TWO MODELS FOR LONGITUDINAL ITEM RESPONSE DATA

Cheryl D. Hill

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology (Quantitative).

Chapel Hill
2006

Approved by

Advisor: David Thissen, Ph.D.

Reader: Patrick J. Curran, Ph.D.

Reader: Andrea M. Hussong, Ph.D.

Reader: Robert C. MacCallum, Ph.D.

Reader: Abigail T. Panter, Ph.D.

ABSTRACT

Cheryl D. Hill: Two Models for Longitudinal Item Response Data

(Under the direction of Dr. David Thissen)

Questionnaires are sometimes administered to the same sample of examinees on more than one occasion. Even when longitudinal data are available, researchers employing item response theory (IRT) often use data only from the first administration for item calibration because there is likely a lack of conditional independence between responses to the same item from the same individual. However, in many longitudinal study designs, the sample size at one occasion is too small for reliable item calibration. Thus, a longitudinal IRT model for use with repeated measures study designs is desirable.

This research develops two distinct approaches to longitudinal IRT. One of these models is based on latent class analysis, while the other is based on full-information bi-factor analysis. Both account for the local dependence among items that are administered twice by introducing parameters that describe how the repeated nature of each item affects the response (separately from the effect of the latent trait). The models include parameters that describe the latent trait distribution at the second administration relative to the standardized distribution at time one and the correlation between the latent traits at two time points. The addition of these model components allows item parameters to be calibrated using available data from two occasions.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. David Thissen, for mentoring me during my graduate career. He always had an answer to my endless questions, and his insight into my research was invaluable. I appreciate his willingness to provide support, be it in the form of a research assistantship, travel funding, or introductions to others in our field. I am so fortunate to have had the opportunity to work with one of the great minds in psychometrics.

Thank you to my committee members: Dr. Patrick Curran, Dr. Andrea Hussong, Dr. Robert MacCallum, and Dr. Abigail Panter. Each of them has a very demanding schedule, and I appreciate their taking the time to serve on my committee. They each brought a unique perspective to our discussions, and it was valuable having this diverse group to help shape my work.

My time as a graduate student in the L.L. Thurstone Psychometric Laboratory was a remarkable experience. The camaraderie among the students is something not found in many graduate programs, and I am so lucky to have had this group supporting me along the way. I especially want to thank Mike Edwards for taking me under his wing. Mike was always willing to lend an ear or offer a suggestion, and I could not have made it without him.

Thank you to my husband, Jeremy, who had incredible patience with me throughout this process. He is a wonderful partner, and I am looking forward to our next phase in life.

Finally, deep appreciation goes to my family. They ingrained in me at an early age the importance of education, and I thank them for giving me a drive to achieve. I would never have made it to this point without their generous support and encouragement.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Motivation for This Research

Questionnaires are sometimes administered to the same sample of examinees on more than one occasion. This is often the case in psychological studies or clinical trials in which the effect of a treatment is modeled longitudinally. Traditionally, even when longitudinal data are available, researchers using item response theory (IRT) to develop or score a measure use the data from the first test administration for item calibration. This inefficient use of the data is due to the likely lack of conditional independence between responses by the same individual, and the fact that implementations of IRT in software require conditional independence.

In many longitudinal study designs, the sample size is too small for reliable item calibration using data from one occasion. Thus, it is desirable that a longitudinal IRT model be developed for use with repeated measures study designs. Using data from more than one occasion may provide additional information, effectively transforming a sample that is too small into one that is adequate for item calibration.

1.2 Literature Review

Several methods exist that could be applied to longitudinal item response data, but none are ideal for the purposes of calibrating items with a longitudinal sample. One long-standing approach is multitrait-multimethod (MTMM) confirmatory factor analysis (CFA). The MTMM technique was developed by Campbell and Fiske (1959) as an approach to assessing the validity of psychological measures. This technique uses different methods to measure different traits and hypothesizes the strength of association between the methods and traits. An example is a measure with multiple subscales measuring multiple outcomes, on which it is expected that the subscale intended to measure a particular outcome is more associated with that outcome than it is with other outcomes, and this outcome should be measured better by its corresponding subscale than by any other subscale. Another example might involve three people responding to a measure for themselves and for each other, in which case the MTMM approach would hypothesize that a person's self-report responses would be more related to their trait than would be a friend's responses about that person, and an acquaintance's responses regarding that person would be even less related to that person's trait.

Kenny and Kashy (1992) suggested that MTMM methods could be evaluated using CFA, where the size of the factor loadings should vary in a predictable pattern. The model they propose that would apply to longitudinal response data is the correlated uniqueness model. In this model, there are multiple traits (i.e., trait factors), but the methods are represented with correlated unique factors across similar methods. For longitudinal item responses, the trait factors would be the latent trait of the measure over time. The unique factors for an item would be correlated over time, but disturbances for the other items would not correlate with that for the first over time. CFA parameter estimates can be converted into IRT parameters, so the results of an MTMM CFA approach to longitudinal data could be translated into item parameters and used in place of an IRT model.

While this approach is appropriate for longitudinal response data, the categorical nature of item response data can be problematic in the context of CFA. Categorical confirmatory factor analysis (CCFA) techniques exist and have improved considerably over the past two decades, but these techniques work best with large samples and simple models. Although this approach is reasonable, it would not be ideal for many longitudinal study designs involving small sample sizes.

Hierarchical linear modeling (HLM) has also been applied in some form to longitudinal item response data. HLM is appropriate for data in which responses are nested within a higher level, for example, children nested within schools. Researchers have proposed combining IRT models with HLM techniques so that the standard errors for the individual latent traits from the IRT model can be used as information available in the higher order model. Often these higher levels involve the variables of interest to researchers, and accounting for measurement error from the first level of the model can improve the accuracy of estimation in higher levels of the model. For the type of data considered here, individuals' responses are nested within time.

Some researchers have used longitudinal Rasch modeling with HLM techniques using penalized quasi-likelihood estimation (e.g., Pastor & Beretvas, 2006; Raudenbush & Sampson, 1999). In such models, the log odds is calculated for one fewer than the number of response categories for each item. Log odds models involve only a location parameter for each item, and assume a slope (or discrimination) parameter that is constant across items, so such models may not suit the needs of researchers who have longitudinal data from scales with varying levels of discrimination among the items.

Fox and Glas (2003) proposed a multilevel IRT model that uses the two-parameter logistic (2PL) model with slope and threshold parameters estimated separately for each item. This model is appropriate for many scales with dichotomous outcomes, and it can also be extended to polytomous responses. These researchers were able to include this more complicated IRT model in the HLM framework by replacing maximum likelihood estimation with Markov Chain Monte Carlo (MCMC) estimation. MCMC is a powerful estimation technique that can work for problems previously unsolvable with maximum likelihood estimation, but many applied researchers are not experienced with this complicated approach and may hesitate to use it.

Other researchers have applied traditional IRT techniques to repeated measures data by making relatively restrictive assumptions about the data. Ferrando, Lorenzo, and Molina (2001) considered the application of IRT to repeated measures, specifically to assess the stability of items over time. However, in the development of their model, they reason that the assumption of local independence is not violated because the latent trait is believed to be stable and the time between test administrations long enough for responses to be assumed independent of each other. Further, their model was designed to assess item stability over time when the items have already been calibrated. As a result, this model, while applicable to longitudinal data under certain conditions, is not relevant to researchers who hope to borrow information gained by multiple administrations for the purposes of item calibration.

Andrade and Tavares (2005) developed an IRT model for longitudinal data intended to estimate parameters of the population distribution, specifically the mean vector and the covariance matrix for multiple latent variables. Their model, however, worked under the assumption that the item parameters were already known (presumably, from some prior calibration). Thus, they did not consider item calibration with repeated measures.

Douglas (1999) developed an item response model that could handle the longitudinal nature of clinical trial data while simultaneously calibrating item parameters. In this model, the likelihood of a response pattern across time points is the product of the probability of the response to each item at each time; thus, independence between the items across the time points is assumed conditional on the entire vector of latent variable values across all times. This model is not an ideal approach because it ignores the possibility that local dependence (LD) appears between responses to the same item at different times. The dependence between the same items at different times may reflect an additional latent trait that is not accounted for in the one-trait-per-time design.

1.3 The Unique Contribution of This Research

The current literature on longitudinal item response data modeling covers some aspects of the problem of building an IRT model for repeated measures data; however, no complete solution has been offered. The existing models require the researcher to make assumptions (e.g., the item parameters are known or the latent trait is stable) that are unlikely to be true for many research designs. Alternative approaches are available, but the associated estimation procedures may not be appropriate for small longitudinal datasets. Areas such as personality measurement, educational testing, and clinical trials can benefit from a longitudinal IRT model that calibrates items while modeling change over time without requiring restrictive assumptions.

1.4 Specific Aims of This Research

The purpose of this research is to develop longitudinal IRT models for use in studies in which the participants are administered the same set of binary test items on two occasions. These models account for the local dependence among items that are administered multiple times by introducing parameters that describe how the repeated nature of each item affects the response (separately from the effect of the latent trait). Additionally, the models include parameters that describe the distribution of the latent trait at the second administration relative to the standardized distribution at time 1, and the correlation between the latent traits at two time points. The addition of these model components allows item parameters to be calibrated using available data from two occasions rather than limiting calibration to data from one time point.

1.4.1 Aim 1

The first set of research goals is to develop two approaches to longitudinal IRT, and to implement estimation algorithms for them. These models will be based on latent class analysis (LCA) and full-information bi-factor analysis, respectively. Maximum marginal likelihood parameter estimation will use the EM algorithm (Bock & Aitkin, 1981). The R statistical system (Ihaka & Gentleman, 1996) and C++ will be used to implement these estimation methods.

1.4.2 Aim 2

The second set of research goals is to check the parameter recovery of the algorithms using simulated data, and to evaluate the models using an empirical dataset. Parameter recovery will be assessed using simulated data of 100, 250, or 500 simulees with 5 or 10 dichotomous items administered twice. Results of this simulation will be compared to parameter estimates from the 2PL model for the LCA approach, and to parameter estimates from a CCFA model for the bi-factor analysis approach. If longitudinal IRT parameter recovery is successful, differences between the results of these models and the existing models will further be investigated using data from a psychological distress scale included on a longitudinal survey of adolescent substance abuse.

CHAPTER 2

PROPOSED MODELS

2.1 A Local Dependence Approach to Longitudinal IRT

Consider the probability of a particular response pattern for a set of items administered twice (without accounting for LD between administrations) as

$$P(\underline{u} \mid \underline{\theta}) = \int_{\theta_2}\int_{\theta_1} \prod_{t=1}^{2}\left(\prod_{i=1}^{I} T(u_{it} \mid \theta_t)\right)\Phi(\underline{\theta})\partial\theta_1\partial\theta_2 \,, \tag{1}$$

where $\underline{u}$ is a vector of responses, $\underline{\theta}$ is a vector of latent trait values at time 1 and time 2, t is administration occasion, and i is item number within an administration (I items per administration). For the 2PL model, useful for binary items that are not affected by guessing, the probability of endorsing an item is

$$T(u_{it} = 1 \mid \theta_t) = \frac{1}{1 + \exp(-a_i(\theta_t - b_i))}, \tag{2}$$

where a_i is the slope, or the strength of relationship between the response and the latent trait, and b_i is the threshold, or the location on θ_t where the examinee has a 50% probability of endorsing the item. Alternatively, the probability of not endorsing an item is

$$T(u_{it} = 0 \mid \theta_t) = 1 - T\left(u_{it} = 1 \mid \theta_t\right). \tag{3}$$

In (1), $\Phi(\underline{\theta})$ is the bivariate normal density, where

$$\begin{aligned}\Phi(\underline{\theta}) = {} & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 \text{-} \rho^2}} \cdot \\ & \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(\theta_1-\mu_1)(\theta_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(\theta_2-\mu_2)^2}{\sigma_2^2}\right]\right].\end{aligned} \tag{4}$$

Here, μ_t is the mean of θ_t, σ_t is the standard deviation of θ_t, and ρ is the correlation between θ_1 and θ_2.

