The Patient Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care, 45,* 3-11.

Chang, C.-H., & Reeve, B. R. (2005). Item response theory and its applications to patent-reported outcomes measurement. *Evaluation & the Healthcare Professions*, *28,* 264-282.

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 24,* 333-341.

Chen, S.-K., & Cook, K. F. (2009). SIMPOLYCAT: An SAS program for conducting CAT simulations based on polytomous IRT models. *Behavior Research Methods, 41,* 499-506.

Chen, S.-K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 53,* 61-77.

Chen, W. (1993). *IRT_LD: A computer program for the detection of pairwise local dependence between test items* (Research Memorandum 93-2). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22,* 265-289.

Chen, C.-T., & Wang, W.-C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement*, *31,* 388-411.

Choi, S. W. (2009). Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT models. *Applied Psychological Measurement, 33,* 644-645.

Choi, S. W., & Swartz, J. R. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33,* 419-440.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardize questionnaire.* Thousand Oaks, CA: Sage Publications.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *The Public Opinion Quarterly, 65,* 230-253.

Craig, S. B., & Harvey, R. J. (2004, April). *Using CAT to reduce administration time in 360° performance assessment.* In Craig, B. (Chair), *360, The next generation: Innovations in multisource performance assessment*. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of Royal Statistical Society. Series B, 46,* 440-464.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.

DeVellis, R. (2003). *Scale Development: Theory and Applications*. (2nd ed.). Sage Applied Social Research Methods Series, Vol. 26. Thousand Oaks, CA: Sage.

De Walt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45,* 12-21.

Dillman, D. A., & Smyth, J. D. (2007). Design effects in the transition to web-based surveys. *American Journal of Preventive Medicine, 32,* 90-96.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129-143.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53,* 61-77.

Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Fennessy, L. M. (1995). *The impact of local dependencies on various IRT outcomes.* Unpublished doctoral dissertation, University of Amherst.

Flaugher, R. (1990). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 103-134). Hillsdale, NJ: Lawrence Erlbaum.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage Publications.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39-53.

Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology*, *23,* 53-57.

Fu, J., Bolt, D. M., & Li, Y. (2005, April). *Evaluating item fit for a polytomous Fusion model using posterior predictive checks.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Canada.

Gershon, R. C. (1989). *Test anxiety and item order: New parameters for item response theory.* Paper presented at the Annual Meeting of the Educational Research Association, San Fransisco, CA.

Gershon, R., Choi, S., Lai, J. S., Wee, H. L., Yoo, H., & Hambleton, R. K. (2009). *Alternative item response theory models for PROMIS.* Paper presented at the 2009 International Society for Quality of Life Research Meeting, New Orleans, LA.

Gershon, R., Rothrock, N. E., Hanrahan, R. T., Jansky, L. J., Harniss, M., & Riley, W. (2010). The development of a clinical outcomes survey research application: Assessment Center[SM]. *Quality of Life Research, 19,* 677-685.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29,* 433-456.

Habing, B., Finch, H., Roberts, J. S. (2005). A Q_3 statistic for unfolding item response theory models. *Applied Psychological Measurement, 29*, 457-471.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Measurement Methods for the Social Sciences Series. Newbury Park, CA: Sage.

Hattie, J. A. (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hays, R. D. (2004, June). *Next steps in use of IRT in the assessment of health outcomes.* Summary paper for the National Cancer Institute and the Drug Information Association co-sponsored "Advances in Health Outcomes Measurement: Exploring the Current State and the Future of Item Response Theory, Item Banks, and Computer Adaptive Testing," Bethesda, MD.

Hombo, C. M. (2003). NAEP and No Child Left Behind: Technical challenges and practical solutions. *Theory into Practice*. *42*, 59-65.

Hoskens, M., & De Boeck, P. (1997). A parametric method for local dependence among test items. *Psychological Methods, 2*, 261-275.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *Journal of Educational Statistics, 1,* 69-82.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous itemresponse models. *Psychometrika, 66*, 109-132.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44,* 1-21.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Jansky, L. J., & Huang, J. C. (2009). A multi-method approach to assess usability and acceptability: A case study of the patient-reported outcomes measurement system (PROMIS) workshop. *Social Science Computer Review, 27,* 267-270.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2010). *Simultaneous modeling of item and person dependence using a multilevel Rasch measurement model.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Jiao, H.**,** Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*, 311-321.

Jiao, H., Wang, S., & Kamata, A. (2007). Modeling local item dependence with the hierarchical generalized linear model. In E. V. Smith & R. M. Smith (Eds.), *Rasch Measurement: Advanced and Specialized Applications.* Maple Grove, MN: JAM press.

Kamata, A. (1999). Some generalizations of the Rasch Model: An application of the hierarchical generalized linear model. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79-93.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests.* Unpublished doctoral dissertation, University of Texas at Austin.