The model in (1) does not account for the LD among responses to the same item across time. To account for such likely LD, some sort of LD parameter must be added for each item pair. In developing an LD index for use in IRT, Chen and Thissen (1997) introduce π_{LD}, the probability that the response to the second item is identical to the response to the first item without consideration of the latent trait. Alternatively, $1-\pi_{\mathrm{LD}}$ is the probability that the response to the second item is based solely upon the process implied by the IRT model without consideration of the response to the first item. When π_{LD} is 1, the second item provides no information about the latent trait that is not available through the response to the first item. When π_{LD} is 0, the two items each provide unique information about the latent trait. When π_{LD} is somewhere between 0 and 1, the second item provides some new information, but some of the information has already been captured through the response to the first item.

The same π_{LD} parameterization may be used to represent the LD between two administrations of the same item. In this context, each item will have an LD parameter, called κ_i. These LD parameters describe the probability that the item administered at time 2 contributes redundant information to what is available at time 1. Alternatively, $1-\kappa_i$ is the probability that the response at time 2 was determined by the IRT model, and this response contributes unique information. The κ_is balance the proportion of information that comes from the model representing items that contribute unique information at both administrations (i.e., the full IRT model) and information that comes from the model representing items that contribute unique information only at time 1 (i.e., the LD model).

Consider a simple 2-item test that is administered twice. For dichotomous items in which 1 represents a correct or endorsed response and 0 represents an incorrect or non-endorsed response, $\underline{u}$ is a vector that refers the responses to item 1 at time 1, item 2 at time 1, item 1 at time 2, and item 2 at time 2. There are 16, or $2^{2 \cdot I}$, possible response patterns:

- Neither item is repeated ($\underline{u}$ = 0011, 0110, 1001, or 1100);
- Item 1 is repeated but item 2 is not ($\underline{u}$ = 0001, 0100, 1011, or 1110);
- Item 2 is repeated but item 1 is not ($\underline{u}$ = 0010, 0111, 1000, or 1101); and
- Both items are repeated ($\underline{u}$ = 0000, 0101, 1010, or 1111).

When neither item is repeated, the examinee must have used the full IRT model when responding to each item at each time point (i.e., the LD model is not considered because neither item response was repeated).[1] Thus, the probability of each item response must be weighed by the probability of using the IRT model for both item 1 and item 2. This probability is written as

$$P(u_{11}, u_{21}, u_{12}, u_{22}) = (1-\kappa_1)(1-\kappa_2)P_{uuuu} . \qquad (5)$$

In equation (5), the first subscript on u refers to the item number, the second subscript refers to the time, κ_i is the probability that item i was repeated because of LD and not by chance, given the IRT model, and

$$P_{uuuu} = \int_{\theta_2}\int_{\theta_1} T(u_{11} \mid \theta_1) \cdot T(u_{21} \mid \theta_1) \cdot T(u_{12} \mid \theta_2) \cdot T(u_{22} \mid \theta_2) \cdot \Phi(\underline{\theta}) \partial\theta_1 \partial\theta_2 . \qquad (6)$$

The calculation of this probability involves all four item responses because they are (conditionally) independent of one another.

[1] In this model, only positive LD is included, where the response at time 2 is identical to response at time 1 because respondent chooses to repeat. An alternative model could incorporate negative LD, where the response at time 2 is different from response at time 1 because respondent chooses *not* to repeat, but such a model is not considered here.

When item 1 is repeated but item 2 is not repeated, the examinee must have used the full IRT model when responding to item 2 at each time point, but either the IRT model or the LD model may have been used when responding to item 1 at time 2. In other words, the response to item 1 may have been repeated because θ_2 led to the response, or because the response at time 1 was duplicated as the response at time 2. The probability equation for this response pattern must incorporate both the probability of using the IRT model for both item 1 and item 2 (first part of the sum), as well as the probability that the IRT model was used for item 2 and the LD model was used for item 1 (second part of the sum). This equation is

$$P(u_{11}, u_{21}, u_{12}, u_{22}) = (1-\kappa_1)(1-\kappa_2)P_{uuuu} + \kappa_1(1-\kappa_2)P_{uu\underline{\bar{x}}u}, \tag{7}$$

where

$$P_{uu\underline{\bar{x}}u} = \int_{\theta_2}\int_{\theta_1} T(u_{11} \mid \theta_1) \cdot T(u_{21} \mid \theta_1) \cdot 1 \cdot T(u_{22} \mid \theta_2) \cdot \Phi(\underline{\theta}) \partial\theta_1 \partial\theta_2 \,. \tag{8}$$

Here, $\underline{\bar{x}}$ represents the fact that the response to item 1 at time 2 does not factor in to the probability calculation. Thus, when a portion of the model represents the possibility that an item was repeated due to LD (e.g., $P_{uu\underline{\bar{x}}u}$), the probability of the response to that item at time 1 is included in the model but the probability of the response to that item at time 2 becomes 1 (i.e., the response at time 2 provides no information about θ_2).

A similar model is seen when the response to item 1 is not repeated but the response to item 2 is repeated, as in

$$P(u_{11}, u_{21}, u_{12}, u_{22}) = (1-\kappa_1)(1-\kappa_2)P_{uuuu} + (1-\kappa_1)\kappa_2 P_{uuu\underline{\bar{x}}} \tag{9}$$

Here, the response pattern probability is a weighted combination of the probability that all four item responses were based on the IRT model (first part of the sum) and the probability that the responses at time 1 and the response to item 1 at time 2 were based on the IRT model while the response to item 2 at time 2 was based on the LD model (second part of the sum). Again, this second probability portion excludes the probability of item 2 at time 2 because, if the response was based on the LD model, then it contributes no additional information about the latent trait, which is seen in

$$P_{uuu\underline{\bar{x}}} = \int_{\theta_2}\int_{\theta_1} T(u_{11} \mid \theta_1) \cdot T(u_{21} \mid \theta_1) \cdot T(u_{12} \mid \theta_2) \cdot 1 \cdot \Phi(\underline{\theta}) \partial\theta_1 \partial\theta_2 \ . \qquad (10)$$

Finally, if both items were repeated at time 2, then the probability of the response pattern is

$$\begin{aligned} P(u_{11}, u_{21}, u_{12}, u_{22}) = (1-\kappa_1)(1-\kappa_2)P_{uuuu} + \kappa_1(1-\kappa_2)P_{uu\underline{\bar{x}}u} \\ + (1-\kappa_1)\kappa_2 P_{uuu\underline{\bar{x}}} + \kappa_1\kappa_2 P_{uu\underline{\bar{x}}\underline{\bar{x}}} \end{aligned} \ , \qquad (11)$$

which is a weighted combination of the probability that the responses at time 2 were based on the IRT model (first part of the sum), the probability that the response to item 1 at time 2 was based on the LD model while the response to item 2 at time 2 was based on the IRT model (second part of the sum), the probability that the response to item 2 at time 2 was based on the LD model while the response to item 1 at time 2 was based on the IRT model (third part of the sum), and the probability that both responses at time 2 were based on the LD model (fourth part of the sum). Here, the fourth probability only includes the responses at time 1 in the model, as in

$$P_{uu\underline{\bar{x}}\underline{\bar{x}}} = \int_{\theta_2}\int_{\theta_1} T(u_{11} \mid \theta_1) \cdot T(u_{21} \mid \theta_1) \cdot 1 \cdot 1 \cdot \Phi(\underline{\theta}) \partial\theta_1 \partial\theta_2 \ . \qquad (12)$$

Because κ_i are probabilities, they cannot be smaller than 0 or larger than 1. It is a useful reasonableness test to consider this model as the κ_i parameters go to extremes. If κ_1 and κ_2 are both 0, then regardless of whether an item response is repeated or not, the response cannot be due to LD. The only term that remains in the four variations of the response probability equation is P_{uuuu}, which indicates that the probability of a response pattern is derived solely from the IRT model.

When either of the κ_i parameters are 1, the response to that item at time 2 is fully dependent on the response at time 1. The response at time 2 cannot be different from the response at time 1, so the eight possible response patterns in which that item is not repeated are not observed. For the eight remaining response patterns, the terms that account for the possibility that the repeated item response is due to the IRT model drop out of the equation (because $1-\kappa_i$ is 0).

When both κ_i parameters are 1, both responses at time 2 are fully dependent on the responses at time 1. The responses at time 2 cannot be different from the responses at time 1, so only the last four possible response patterns in which both items are repeated are observed. In this case, the response pattern probability equation is simply $P_{uu\underline{\bar{x}}\underline{\bar{x}}}$, which is the IRT model that includes only the responses at time 1.

To extend this approach to scales with I items (more than two), the probability of a response pattern, which involves comparing the response pattern, $\underline{u}$, to each $\underline{p}$ of the 2^I possible combinations of repeat patterns, is

$$P(\underline{u}) = \sum_{\underline{p}} \left[I\{A\}_{\underline{p}} \cdot \left[\prod_{i=1}^{I} I\{B\}_{\underline{p}i} \right] \int_{\theta_2} \int_{\theta_1} \left[\prod_{i=1}^{I} T\left(u_{i1} \middle| \theta_1\right) \right] \cdot \left[\prod_{i=1}^{I} T\left(u_{i2} \middle| \theta_2\right)^{I\{C\}_{\underline{p}i}} \right] \Phi(\underline{\theta}) \partial\theta_1 \partial\theta_2 \right], \quad (13)$$

where

$$I\{A\}_{\underline{p}} = \begin{cases} 1 & \text{if all items repeated in } \underline{p} \text{ are repeated in } \underline{u} \\ 0 & \text{otherwise} \end{cases}, \quad (14)$$

$$I\{B\}_{\underline{p}i} = \begin{cases} \kappa_{\mathrm{i}} & \text{if } i \text{ is repeated in } \underline{p} \\ 1-\kappa_{\mathrm{i}} & \text{otherwise} \end{cases}, \quad (15)$$

and

$$I\{C\}_{\underline{p}i} = \begin{cases} 0 & \text{if } i \text{ is repeated in } \underline{p} \text{ and } \underline{u} \\ 1 & \text{otherwise} \end{cases}. \tag{16}$$

In this generalized equation, indicator function *A* controls which of the possible probability portions are included in the total probability sum for that response pattern, indicator function *B* controls the LD weights for that probability portion, and indicator function *C* controls the inclusion of the item response probabilities in the time 2 product.

2.2 A Description of the Model as Latent Class Analysis

An alternative description of the model is as LCA. In the two items twice example, there are four latent classes of persons:

- The class that responds at time 2 by using the IRT model for both items ($\underline{u}$ = 0011, 0110, 1001, or 1100);
- The class that responds at time 2 by repeating the response to item 1 because of LD and using the IRT model for item 2 ($\underline{u}$ = 0001, 0100, 1011, or 1110);
- The class that responds at time 2 by repeating the response to item 2 because of LD and using the IRT model for item 1 ($\underline{u}$ = 0010, 0111, 1000, or 1101); and
- The class that responds at time 2 by repeating the responses to both items because of LD ($\underline{u}$ = 0000, 0101, 1010, or 1111).

The corresponding class probabilities can be represented as π_0, π_1, π_2, and π_3, respectively, and they must sum to 1. Because the parts of the models that are associated with these class probabilities are P_{uuuu}, $P_{uu\underline{\bar{x}u}}$, $P_{uuu\underline{\bar{x}}}$, and $P_{uu\underline{\bar{x}}\underline{\bar{x}}}$, respectively, the relation between the LD parameters and the LCA parameters is

$$\begin{aligned}\pi_0 &= (1-\kappa_1)(1-\kappa_2)\\ \pi_1 &= \kappa_1(1-\kappa_2)\\ \pi_2 &= (1-\kappa_1)\kappa_2\\ \pi_3 &= \kappa_1\kappa_2\end{aligned}\quad, \tag{17}$$

which imply that $\pi_1 + \pi_3 = \kappa_1$ and $\pi_2 + \pi_3 = \kappa_2$. The LCA parameters, $\pi_{\underline{p}}$, can be estimated using traditional LCA methods, and the κ_i parameters can be obtained from the values of $\pi_{\underline{p}}$. A general translation from κ_i to $\pi_{\underline{p}}$ is

$$\kappa_i = \sum_{\underline{p}} I\{D\}_{\underline{p}i} \cdot \pi_{\underline{p}}\ , \tag{18}$$

where

$$I\{D\}_{\underline{p}i} = \begin{cases} 1 & \text{if } i \text{ is repeated in } \underline{p} \\ 0 & \text{otherwise} \end{cases}. \tag{19}$$

2.3 An Alternative Model: A Bi-factor Analysis Approach to Longitudinal IRT

While the LCA approach to longitudinal IRT contains all of the components necessary for accounting for the LD among repeated items, it has the potential to create estimation problems. Each item has an additional parameter for LD, and three more parameters are included for the $\underline{\theta}$ distribution, so the number of parameters increases by over 50% as compared to a traditional 2PL model. More importantly, the E-step becomes computationally demanding because the probability that corresponds to each latent class must be calculated for each response pattern. Thus, the size of the problem grows exponentially when a test becomes long, and sparseness in the latent classes may make $\underline{\kappa}$ estimation difficult.

Because of these potential problems, a second approach to longitudinal IRT is considered. This approach borrows from full-information item bi-factor analysis (Gibbons & Hedeker, 1992). A bi-factor model is an approach to simplifying an s-dimensional model into a model with one primary dimension and s–1 secondary dimensions. Each item loads on the primary dimension and has a non-zero loading on no more than one secondary dimension. This bi-factor structure allows for simplified likelihood equations by reducing the integrations to two dimensions. Gibbons and Hedeker (1992) recognize that the bi-factor solution is an alternative model for tests with locally dependent items.

In the bi-factor analysis approach to longitudinal IRT, instead of one primary factor and a collection of secondary factors, the model includes two primary factors (one for each θ_t) and I secondary factors (one for each item). This bi-factor analysis approach permits the item parameters to be estimated using data from both time points (by constraining the item parameters to be equal within items across primary factors). Additionally, the LD is accounted for by the secondary factors that capture the relationship between the responses for each item pair at time 1 and time 2.

In this approach, the probability of response pattern $\underline{u}$ is

$$P(\underline{u}) = \int_{\underline{\theta}} \left\{ \prod_{t=1}^{2} \left[\prod_{j=3}^{F} \int_{\theta_j} \prod_{i=1}^{I} T(u_{it}|\theta_t, \theta_j)^{I\{E\}_{ij}} \right]\right] \right\} \Phi(\underline{\theta}) \partial \underline{\theta}. \tag{20}$$

Here, t is time, j is the secondary factor, F is the total number of factors ($F = 2 + I$),

$$T(u_{it} = 1|\theta_t, \theta_j) = \frac{1}{1 + \exp[-(a_{it}\theta_t + a_{ij}\theta_j + d_i)]}, \tag{21}$$

$$I\{E\}_{ij} = \begin{cases} 1 \text{ when } j = i + 2 \\ 0 \text{ otherwise} \end{cases}, \tag{22}$$

and

$$\Phi(\underline{\theta}) = \frac{1}{\sqrt{(2 \cdot \pi)^3 \cdot |\mathbf{\Sigma}|}} \cdot \exp\left[-\frac{1}{2} (\underline{\theta} - \underline{\mu})' \mathbf{\Sigma}^{-1} (\underline{\theta} - \underline{\mu}) \right], \tag{23}$$

where $\underline{\mu}$ is a vector of means for $\underline{\theta}$ and $\mathbf{\Sigma}$ is the covariance matrix for $\underline{\theta}$. In (21), $d_i = -(a_{it}b_{it} + a_{ij}b_{ij})$, combining the threshold parameters, that are not separately identified, into a single intercept parameter. Thus, in this alternative approach, there is one intercept parameter for each item and two slope parameters, one that corresponds to the construct of interest (e.g., ability or proficiency) and one that corresponds to the LD of that item. Indicator function *E* is used to ensure that the probability of item response u_{it} is only calculated for the secondary factor that corresponds to *i*, (i.e., θ_j or θ_{i+2}), which ensures simple structure on the secondary dimensions.

CHAPTER 3

ESTIMATION METHODS

The parameters of either of these models for longitudinal IRT can be estimated using direct maximum likelihood, which maximizes the function

$$l = \sum_{\underline{\mathrm{u}}} r_{\underline{u}} \log\left(\mathrm{P}\left(\underline{u}\right)\right), \tag{24}$$

where $r_{\underline{u}}$ is the observed number of examinees with response pattern $\underline{u}$. However, direct maximum likelihood is not a practical estimation method for most data problems because computing time becomes excessive as parameters are added to the model (i.e., as test length increases). An alternative approach uses Dempster, Laird, and Rubin’s (1977) EM algorithm, as refined by Bock and Aitkin (1981) for IRT, by Mooijaart and van der Heijden (1992) for LCA, and by Gibbons and Hedeker (1992) for item bi-factor analysis.

The EM algorithm consists of two iterative steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, the current values for the model parameters are used to estimate the number of people expected to be at each quadrature point, q, along $\underline{\theta}$, the proportion of these people expected to endorse each item, and, in the LCA approach, the proportion of people expected to be in a particular latent class. In the M-step, these estimates are used as if they were observed data to obtain updated parameters. These two steps alternate until some convergence criterion is reached. This process is described in detail below for each of the proposed approaches.

3.1 Latent Class Analysis Estimation

3.1.1 The E-step

For the IRT portion of the estimation procedure, the goal of the E-step is to estimate the number of people at each quadrature point who endorse each item given the current values of the model parameters. This information is stored in a series of $R*$ tables. R_{i1}^{*} contains the expected number of examinees at each quadrature point that would endorse item *i*, while R_{i0}^{*} contains the expected number of examinees at each quadrature point that would not endorse item *i*. For dichotomous items, there are two $R*$ tables for each item.

One of the goals in creating a longitudinal IRT model is to be able to combine data from two time points to calibrate one set of item parameters. Thus, the expected number of examinees at quadrature point *q* who respond correctly to item *i* is a combination of the examinees at quadrature point *q* at time 1 and the examinees at quadrature point *q* at time 2, as in

$$r_{qi1}^{*} = \sum_{\underline{u}} \left[\frac{r_{\underline{u}}}{\widetilde{P}_{\underline{u}}} \cdot \sum_{\underline{p}} \left[\begin{matrix} u_{i1} \cdot \sum\limits_{q_2}^{Q_2} L_{\underline{u}\,\underline{p}}\left(\theta_q, \theta_{q_2}\right) \cdot \Phi\left(\theta_q, \theta_{q_2}\right) + \\ I\{C\}_{\underline{pi}} \cdot u_{i2} \cdot \sum\limits_{q_1}^{Q_1} L_{\underline{u}\,\underline{p}}\left(\theta_{q_1}, \theta_q\right) \cdot \Phi\left(\theta_{q_1}, \theta_q\right) \end{matrix} \right] \right], \tag{25}$$

where

$$L_{\underline{u}\,\underline{p}}\left(\theta_{q_1}, \theta_{q_2}\right) = I\{A\}_{\underline{p}} \cdot \left[\prod_{i=1}^{I} I\{B\}_{\underline{pi}} \right] \cdot \left[\prod_{i=1}^{I} T\left(u_{i1} \middle| \theta_{q_1}\right) \right] \cdot \left[\prod_{i=1}^{I} T\left(u_{i2} \middle| \theta_{q_2}\right)^{I\{C\}_{\underline{pi}}} \right] \tag{26}$$

and

$$\widetilde{P}_{\underline{u}} = \sum_{\underline{p}} \sum_{q_2}^{Q_2} \sum_{q_1}^{Q_1} \left[L_{\underline{u}\,\underline{p}}\left(\theta_{q_1}, \theta_{q_2}\right) \right] \cdot \Phi\left(\theta_{q_1}, \theta_{q_2}\right). \tag{27}$$

Summation over quadrature points replaces integration over $\underline{\theta}$ in (13). $\widetilde{P}_{\underline{u}}$ is used to normalize the sum so that when it is multiplied by $r_{\underline{u}}$, the resulting value represents a number of persons.

For the time 1 responses, the values at each quadrature point of time 1 are summed across the quadrature points of time 2 and are included in the R_{i1}^{*} calculation when item *i* was endorsed. For the time 2 responses, the values at each quadrature point of time 2 are summed across the quadrature points of time 1, but they are only included in the R_{i1}^{*} calculation when that item at time 2 was endorsed and used in the calculation of the probability for a particular part of the model (as controlled by indicator function *C*).

Conversely, the R_{i0}^{*} calculation includes the response probability for item *i* when it is not endorsed at time 1, as well as when it is not endorsed but is used in the probability calculation at time 2, which is written as

$$r_{qi0}^{*} = \sum_{\underline{u}} \left[\frac{r_{\underline{u}}}{\widetilde{P}_{\underline{u}}} \cdot \sum_{\underline{p}} \left[\begin{array}{l} \left(1 - u_{i1}\right) \cdot \sum_{q_2}^{Q_2} L_{\underline{u}\,\underline{p}}\left(\theta_q, \theta_{q_2}\right) \cdot \Phi\left(\theta_q, \theta_{q_2}\right) + \\ I\{C\}_{\underline{p}i} \cdot \left(1 - u_{i2}\right) \cdot \sum_{q_1}^{Q_1} L_{\underline{u}\,\underline{p}}\left(\theta_{q_1}, \theta_q\right) \cdot \Phi\left(\theta_{q_1}, \theta_q\right) \end{array} \right] \right] \tag{28}$$

For the LCA portion of the estimation procedure, the goal of the E-step is to calculate the expected number of people with each response pattern in each latent class. This is calculated by

$$n_{\underline{u}\,\underline{p}}^{*} = \frac{r_{\underline{u}}}{\widetilde{P}_{\underline{u}}} \cdot \sum_{q_2}^{Q_2} \sum_{q_1}^{Q_1} L_{\underline{u}\,\underline{p}}\left(\theta_{q_1}, \theta_{q_2}\right) \cdot \Phi\left(\theta_{q_1}, \theta_{q_2}\right), \tag{29}$$

where the latent class parameters are in the form of $\pi_{\underline{p}}$ rather than $\prod_{i=1}^{I} I\{B\}_{\underline{p}i}$ and are represented in (26) as

$$L_{\underline{u}\,\underline{p}}\left(\theta_{q_1},\theta_{q_2}\right)=I\{A\}_{\underline{p}}\cdot\pi_{\underline{p}}\cdot\left[\prod_{i=1}^{I}T\left(u_{i1}\middle|\theta_{q_1}\right)\right]\cdot\left[\prod_{i=1}^{I}T\left(u_{i2}\middle|\theta_{q_2}\right)^{I\{C\}_{\underline{p}i}}\right] \tag{30}$$