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (April, 2007). *Assessing relative performance of local item dependence (LID) indexes.* Paper presented at the National Council on Measurement in Education Annual Meeting. Chicago, IL.

Kim, S.-H., Cohen, A. S., & Lin, Y.-H. (2006). *LDIP*: A computer program for local dependence indices for polytomous items. *Applied Psychological Measurement, 30,* 509-510.

Kim, S.-H., Cohen, A. S., & Lin, Y.-H. (in press). *LDID*: A computer program for local dependence indices for dichotomous items. *Applied Psychological Measurement.*

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241-261.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8,* 147-154.

Lai, J.-S. (2007). *PROMIS wave 1 analysis summary: Fatigue domain*. Retrieved December 23, 2010 from the PROMIS website: http://www.nihpromis.org/Calibration%20Analysis%20Summaries/Calibration_Analysis_Summary_for_Fatigue.doc.

Lai, J.-S., & Chen, W.-H. (2006). *Fatigue archival analysis report*. Retrieved December 23, 2010 from the PROMIS website: http://www.nihpromis.org/Data%20Analysis/FatigueArchivalAnalysisReport.doc.

Lee, Y.-S. (2007). A Comparison of Methods for Nonparametric Estimation of Item Characteristic Curves for Binary Items. *Applied Psychological Measurement*, *31,* 121-134.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 519-537.

Lin, Y.-H., Kim, S.-H., & Cohen, A. S. (June, 2006). *Local dependence indices and detection investigation for polytomous items.* Paper presented at the International Meeting of the Psychometric Society, Montreal, Canada.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometricka, 47,* 149-174.

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10,* 355-366.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics, 11,* 204-209.

McDonald, R. P., & Mok, M. M. –C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30,* 23-40.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (RR ONR82-1). Iowa City: American College Testing Program.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23,* 187-194.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91-100.

Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.). *Computerized adaptive testing: From inquiry to operation* (pp. 169-179). Washington, DC: American Psychological Association.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*, 1-20.

Pommerich, M., & Ito, K. (2008). *An examination of the properties of local dependence measures when applied to adaptive data.* Annual meeting of the National Council on Measurement in Education, New York, NY.

Pommerich, M., & Segall, D. (2008). Local dependence in an operational CAT: Diagnosis and Implications. *Journal of Educational Measurement, 45*, 201-223.

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J.-S, & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care, 45*, 22-31.

Reeve, B. B., & Mâsse, L. C. (2004). Item response theory modeling for questionnaire evaluation. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: John Wiley & Sons, Inc.

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

Rock, D. A., & Nelson, J. (1992). Chapter 8: Applications and extensions of NAEP concepts and technology. *Journal of Educational and Behavioral Statistics, 17,* 219-232.

Roussos, L. A., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1-30.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph Supplement,* No. 17.

Scknipke, D. L., & Reese, L. M. (1997, March). *A comparison of testlet-based test designs for computerized adaptive testing.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording, and context*. New York: Academic Press.

Severo, N. C., & Zelen, M. Normal approximation to the chi-square and non-central F probability functions. *Biometrika, 47,* 411-416.

Singh, J., Howell, R. D., & Rhoads, G. K. (1990). Adaptive designs for Likert-type data: An approach for implementing marketing surveys. *Journal of Marketing Research*, *27,* 304-321.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of using visual design principles to group response options in web surveys. *International Journal of Internet Science, 1,* 6-16.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11,* 402-515.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selected in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485-518.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang. J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 21,* 195-213.

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory and local dependence: A preliminary report* (Research Memorandum 92-2). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 103-134). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16,* 109-119.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

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

Thompson, T. D., & Pommerich, M. (1996, April). *Examining the sources and effects of local dependence.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R. (1987). Attitude measurement: A cognitive perspective. In H. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 149-162). New York: Springer-Verlag.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103,* 299-314.

Tourangeau, R., Rips, L. J., Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.

Tsai, T. H., & Hsu, Y. C. (2005). *The use of information entropy as a local item dependence assessment.* Paper presented at Annual Meeting of the American Educational Research Association, Montreal Quebec Canada.

Tuerlinckx, F., and De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181-195.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181-196.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston: Kluwer.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42,* 283-302.

van der Linden, W. J., & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27,* 107-120.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Boston: Kluwer Academic.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22,* 259-270.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22,* 203-226.

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analogue for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C.

A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Dordrecht; Boston: Kluwer Academic.

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

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57,* 749-766.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15,* 22-29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203-220.

Wang, L., Fan, X., & Wilson, V. L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling, 3,* 228-247.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29,* 126-149.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 109-128.

Weiss, D. J. (1982) Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Weissman, A., (2002). Assessing the efficiency of item selection in computerized adaptive testing. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). *IRTGEN*: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 27,* 299-300.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Wise, S. L. (1997, March). *Overview of practical issues in a CAT program.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8,* 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30,* 187-213.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test, and ability statistics. MCAT Monograph, 5.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*, 291-309.