For the distributional parameters of the population, the goal of the E-step is to create a matrix with the expected number of examinees at each quadrature point on 2-dimensional $\underline{\theta}$. This is calculated as the sum of each of the latent class probabilities times the observed number of examinees with that response pattern summed across response patterns, which is written as

$$N^*_{q_1q_2}=\sum_{\underline{u}}\left[\frac{r_{\underline{u}}}{\tilde{P}_{\underline{u}}}\cdot\sum_{\underline{p}}\left[L_{\underline{u}\,\underline{p}}\left(\theta_{q_1},\theta_{q_2}\right)\cdot\Phi\left(\theta_{q_1},\theta_{q_2}\right)\right]\right]. \tag{31}$$

3.1.2 The M-step

In the M-step, the item parameters, the population parameters, and latent class parameters are calculated using the estimated data created in the E-step. The calculation of the item parameters takes the same form as it does for traditional unidimensional 2PL EM estimation. The log of the M-step likelihood function is maximized to obtain estimates for the item parameters, which is written as

$$l_i=\sum_{q}^{Q}r^*_{qi1}\cdot\log\left(T_i\right)+\sum_{q}^{Q}r^*_{qi0}\cdot\log\left(1-T_i\right), \tag{32}$$

where T_i is from (2) and the r^*s are from (25) and (28). For each item, estimates of the item parameters are obtained where the first derivative of (32) is equal to zero, and change is monitored through the values of the second derivatives (Bock & Aitkin, 1981). These computations are better-conditioned when the trace line function is parameterized in slope-intercept form, rather than the typical slope-threshold form presented in (2). Thus, T_i is written as

$$T_i = \frac{1}{1 + \exp\left(-\left(a_i\theta_t + d_i\right)\right)}, \tag{33}$$

where d_i is $-a_ib_i$. Using this slope-intercept form, the first derivative for the slope parameter is

$$\frac{\partial l}{\partial a_i} = \sum_q^Q \left(\theta_q \cdot \left(r^*_{qi1} - N^*_{qi} \cdot T_i\right)\right), \tag{34}$$

and the first derivative for the intercept parameter is

$$\frac{\partial l}{\partial d_i} = \sum_q^Q \left(r^*_{qi1} - N^*_{qi} \cdot T_i\right), \tag{35}$$

where

$$N^*_{qi} = r^*_{qi1} + r^*_{qi0}. \tag{36}$$

The second derivative for the slope parameter is

$$\frac{\partial^2 l}{\partial a_i^2} = \sum_q^Q \left[-\theta_q^2 \cdot N^*_{qi} \cdot T_i \cdot \left(1 - T_i\right)\right], \tag{37}$$

the second derivative for the intercept parameter is

$$\frac{\partial^2 l}{\partial d_i^2} = \sum_q^Q \left[-N^*_{qi} \cdot T_i \cdot \left(1 - T_i\right)\right], \tag{38}$$

and second cross derivative for the slope parameter and intercept parameter is

$$\frac{\partial^2 l}{\partial a_i \partial d_i} = \sum_q^Q \left[-\theta_q \cdot N^*_{qi} \cdot T_i \cdot \left(1 - T_i\right)\right] \tag{39}$$

The M-step calculation of the latent class parameter uses an approach suggested by Mooijaart and van der Heijden (1992), which is

$$\hat{\pi}_{\underline{p}} = \sum_{\underline{u}} \frac{n^*_{\underline{u}\underline{p}}}{N}, \tag{40}$$

where $n^{*}_{\underline{u}\underline{p}}$ is from (29) and N is the total number of examinees. The latent class parameter estimates, $\hat{\pi}_{\underline{p}}$,are then transformed into local dependence parameters, $\hat{\kappa}_{\underline{i}}$, by (18), which are used in the subsequent E step.

The M-step calculation of the population parameters uses the expected number of examinees at each quadrature point, $N^{*}_{q_1 q_2}$ from (31). The mean for the population at time t is

$$\mu_t = \frac{\sum_{q_1}^{Q_1}\sum_{q_2}^{Q_2} N^{*}_{q_1 q_2} \cdot \theta_{q_t}}{\sum_{q_1}^{Q_1}\sum_{q_2}^{Q_2} N^{*}_{q_1 q_2}}, \tag{41}$$

the standard deviation for the population at time t is

$$\sigma_t = \sqrt{\frac{\sum_{q_1}^{Q_1}\sum_{q_2}^{Q_2} N^{*}_{q_1 q_2} \cdot \left(\theta_{q_t} - \mu_t\right)^2}{\sum_{q_1}^{Q_1}\sum_{q_2}^{Q_2} N^{*}_{q_1 q_2}}}, \tag{42}$$

and the population covariance is

$$\rho = \frac{\dfrac{\sum_{q_1}^{Q_1}\sum_{q_2}^{Q_2} N^{*}_{q_1 q_2} \cdot \theta_{q_1} \cdot \theta_{q_2}}{\sum_{q_1}^{Q_1}\sum_{q_2}^{Q_2} N^{*}_{q_1 q_2}} - \mu_1 \cdot \mu_2}{\sigma_1 \cdot \sigma_2} \tag{43}$$

as formulated by Bock (1985). The values for μ_2, σ_2, and ρ are then used as estimates in the next E-step, while the values for μ_1 and σ_1 are replaced with fixed values, usually 0 and 1, respectively, to identify the scale of the latent variables.

3.2 Bi-factor Analysis Estimation

3.2.1 The E-step

In the E-step, again, the number of people at each quadrature point who endorse each item given the current values of the model parameters is estimated and stored in *R** tables.

However, the resulting *R** tables are 2-dimensional (θ_t x θ_j), where the time 1 and time 2 responses are collapsed into one primary factor, θ_t. Repeated responses do not affect the inclusion or exclusion of data entered into the *R** tables in the bi-factor approach.

The expected number of examinees at quadrature point q_t x q_j who respond correctly to item *i* is a combination of the examinees at quadrature point q_1 x q_j (at time 1) and the examinees at quadrature point q_2 x q_j (at time 2), which is written as

$$r^*_{q_t q_j i1} = \sum_{\underline{u}} r_{\underline{u}} \cdot \left[\begin{array}{l} \displaystyle\sum_{q_2}^{Q_2} u_{i1} \cdot \frac{L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right) \cdot \Phi\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right)}{\widetilde{P}_{\underline{u}}} + \\ \displaystyle\sum_{q_1}^{Q_1} u_{i2} \cdot \frac{L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right) \cdot \Phi\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right)}{\widetilde{P}_{\underline{u}}} \end{array} \right], \tag{44}$$

where

$$L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right) = \prod_{t=1}^{2} \left[\prod_{j=3}^{F} \prod_{i=1}^{I} T_i\left(\theta_{q_t},\theta_{q_j}\right)^{I\{E\}_{ij}} \right], \tag{45}$$

and

$$\widetilde{P}_{\underline{u}} = \sum_{q_j}^{Q_j} \sum_{q_2}^{Q_2} \sum_{q_1}^{Q_1} \left[L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right)\right] \cdot \Phi\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right). \tag{46}$$

Thus, for each response pattern, a 3-dimensional array of probabilities, representing a grid of points in the θ_1 x θ_2 x θ_j space, is calculated. The array of probabilities is multiplied by the response at time 1 (i.e., the probabilities are included when the response is correct) and summed across the quadrature points of time 2. Added to those values is the same 3-dimensional array of probabilities multiplied by the response at time 2 and summed across the quadrature points of time 1. The resulting 2-dimensional array is then multiplied by the observed number of examinees with that response pattern, and these arrays are summed across response patterns.

A similar calculation is used for the expected number of examinees at quadrature point q_t x q_j who respond incorrectly to item i, where probabilities are included for incorrect responses, which is written as

$$r^{*}_{q_t q_j i0} = \sum_{\underline{u}} r_{\underline{u}} \cdot \left[\begin{array}{l} \sum\limits_{q_2}^{Q_2} \left(1-u_{i1}\right) \cdot \dfrac{L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right) \cdot \Phi\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right)}{\widetilde{P}_{\underline{u}}} + \\ \sum\limits_{q_1}^{Q_1} \left(1-u_{i2}\right) \cdot \dfrac{L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right) \cdot \Phi\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right)}{\widetilde{P}_{\underline{u}}} \end{array} \right] \qquad (47)$$

Additionally in the E-step, information about the $\underline{\theta}$ distribution is obtained for use in the M-step. The 3-dimensional array of probabilities is multiplied by the observed number of examinees with that response pattern and summed across the quadrature points of the secondary factor, θ_j. The resulting 2-dimensional array is summed across response patterns, creating an array of expected counts at each quadrature point of $\theta_{1,2}$, as in

$$N^{*}_{q_1 q_2} = \sum_{\underline{u}} \left[\sum_{q_j}^{Q_j} r_{\underline{u}} \cdot \frac{L_{\underline{u}}\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right) \cdot \Phi\left(\theta_{q_1},\theta_{q_2},\theta_{q_j}\right)}{\widetilde{P}_{\underline{u}}} \right], \qquad (48)$$

3.2.2 The M-step

The M-step in the bi-factor approach is similar to the M-step in the LCA approach. For the item parameters, maximum likelihood multiple logit analysis is employed with the *R** tables from the E-step. Again, the log of the likelihood function is maximized to obtain estimates for the item parameters, which is written as

$$l_i = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} r^{*}_{q_t q_j i1} \cdot \log\left(T_i\right) + \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} r^{*}_{q_t q_j i0} \cdot \log\left(1 - T_i\right), \qquad (49)$$

where T_i is from (21) and the r*s are from (44) and (47). For each item, estimates of the item parameters are obtained where the first derivative of (49) is equal to zero, and change is monitored through the values of the second derivatives (Bock & Aitkin, 1981).

The first derivative for either of the slope parameters is

$$\frac{\partial l}{\partial a_{fi}} = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} \left(\theta_{q_t q_j} \cdot \left(r^*_{q_t q_j i 1} - N^*_{q_t q_j i} \cdot T_i \right) \right), \tag{50}$$

and the first derivative for the intercept parameter is

$$\frac{\partial l}{\partial d_i} = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} \left(r^*_{q_t q_j i 1} - N^*_{q_t q_j} \cdot T_i \right), \tag{51}$$

where

$$N^*_{q_t q_j i} = r^*_{q_t q_j i 1} + r^*_{q_t q_j i 0} . \tag{52}$$

The second derivative for either of the slope parameters is

$$\frac{\partial l^2}{\partial a^2_{fi}} = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} \left(-\theta^2_{q_t q_j} \cdot N^*_{q_t q_j i} \cdot T_i \cdot (1 - T_i) \right), \tag{53}$$

the second derivative for the intercept parameter is

$$\frac{\partial l^2}{\partial d^2_i} = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} \left(-N^*_{q_t q_j i} \cdot T_i \cdot (1 - T_i) \right), \tag{54}$$

the second cross derivative for both slope parameters is

$$\frac{\partial l^2}{\partial a_1 \partial a_2} = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} \left(-\theta_{q_1 q_j} \cdot \theta_{q_2 q_j} \cdot N^*_{q_t q_j i} \cdot T_i \cdot (1 - T_i) \right), \tag{55}$$

and the second cross derivative for either of the slope parameters and the intercept parameter is

$$\frac{\partial l^2}{\partial a_{fi} \partial d_i} = \sum_{q_t}^{Q_t} \sum_{q_j}^{Q_j} \left(-\theta_{q_t q_j} \cdot N^*_{q_t q_j i} \cdot T_i \cdot (1 - T_i) \right). \tag{56}$$

For the population parameters, the same equations that were used for obtaining μ_2, σ_2, and ρ in the LCA approach are also used in the bi-factor analysis approach.

3.3 Bi-factor Analysis in Limited-Information Item Factor Analysis

Because the item bi-factor analysis model is a full-information approach to Holzinger and Swineford's (1937) bi-factor method, it is sensible to consider this approach in a CCFA framework as mentioned in the introduction in the context of MTMM techniques. This model is depicted as a structural equation model in Figure 1 for the simple case of two items administered twice. This model includes two factors for the two administrations, θ_1 and θ_2, and the items load only on the factor that corresponds to the time at which they were administered. Factor loadings for the same item at different time points are constrained to be equal so that each item has one loading on the outcome of interest. The first factor is standardized with a mean of 0 and a variance of 1, while the second factor has a mean of α_2 and a variance of φ_{22}. The two factors have a covariance of φ_{21}.

The LD between the same item at different time points is captured in an error covariance (or correlation when the measured variables are standardized) between the two items. This error covariance could also be described as the square of the factor loading if each item pair loaded on its own LD factor with the two factor loadings constrained to be equal. Both parameterizations highlight the fact that the LD describes variance that the item pairs have in common above and beyond what is explained by the primary construct of interest.

Not depicted in Figure 1 is the threshold parameter, τ_i, for each item. This parameter is similar to the threshold parameter of the IRT model in that it specifies the location on the latent construct above which the item is endorsed and below which the item is not endorsed.

For the relationship between the IRT parameters in (21) and CCFA parameters, the factor loading, λ_i, can be translated into a slope for the primary factor, a_{it}, by the equation

$$a_{it} = \frac{1.7 \cdot \lambda_i}{\sqrt{1-\left(\lambda_i^2+\delta_i\right)}}. \qquad (57)$$

The covariance between the responses to the same item at time 1 and time 2, δ_i, can be translated into a slope for the secondary factor, a_{ij}, by the equation

$$a_{ij} = \frac{1.7 \cdot \sqrt{\delta_i}}{\sqrt{1-\left(\lambda_i^2 + \delta_i\right)}}. \tag{58}$$

The threshold, τ_i, can be translated into an intercept, d_i, by the equation

$$d_i = \frac{-\tau_i}{\sqrt{1-\left(\lambda_i^2 + \delta_i\right)}}. \tag{59}$$

In these equations, the value 1.7 converts the estimates from the normal metric to the logistic metric which is specified in (21) (McLeod, Swygert, & Thissen, 2001).

This CCFA model can be estimated using the latent variable software M*plus* (Muthén & Muthén, 2003) using weighted least squares (WLS) estimation when the sample is large or using diagonally weighted least squares with a mean- (and variance-) adjusted chi-square test statistic (WLSM/V) when the sample is small (Oranje, 2003). The CCFA converted parameters should be similar to the IRT parameters when samples are large, but some differences may be observed because IRT is a full-information estimation method and CCFA is a limited-information estimation method.

CHAPTER 4

METHODS FOR EVALUATING THESE APPROACHES

When a new model is introduced, it is important to ask two questions: (1) can the parameters be estimated, and (2) are the estimates interpretable? These questions were considered for each of the two proposed approaches to modeling longitudinal IRT data. The first question was answered by using simulated data to evaluate the parameter recovery of each model. By using simulated data in which the true parameter values are known, the estimates can be compared to the true values to determine how well the algorithms estimate the parameters. The second question was answered using empirical data. The content of the items can indicate how stable the responses and the latent trait should be over time, and parameter estimates can be examined to determine if the hypothesized properties are revealed in the magnitude and sign of the estimates.

4.1 Parameter Recovery with Simulated Data

4.1.1 Simulation Parameters and Methods

First, data were simulated in the R statistical system for a large number of simulees (N = 5000) and a small number of dichotomous items (I = 4) at two time points. These data were used to evaluate the estimation procedures to ensure that they had been implemented correctly in C++. Because the sample was very large, the parameter estimates are expected to be very close to the true values. An additional use of these data was to explore if the two models produce comparable results. Data generated with one model were fit with the alternative model as an attempt to identify a link between the two approaches by comparing the parameter estimates.

Once the estimation methods were verified, data were simulated with smaller numbers of simulees ($N = 100$, 250, or 500) and larger numbers of items ($I = 5$ or 10). These conditions were chosen because they are comparable to the characteristics of data collected in longitudinal study designs. The parameter estimates were compared to the true values to evaluate if the model can capture the parameters under these realistic conditions. It is important to demonstrate that the longitudinal IRT approach can recover the true parameter values for the data problems for which it is intended (i.e., many items with limited examinees).

Further, it is desirable that the proposed models outperform existing methods in terms of parameter recovery. The parameter estimates of the LCA model were compared to those of the unidimensional 2PL model using data from one time point, and the parameter estimates of the bi-factor model were compared to that of the limited-information CCFA model using data from both time points.

For the LCA approach, true theta values were drawn from a 2-dimensional normal distribution with a mean of 0 and a variance of 1 at time 1, a mean of 0.2 and a variance of 1 at time 2, and a correlation of 0.5 between the two times. True a values varied between 1 and 2, true b values varied between -1 and 1, and true κ values varied between 0.1 and 0.4.

To simulate the response at time 1, the probability of endorsement is calculated using (2) with the person's true θ_1 value and the slope and threshold for that item. A random number is drawn from a rectangular distribution between 0 and 1, and if this random number is less than the calculated probability, then the response is scored as positive. If the random number is greater than or equal to the calculated probability, then the response is scored as negative. A provisional response at time 2 is simulated in the same manner using a new random number and a probability calculated with the person's true θ_2 value. Because the κ parameters can be considered as probabilities of repeating the response from time 1 at time 2, another random number is drawn from a rectangular distribution between 0 and 1. If the random number is less than the κ value for that item, then the provisional response at time 2 is replaced with the person's response at time 1. If the random number is greater than the κ value, then the provisional response is retained as the response at time 2.

For the bi-factor analysis approach, true theta values were drawn from a 3-dimensional normal distribution with a mean of 0 and a variance of 1 at time 1 and for the specific factor, and a mean of 0.2 and a variance of 1 at time 2. The correlation between θ_1 and θ_2 was .5, while θ_1 and θ_2 did not covary with the specific factor. True *a* values on the dimension of interest varied between 1 and 2, true *a* values on the LD dimension varied between 0.5 and 1.5, and true *b* values varied between -1 and 1 (thresholds were translated into intercepts given the slope values).

To simulate the response at time 1, the probability of a correct response is calculated using (21) with the person's true θ_1 and true θ_j, and the slopes and intercept for that item. A random number is drawn from a rectangular distribution between 0 and 1, and if this random number is less than the calculated probability, then the response is scored as correct. If the random number is greater than or equal to the calculated probability, then the response is scored as incorrect. The response at time 2 is simulated using the same process by calculating the probability with θ_2 instead of θ_1 and using a new random number draw.

Within the estimation software, conservative starting values were chosen for the parameters. The starting values for the distributional parameters were 0 for the mean at time 2, 1 for the variance at time 2, and 0.75 for the correlation between the two time points. For the LCA approach, *a* values started at 1, *b* values started at 0, and *κ* values started at 0.1. For the bi-factor approach, a_1 values started at 1, a_2 values started at 0.5, and *d* values started at 0. Because little was known about the properties of the likelihood surface for these models, strict convergence criteria were chosen to provide the routine ample opportunity to "climb" to the maximum value. Ten-thousand cycles were allowed, with estimation ending if the maximum change in estimated values from one cycle to the next dropped below 1.0e–07.

4.1.2 2PL Method

Parameter estimates from the LCA model were compared to those obtained using the 2PL model in Multilog (Thissen, Chen, & Bock, 2003). Data from the first administration alone were used for parameter calibration. Distributional statistics were calculated in SAS 9.1 (SAS Institute, Inc., 2005) using the sum of the item responses within administration and standardizing the sample statistics at time 2 using the statistics from time 1.

4.1.3 CCFA Method

Parameter estimates from the bi-factor model were compared to those obtained using a longitudinal CCFA model in M*plus* 3.13 (Muthén & Muthén, 2003). Although WLS estimation is the gold-standard approach to CCFA model estimation, it is known to perform poorly with small samples or complex models (Oranje, 2003). Anticipating these conditions, WLSM/V estimation was used when the sample size was 500 or less. When data conditions are appropriate for WLS, the WLS parameter estimates should be similar to those obtained with WLSM/V. Thus, even if WLSM/V estimation was used for data with which WLS would be suitable, no error should be induced in the parameter estimates.

The model depicted in Figure 1 was specified in M*plus* by letting each item load on the latent factor particular to its administration time, while constraining the factor loadings to be consistent across time within item. Means and thresholds were included in the model by specifying a mean structure analysis, and thresholds were also constrained equal across time within item. Error correlations were introduced between each item pair. Theta parameterization, which allows the residual variances to be estimated in the model, was used. Each analysis was conducted on an inter-item tetrachoric correlation matrix by indicating that the data were categorical. Standardized estimates for factor loadings, error correlations, and thresholds were used in comparing CCFA results to bi-factor method results because IRT assumes that the underlying response variable is standard normal. CCFA estimates were translated into the IRT metric, and standard errors were translated using the delta method.

4.2 Model Evaluation with Empirical Data

Empirical data from the Understanding Adolescent Health Risk Behaviors survey from the study *The Context of Adolescent Substance Use* (the Context Study) were used for evaluating the interpretability of these proposed models, as well as to compare the results obtained through the proposed methods to results available through existing methods (NIDA Grant No. R01 DA13459).[2] The Context Study collects data longitudinally on adolescents in grades 6 through 12 in the fall and spring of each year. Items on the survey assess tobacco, alcohol, and drug use, aggression, and family relationships. For the purpose of this research, a set of items measuring behaviors indicative of psychological distress were chosen. These 10 items

[2] Great appreciation goes to Dr. Susan Ennett for making this substantial dataset available for these analyses.

were measured on a 5-point Likert response scale ranging from “Strongly agree” to “Strongly disagree”, and the item text is presented in Figure 2.

Data from adolescents at the first fall administration and the second spring administration (waves 2 and 3 of data collection) were chosen for the present analyses. Only participants who had no missing data on these 10 items across the two time points were included, which resulted in a sample of 3,788. This sample was 53% female and 57% white, with an average age of 13.5 (*sd* = 0.97) for the fall administration and 14.0 (*sd* = 0.97) at the spring administration.

Prior to using these data to assess the proposed models, the dimensionality of this 10-item scale was assessed through a 1-factor CCFA model in M*plus*. Local item dependencies were examined using modification indices for the error correlations between the measured variables. It was established that several of the items were locally dependent, and the model was trimmed until one set of unidimensional, locally independent items was identified. After this item reduction, items 1, 2, 7 and 9 remained, and all subsequent analyses were conducted using this set of 4 items.

Because the proposed models have been parameterized for binary data, it was necessary to dichotomize the Likert response scale. “Strongly agree”, “Agree somewhat”, and “Neither” were grouped together to indicate endorsement, and “Disagree somewhat” and “Strongly disagree” were group together to indicate non-endorsement. While these items would be best analyzed using their original response categories, this categorization is appropriate for these demonstrative analyses.

CHAPTER 5

RESULTS

5.1 Simulated Data

5.1.1 Latent Class Analysis

The programming of the LCA approach was evaluated using a simulated dataset of 5000 simulees and 4 items. The generating values, the parameter estimates, and the difference between the estimated and true values are presented in Table 1. The estimation procedure appears to work well for the LCA model, with differences between the estimated value and the true value being no more than 0.1 for the slope parameter, no more than 0.04 for the threshold parameter, and no more than 0.03 for the LD parameter. The difference between the estimated and true values was 0.01 for the correlation between θ_1 and θ_2, 0.03 for the mean of θ_2, and 0.02 for the standard deviation of θ_2. These differences are very small and indicate good parameter recovery.

With the LCA algorithm implementation evaluated as correct, the analyses proceeded to evaluating parameter recovery under realistic data conditions. It is important to stress that parameter recovery was evaluated using one simulated dataset per data condition. Large parameter recovery simulation studies often use 1000 or more simulated datasets per condition and summarize results across those datasets using statistics like bias and root mean square error. Such large simulations aim to reduce the effect of sampling error using these summary statistics. Some of the simulated data may be drawn from extreme parts of the underlying distribution, creating datasets with sample statistics different from the population values. Aggregating over the whole set of data for one condition allows the true recovery to stand out while minimizing the effect of the odd samples. However, the present parameter recovery examination is intended to evaluate the potential of the proposed approaches, not to declare one method superior to another under specific conditions. If one or both approaches show value, then a larger simulation study to determine the conditions under which these methods are appropriate would be a logical next research step.

Responses to 5 items for samples of 100, 250, and 500 simulees were generated using the properties of the LCA model, and the results are presented in tables that include the true generating values, the estimates from the LCA model and the 2PL model, and the difference between the estimated values and the true values. For the sample of 100 simulees in Table 2, the LCA model had a smaller range for the difference between the estimates and true values as compared to the 2PL model. For the slope parameter, differences for the LCA model ranged from –0.1 to +1.4, while differences for the 2PL model ranged from –0.3 to +2.3. Both models overestimated a_4, so this large range ceiling is probably due to an oddity in the sampling for that parameter rather than being a true error. For the threshold parameter, differences for both models ranged from –0.2 to +0.3. For the LD parameter, estimated and true values differed from –0.1 to +0.01. Considering the small sample size, the errors for each of the parameter estimates are reasonable for both models. Each model also captured the mean and standard deviation of the latent trait at time 2, but the 2PL model showed more error in the correlation between latent traits than did the LCA model (+0.3 versus +0.2).

The estimates for the 250 simulees in Table 3 had less error than did those for the sample of 100, while the 2PL model showed slightly narrower error ranges. For the slope parameter, the differences ranged from –0.2 to +0.6 for the LCA model and –0.1 to +0.6 for the 2PL model, and for the threshold parameter, the differences ranged from –0.1 to +0.2 for the LCA model and 0 to +0.2 for the 2PL model. Again, both models captured the true parameter values suitably, and the more extreme difference in slope estimates was consistent between models, indicating randomness in sampling. The LD parameter estimates differed by –0.1 to +0.2 as compared to the true values. As in the 100 simulee dataset, both models captured the mean and standard deviation of the time 2 latent trait, but the 2PL model showed much more error in the correlation as compared to the LCA model (+0.4 versus +0.1).

In Table 4, the estimates for the sample of 500 simulees were similar to those for 250 simulees. Slope parameter errors ranged from –0.1 to +0.4 for the LCA model and from –0.2 to +0.6 for the 2PL model. Threshold parameter errors ranged from –0.2 to +0.2 for both models. The errors in the LD estimates ranged from –0.1 to +0.02. Again, both models provided good estimates for the mean and standard deviation of the latent trait at time 2, but the 2PL model produced a poorer estimate for the correlation between the latent traits (+0.3 versus –0.1). The consistency between the sample with 250 simulees and 500 simulees suggests that for 5 items, sample size greater than 250 has little impact on parameter recovery. Even with a small sample of 100 simulees, both models recovered the true parameters acceptably, with the LCA model producing more accurate estimates than the 2PL model.

To examine the effect of test length on parameter recovery, samples of the same size were simulated for tests with 10 items. For these 10-item results, error in the parameter estimates was summarized with mean square error (MSE), where the difference between the estimated value and the true value is squared and averaged across the 10 items for that model. In Table 5, results are presented for the 100 simulee sample. Slope parameter error ranged from –0.7

to +0.8 (MSE = 0.17) for the LCA model and from –0.2 to +0.8 (MSE = 0.14) for the 2PL model. Threshold parameter error ranged from –0.4 to +0.3 (MSE = 0.04) for the LCA model and from –0.3 to +0.5 (MSE = 0.05) for the 2PL model. Neither model recovered any parameter perfectly, but the MSE was relatively small for the slope parameter and very small for the threshold parameter. Surprisingly, there was little difference between the two models in terms of parameter recovery of 2PL parameters. For the LD parameter, error ranged from –0.1 to +0.2 (MSE = 0.01), which is no worse than the 5-item samples. The two models showed little error in the θ_2 mean (no error for the LCA model, –0.05 for the 2PL model), but they showed more error in the θ_2 standard deviation (–0.21 for the LCA model, –0.17 for the 2PL model). As with the 5-item samples, the 2PL model poorly recovered the correlation parameter (error of +0.25) as compared to the LCA model (error of –0.05).

The results for the sample of 250 simulees in Table 6 showed better parameter recovery for both models. The slope parameter error for the LCA model ranged from –0.3 to +0.3 (MSE = 0.04) and ranged from –0.5 to +0.6 (MSE = 0.12) for the 2PL model. The threshold parameter error ranged from –0.1 to +0.2 (MSE = 0.01) for the LCA model and from –0.2 to +0.2 (MSE = 0.01) for the 2PL model. These ranges and MSE values show that both models recovered the 2PL parameters well with a sample of 250, with the LCA model providing more accuracy in the slope parameters than the 2PL model. The error in the LD parameters ranged from –0.04 to +0.1 (MSE = 0.003). This sample exhibited the same trend of good estimation for the mean and standard deviation of θ_2, with the 2PL model failing to recover the correlation as well as the LCA model (+0.3 versus +0.1).

The results obtained with a sample of 500 simulees in Table 7 continued to show reduced error with increased sample size. The slope parameter error ranged from –0.4 to +0.1

(MSE = 0.04) for the LCA model and from –0.4 to +0.3 (MSE = 0.07) for the 2PL model. The threshold parameter error ranged from –0.3 to +0.02 (MSE = 0.04) for the LCA model and from –0.3 to 0 (MSE = 0.03) for the 2PL model. With a sample of 500, the 2PL parameters are recovered well using either model. The error for the LD parameters ranged from –0.04 to +0.06 (MSE = 0.001). The distributional parameters were recovered in the same manner as before, with the main difference being in the correlation for the 2PL model (+0.3) and the LCA model (–0.02).

Overall for the LCA model, parameter recovery is quite good. The approach shows promise for small samples and longer tests. Although there were no drastic differences in slope and threshold recovery between the LCA model and the 2PL model, the LCA model outperformed the 2PL model in distributional parameter recovery with its ability to include the distributional parameters as part of the full-information model. The LCA model also had success in recovering the LD parameters across all data conditions.

5.1.2 Bi-factor Analysis

As with the LCA model, prior to examining parameter recovery of the bi-factor model, the implementation of the EM algorithm was verified with a sample of 5000 simulees and 4 items. Results are presented in Table 8. Error in the a_1 estimates ranged from –0.1 to –0.01, in the a_2 estimates ranged from +0.03 to +0.2, and in the d estimates ranged from –0.1 to –0.02. These differences between the estimated values and true values are small and imply correct programming. Errors of estimation were also small for the distributional parameters, –0.02 for the θ_2 mean, –0.04 for the θ_2 standard deviation, and –0.07 for the correlation between θ_1 and θ_2.

With the programming verified, analyses switched to parameter recovery evaluation using samples generated with the bi-factor model with the same characteristics as those generated using the LCA model. Responses to a 5-item test were simulated for samples of 100, 250, and 500. Results for the 100 simulee sample are presented in Table 9. Estimation errors in the slopes on the primary dimension ranged from –0.4 to +3.0 for the bi-factor model and from +0.02 to +2.9 for the CCFA model. The items for which the largest difference between the estimates and the true values were observed were not consistent between models, suggesting that the error is not a result of an odd sample. The errors for the LD slopes ranged from –0.4 to +0.6 for the bi-factor model and from –0.3 to +0.6 for the CCFA model. The difference between the estimated intercept values and their true values ranged from –0.7 to +0.7 for the bi-factor model and from –2.0 to –0.03 for the CCFA model. The CCFA estimation algorithm did not capture the threshold parameters as well as the bi-factor model, while both models exhibited similar levels of error in their slope estimates. In comparing distributional parameters, the bi-factor model showed slightly less error in the θ_2 mean and in the correlation between θ_1 and θ_2 (–0.1 versus –0.2 for both parameters), while the bi-factor model showed slightly more error in the θ_2 standard deviation (–0.1 versus –0.03).

Parameter recovery does not change dramatically when the sample size increases to 250, the results for which are presented in Table 10. Errors of estimation in the slopes on the primary dimension ranged from –0.6 to +1.8 for the bi-factor model and from –0.02 to +1.3 for the CCFA model. Errors in the slopes on the LD dimension ranged from –0.6 to +0.6 for the bi-factor model and from –0.6 to +0.8 for the CCFA model. Errors in the intercept parameters ranged from –0.4 to +0.2 for the bi-factor model and from –0.2 to +0.8 for the CCFA model. Again, the CCFA model produced poorer threshold parameter estimates, while slope parameter recovery for the primary dimension was somewhat better for the CCFA model than the bi-factor model in this sample. The bi-factor model had larger estimation errors than the CCFA model for the θ_2 mean (error of +0.2 versus +0.01 for CCFA) and the correlation between θ_1 and θ_2 (error of –0.3 versus +0.1 for CCFA). Error was similar for the θ_2 standard deviation for both models.

The results for the sample of 500 are presented in Table 11. Error of estimation in the slope parameters on the primary dimension ranged from +0.1 to +1.7 for the bi-factor model and from +0.3 to +0.8 for the CCFA model. Error in the slope parameters on the LD dimension ranged from –1.3 to –0.1 for the bi-factor model and from –0.7 to –0.1 for the CCFA model. Error in the intercept parameters ranged from 0 to +0.1 for the bi-factor model and from –0.3 to +0.5 for the CCFA model. Although the error decreases for this sample with 500 simulees as compared to the samples with 100 and 250 simulees, the CCFA model did not show improvement in capturing the intercept parameters, while the bi-factor model (more so than the CCFA model) did not show improvement in slope parameter estimation. The distributional parameters continue to have relatively large errors for this sample with both methods. For both models error was –0.1 for the mean and standard deviation of θ_2 and +0.1 for the correlation between θ_1 and θ_2.

Again, the simulated test was extended to 10 items with the same sample sizes as the 5-item tests, and parameter recovery was evaluated for this longer test condition with MSE providing a summary for each parameter. The results with 100 simulees are presented in Table 12. Errors of estimation in the slope parameter on the primary dimension ranged from –0.4 to +1.0 (MSE = 0.24) for the bi-factor model and from +0.01 to +3.6 (MSE = 2.35) for the CCFA model. Errors for the slope parameter on the LD dimension were between –0.3

and +1.1 (MSE = 0.23) for the bi-factor model and between –1.0 and +2.0 (MSE = 0.72) for the CCFA model. As with the 5-item examples, neither model precisely captures the slope parameters with 100 simulees, but the MSE values show that the bi-factor model provided more accurate estimates. Error in the intercept parameters ranged from –0.8 to +0.9 (MSE = 0.25) for the bi-factor model and from –4.3 and +0.6 (MSE = 2.26) for the CCFA model. Again, the CCFA model produces poor intercept parameter estimates with a small sample. Estimates for the distributional parameters did not worsen for this longer test with the mean and standard deviation error being –0.1 for both models, while the correlation estimation error was –0.02 for the bi-factor model and +0.2 for the CCFA model.

Results for 250 simulees and 10 items are presented in Table 13. Parameter estimation improved considerably with this increase in sample size. Primary slope estimation error was between –0.6 and +0.3 (MSE = 0.09) for the bi-factor model and between –0.2 and +1.0 (MSE = 0.30) for the CCFA model. LD slope error ranged from –0.03 to +0.8 (MSE = 0.10) for the bi-factor model and from –1.0 and +1.4 (MSE = 0.54) for the CCFA model. Intercept error of estimation varied from 0 to +0.4 (MSE = 0.04) for the bi-factor model and from –0.1 to +1.2 (MSE = 0.18) for the CCFA model. These MSE values were acceptable for the bi-factor model, while they remained inflated for the CCFA model. The distributional parameter estimation error was similar between the models and accuracy did not improve using this sample. Mean error was –0.2 for both items, standard deviation error was –0.04 for the bi-factor model and +0.1 for the CCFA model, and correlation error was –0.1 for the bi-factor model and +0.2 for the CCFA model.

Table 14 contains the results for the simulation involving a sample of 500 with 10 items. Again, the increased sample size improved parameter recovery. Error of estimation in the primary slope was between –0.2 and +1.0 (MSE = 0.16) for the bi-factor model and between +0.1 and +0.6 (MSE = 0.22) for the CCFA model. Estimation error in the LD slope varied from –0.4 to –0.1 (MSE = 0.07) for the bi-factor model and from –1.2 to +0.4 (MSE = 0.40) for the CCFA model. Error in the intercept ranged from –0.2 to +0.4 (MSE = 0.03) for the bi-factor model and from –0.3 to +0.2 (MSE = 0.04) for the CCFA model. The bi-factor model recovered the parameters better than the CCFA model, though there was little difference in the intercept errors between the models. Results for the distributional parameters are consistent with what was seen in other examples. The mean, standard deviation, and correlation errors were +0.01, –0.2, and +0.1, respectively, for the bi-factor model and were –0.01, –0.04, and +0.2, respectively, for the CCFA model.

Overall, the bi-factor model estimation algorithm did not capture the true parameters as well as the LCA model. Both the bi-factor model and the CCFA model performed well in estimating distributional parameters, but both models were sensitive to sample size and test length in estimating slopes and intercepts. Although it is difficult to outline data conditions necessary for accurate estimation using these few examples, the bi-factor model showed improvement in parameter recovery for the test with 10 items and 250 simulees.

5.1.3 Comparison Between Latent Class Analysis and Bi-factor Analysis

In addition to evaluating the parameter recovery of the LCA and bi-factor models, simulated data were used to investigate the relation between the two models. In Table 1 parameter estimates from the bi-factor model are presented for the example with 5000 simulees and 4 items simulated using the properties of the LCA model. It was anticipated that the slope on the primary dimension for the bi-factor model would correspond to the LCA slope and that the threshold for the bi-factor model (i.e., the intercept translated into a threshold using the formula $b_i = –d_i/(a_{i1} + a_{i2})$) would correspond to the LCA threshold. The connection between the LD parameter in the LCA model and the LD slope in the bi-factor model was unknown; the first is a probability of repeating the response from time 1 at time 2 and the second is a correlation between the responses at time 1 and time 2. The slopes and thresholds appear to correspond as expected, but some are estimated more precisely than others in the bi-factor model. The differences between the estimated bi-factor primary slope and the true LCA slope are –0.3, +2.6, –0.4, and –0.01, respectively, for each of the four items. The bi-factor intercept estimates are –0.88, 0.53, 0.41, and –0.32, respectively, for each of the four items, in terms of threshold reparameterization. These values differ from the true LCA threshold values by +0.1, –0.5, –0.1, and +0.2. Thus, for these known parameters of the bi-factor model, the estimates do not correspond to the true values as closely as would be expected with a sample of 5000. It is possible that generating data according to the LCA model and fitting it with the bi-factor model induces error in the parameter estimates that does not appear when the same models are used for simulation and estimation. Additionally, the distributional parameters for the bi-factor model did not correspond exactly to the true values. The mean differed by –0.06, the standard deviation by –0.15, and the correlation by +0.32.

In terms of the connection between the LD parameters in the two models, the order of increasing LD slopes for the bi-factor model is similar to the order of increasing LD parameters for the LCA model. The LCA LD parameters increase from item 1, to item 3, to item 4, to item 2, while the bi-factor LD slopes increase from item 1 to item 3, to item 2, to item 4. However, the primary slope estimate for item 2 had a large amount of error, so the LD slope estimate is probably also affected, which might affect this ordering. Still, any connection between these two conceptualizations of LD is unclear because the LCA model is accounting only for repeated responses, while the bi-factor model accounts for a correlation between underlying response processes.

A similar comparison was made for data generated using the bi-factor model and estimated using the LCA model. Results for this evaluation are presented in Table 8 for 5000 simulees and 4 items. Again, the LCA slope was expected to correspond to the bi-factor slope on the primary dimension, and the LCA threshold was expected to correspond to the bi-factor threshold (translated from the intercept using the slope values). There was some error in estimating the slopes, with LCA slope estimation error of +0.15, +0.46, +0.72, and –0.03 for the four items, respectively. Error also occurred in threshold estimation, with LCA thresholds being different from the bi-factor thresholds (–1.0, 1.0, 0.5, and –0.5) by +0.34, +0.35, +0.22, and –0.17, respectively. As in the previous example, this error suggests that generating data with one model and fitting it with the other induces inaccuracy in the parameter estimates. Some error was also present in the distributional parameters, which differed from the true parameters by –0.02 for the mean, –0.04 for the standard deviation, and +0.15 for the correlation.

The LD parameters for the LCA model increase in the same order as the true bi-factor LD slope values. The LCA LD parameters increase from item 1 and item 4, to item 2, to item 3, while the bi-factor LD slopes increase from item 1, to item 4, to item 2, to item 3. However, the magnitude of the LCA LD parameters is surprisingly small. When data were generated with the LCA model and fit with the bi-factor model, LD parameters of 0.1, 0.2, 0.3, and 0.4 were estimated as LD slopes[3] of –0.4,–0.6,–1.5, and 1.1, respectively. Here, when data were

[3] The sign of these slopes is arbitrary and should not be interpreted.

generated with the bi-factor model and fit with the LCA model, LD slopes of 0.6, 0.7, 1.0, and 1.5 were estimated as LD parameters of 0.01, 0.01, 0.02, and 0.05, respectively.

To explore the possibility of the generating LD slope values for the bi-factor model indicating weak LD, an additional dataset was simulated with the LD slope values doubled to range from 1.2 to 3.0. These increased true values had little effect on the LD parameter estimates for the LCA model. The estimated values ranged from 0 to 0.08, and although they increased in the same order as the estimated LD slopes, their magnitude did not reflect the additional LD that was assumed to be added to the data by increasing the slope values.

The LCA model and the bi-factor model are not two different parameterizations of the same IRT model. The LCA model includes LD parameters, which measure the probability that the response to an item at time 1 is repeated as the response to that item at time 2. The bi-factor model includes LD slopes, which parameterize the correlation between an item's responses at time 1 and time 2. While these parameters are intended to model LD, they do not correspond to identical conceptualizations of LD. LD parameters in the LCA model answer the question, "What is the probability that examinees will provide the same response to an item at time 2 that was given at time 1 beyond what is implied by the latent trait simply because the item had already been considered by the examinees at an earlier time?" This type of LD corresponds to what Chen and Thissen (1997) term "surface local dependence." LD slopes in the bi-factor model answer the question, "How related to each other are the responses to this item at two time points beyond their relation to the latent trait?" This type of LD is similar to what Chen and Thissen (1997) label "underlying local dependence." The identification of these two types of LD by these researchers highlights the fact that there may exist different forces behind observed LD between item responses. It is logical that when one type of LD is induced in simulated data, the model that was used to simulate the data demonstrates better parameter recovery than does the alternative model. This result was seen here, and it indicates that a model should be chosen that corresponds to the assumed underlying LD mechanism. This point will be revisited after examination of the empirical data.

5.2 Empirical Data

Empirical data from the Context Study were used to examine the performance of the proposed models with real data. The four unidimensional psychological distress items (1, 2, 7, and 9 in Figure 2) were modeled with the proposed LCA and bi-factor models, as well as with the existing 2PL and CCFA models. Although the sample was substantial and excess model error was not anticipated, the opportunity was taken to compare the standard errors from each of the four models using one dataset. Standard errors for the proposed model are not available through the EM algorithm implementation[4], so R's non-linear minimizing function was used instead. This function calculates standard errors from the square root of the inverse of the Hessian matrix. The point estimates produced by this function in R are identical to those from the EM algorithm estimation approach in C++.

5.2.1 Latent Class Analysis

Results for the LCA model and the 2PL model are presented in Table 15. The slope parameters range from 2.0 to 3.3 for the LCA model and from 2.0 to 3.2 for the 2PL model, indicating that these four items are good measures of psychological distress. The slope estimates are very similar between models, with the largest difference being 0.13. The standard errors for the slope parameter estimates are similar between the approaches. The

[4] Recent work borrowing on Meng and Rubin's (1991) supplemented EM algorithm has improved standard error calculation in IRT, and this method may be incorporated into future versions of this software.

threshold parameters are also consistent across models, and the largest difference is 0.03. These threshold values range from –0.44 to 0.71 for the LCA model and from –0.42 to 0.71 for the 2PL model, indicating that these items provide information about psychological distress around the center of the θ distribution. The standard errors for these threshold estimates are identical across models, likely because they are so small that there is no precision to be gained with the LCA model. The LCA model estimated a smaller mean, a greater standard deviation, and a smaller correlation for the distributional parameters than did the 2PL model, but any discrepancy is due to the 2PL model using summed scores to calculate these values. The distributional parameters suggest that the center of the psychological distress distribution is stable over time, but the variability increases. Although there is no standard error to associate with the 2PL distribution values, the standard errors for the distribution parameters for the LCA model are small.

The LD parameters as estimated by the LCA model range from 0.11 to 0.21, and each had a small standard error of 0.02. The smallest LD values are associated with item 1 ("I had trouble getting my breath") and item 9 ("I was a bad person"), while the largest value was for item 2 ("I got mad easily"). Item 7 ("I often worried about bad things happening to me") had an LD value equidistant from the largest and smallest values. This is to say that the probability of examinees repeating their time 1 response to items 1 and 9 at time 2 is lower than the probability of repeating item 7, and item 2 has a higher probability of repeating than items 1, 7, and 9. This pattern does make intuitive sense. Having trouble getting one's breath is a condition which may vary for reasons other than psychological distress (e.g., weather conditions or activity level of the child). Variability in getting one's breath within a half a year could be expected, and this appears in the small LD parameter value. Believing that one is a bad person is more difficult to explain, but a response to this item may be based upon recent actions and could also vary over time. Worrying about bad things happening can be seen as more of a trait condition, where some people have a tendency to worry about things more than others. The larger LD parameter for this item is consistent with this more stable condition. Getting mad easily is the most trait-like condition of the four items. Some people have a propensity to get mad easily, which may not be related to psychological distress, and it makes sense that this item would have a higher LD parameter estimate than the other items.

5.2.2 Bi-factor Analysis

Results for the bi-factor model and the CCFA model in fitting the four psychological distress scale items are presented in Table 16. The primary slope parameters range from 2.13 to 3.76 for the bi-factor model, and from 2.84 to 4.22 for the CCFA model. These slope estimates do not correspond closely for the two models and are not in the same order of increasing magnitude. This is likely an extension of the error seen estimates using the CCFA model observed in the simulated data examples. The difference between the standard errors for the two models is striking. The bi-factor model has primary slope standard errors ranging from 0.09 to 0.38, while the CCFA model has primary slope standard errors ranging from 0.47 to 1.72. Because the sample size was so large, WLS estimation was used for this CCFA model, so the full weight matrix is used in calculating standard errors. These standard error differences show the value that is to be gained by using the full-information approach of the bi-factor model.

The threshold values for the bi-factor model ranged from –1.92 to 1.09 and from –3.41 to 1.13 for the CCFA model. These differences are consistent with what was observed with the simulated data, that the CCFA model tends to poorly estimate large thresholds. The standard errors are closer between the models, with those of the bi-factor model between 0.06 and 0.10 and those of the CCFA model between 0.07 and 0.17. The largest CCFA standard error corresponds to the most extreme threshold estimate, which highlights difficulty associated with estimating that parameter.

The distributional parameters were more consistent between the bi-factor model and the CCFA model than were those between the LCA model and the 2PL model, likely because the bi-factor model and the CCFA model both estimate these values by incorporating latent traits for the local dependence. The models provided identical mean estimates, standard deviations that differed by 0.01, and correlations that varied by 0.06. The standard errors for these estimates were also similar between models.

Large differences were also observed between the LD slopes for the bi-factor model and those for the CCFA model. The bi-factor model produced LD slopes between 0.39 and 1.47 (absolute values), while the CCFA model produced LD slopes between 1.28 and 1.96. The standard errors of these estimates were not so discrepant, with the bi-factor model having larger standard errors than the CCFA model for 3 of the 4 items, and the largest standard error difference was 0.05.

It is also interesting to compare the estimates and standard errors between the two proposed models. The slope for item 1 is similar between the two models, but the slopes are more discrepant for the remaining three items. The slope estimates were not in the same order of increasing magnitude between the two models. The standard error for item 1 was identical between models (0.09), and the LCA model had much smaller standard errors for items 2 and 7 (0.09 and 0.11 versus 0.22 and 0.38), while the bi-factor model had a slightly smaller standard error for item 9 (0.13 versus 0.18). The intercepts of the bi-factor model translate to 0.59, –0.27, 0.15, and 0.59 for the four items, respectively. These estimates are close to those of the LCA model, with the biggest difference being 0.17, and the standard errors for the threshold parameters for each of these models are near 0.02. The LCA model estimates the mean of θ_2 as slightly smaller than that of the bi-factor model (0.04 versus 0.07), the standard deviation as larger (1.39 versus 1.17), and the correlation as smaller (0.50 versus 0.61). The standard errors for these estimates are similar.

The LD parameters are in the same order of increasing magnitude as the LD slopes. The magnitude of the standard errors for these estimates is much smaller for the LCA model than for the bi-factor model (near 0.02 versus near 0.20). Again, any connection between the two models is difficult to identify, but the models do appear to agree with each other in terms of magnitude and direction.

It is worth revisiting the purpose of this research: to model the LD between the same items at different time points so that parameter calibration can be conducted using longitudinal item response data. The LD in this model is a nuisance parameter that must be included so that the other parameters can be estimated accurately. We would want the parameter estimates for the longitudinal data modeled with an IRT model accounting for LD to be similar to the parameter estimates for the data at one time point modeled with a traditional IRT model (here, the 2PL model), given that the sample size at one time point is adequate for accurate estimation. This correspondence between the two models was seen using the empirical data with the LCA model and the 2PL model. This suggests that the LCA model adequately accounts for LD between item responses at two time points to be able to combine data across administrations and perform item calibration. The estimates from the bi-factor model did not correspond to those of the 2PL model, suggesting that the bi-factor model may not be so appropriate for this purpose. From this empirical example, we might infer that the nuisance LD that exists between item responses over time is more like surface local dependence, which is accounted for in the LCA model. The bi-factor model does not account for this nuisance LD as well as the LCA model, but the bi-factor model may be useful for other purposes not considered as a motivation for this research.

Throughout these simulated and empirical data examples, the amount of time required for each model to converge varied. The LCA model took between 15 seconds and 1 hour, 39 minutes. The bi-factor model took between 2 minutes and 1 hour, 24 minutes. In general, larger samples and longer tests required more time until convergence. These proposed models do require increased computing time as compared to those of the existing models, which converge within seconds regardless of sample size or test length.

CHAPTER 6

DISCUSSION

The purpose of this research was to propose, implement, and evaluate two approaches to modeling longitudinal data with IRT. Models based on latent class analysis and bi-factor analysis were introduced, and parameter estimation using the EM algorithm was implemented in C++. Simulated data and empirical data were employed to assess parameter recovery and the potential value of these models.

6.1 Evaluation of the Proposed Approaches

A number of simulated datasets were created using each of the proposed models to compare estimated parameters to true parameters. The LCA model achieved better parameter recovery than did the bi-factor model in these sets of simulated data. The LCA model recovered the parameters well across the various conditions and its performance did not appear to be limited by sample size or test length. While it was desired that the LCA model demonstrate better slope and threshold recovery than the 2PL model, the similarity in the results for these models is encouraging because the LCA model has more than twice as many additional parameters to estimate than does the 2PL model. Thus, even if the LCA model does not improve parameter estimation above and beyond what is currently available, it does enhance value because of information provided by the additional parameters it includes.

The LD parameters of the LCA model were well-estimated across the various sample sizes and test lengths. This was an encouraging result because it was feared that sparseness in latent classes (resulting from small samples) or excessive numbers of potential latent classes (resulting from increased test length) would render the model computationally difficult. Although computing time did increase as test length increased (i.e., as more latent classes were considered in the model), parameter recovery did not suffer.

There was also good parameter recovery of the distributional parameters for the LCA model. Parameter recovery was reasonable across the data conditions, but increased sample size improved distributional parameter estimation more than it did for estimation of the item-level parameters. By including the distributional parameters as part of the model, thus modeling the 2-dimensional latent distribution, the distributional parameter estimation was better for the LCA model than for the 2PL model, which does not have a mechanism for including these parameters with longitudinal designs.

The results for the bi-factor model, while encouraging, were not as positive as those for the LCA model. The bi-factor model, with its inclusion of an additional dimension, required larger sample sizes and longer tests to achieve satisfactory parameter recovery than did the LCA model. The primary motivation for this research was to develop a model that could accurately estimate IRT parameters with a small sample, and the bi-factor model does not appear to satisfy this goal. However, it is difficult to generalize using these simulated examples; it appears that the bi-factor model may be appropriate for longitudinal data designs, and it should be considered in future LD research.

While the LCA model and the bi-factor model were proposed with the sample size issue in mind, they do offer several benefits that have not been available through existing IRT data models. That is, even when sample size is adequate for use with existing IRT models, these new models have properties that may make them a preferred choice.

First, these models allow item parameters to be calibrated using data from more than one time point. An assumption of IRT is that item parameters are invariant across samples, but there may be concern that a particular sample is not representative of the population. For example, in a clinical drug trial, the time point chosen for parameter calibration may be sensitive to treatment, as in the presence or absence of treatment may dictate the region on θ in which the examinees are located at that particular time. By including multiple time points, the effective sample may be located on a wider θ region, thus improving the generalizability of the results to the whole population.

Second, these models permit the latent distribution parameters to be estimated over two time points while properly accounting for item local dependence between time points. Previous research has introduced the latent distribution parameters under the assumption that the item responses are independent between time points, but with these models this restrictive assumption is unnecessary. With item parameter calibration conducted simultaneously with distributional parameter estimation, these models can obtain this information with one dataset rather than collecting two datasets, one for item parameter calibration and one for distributional parameter estimation.

Third, these models introduce two unique methods for evaluating the local dependence between item responses at two time points. The LCA model includes LD parameters, which measure the probability that the response to an item at time 1 is repeated as the response to that item at time 2. The bi-factor model includes LD slopes, which measure the correlation between an item’s responses at time 1 and time 2. While these parameters are intended to model LD, they do not correspond to identical conceptualizations of LD. The LCA model accounts for LD observed in repeated responses. Alternatively, the bi-factor model accounts for LD as a correlation between underlying response process factors. These two conceptualizations of LD indicate that the source of the LD must be known before a model can be chosen. In the case of longitudinal item response data analyzed here, it appears that the better fitting model is the LCA model, where LD is treated as a nuisance and accounted for in parameter calibration.

6.2 Future Directions

This research proposed and evaluated two new approaches to longitudinal IRT data modeling, and as with any novel method, there is much to be considered in future research. An obvious next step is a large simulation study design, where many samples in each cell of the design are generated and results are aggregated across these samples using statistics such as bias and root mean square error. The examples used here varied the sample size between 100 and 500 simulees and the test length between 5 and 10 items. These conditions could be further varied to determine how small a dataset can be to achieve accurate estimation, how large a dataset can be before additional data adds little accuracy, and how these two conditions interact with each other. Other properties of the simulated datasets in this research were stable, with one set of item parameters, LD magnitudes, and distributional parameters. Item parameters could be varied to see how parameter recovery is affected by items that are weakly or strongly related to the latent trait or by items that are of varying degrees of difficulty. The amount of LD between items administered at two different times should be examined across the spectrum from no LD (response at time 2 is completely unique) to total LD (response at time 2 is completely redundant). The distributional parameters should also be varied to examine how large shifts in the latent trait, increased variability, or small or large correlation between the latent traits at the two administrations may affect the parameter recovery.

Another clear future goal is to implement these approaches to modeling LD with IRT models that account for polytomous data. The present models only allow for dichotomous data without guessing. It was necessary to establish that these models work with dichotomous data before extending them to polytomous data, but this extension would be very useful for the type of data for which the models are intended (e.g., psychological questionnaires that often employ Likert response scales).

A subsequent extension of the proposed models would be to account for more than 2 time points within the model. Including data from additional administrations may further ease the sample size burden for parameter calibration, or it may require larger samples because of the increased dimensionality. Even if a model that includes data from more than 2 administrations requires larger datasets, modeling the latent trait over multiple time points may be of interest to many researchers.

For more complicated data problems, the parameters of these models may be better estimated with MCMC techniques. Many data problems can be solved with maximum likelihood estimation, but as dimensionality increases, maximum likelihood solutions are more difficult to obtain. MCMC estimation is an alternative that has been gaining popularity, especially for multidimensional IRT models. As additional time points are added to the model, MCMC may be a better alternative for parameter estimation.

An interesting use of these models would be to use them with multiple-reporter data. For example, data may be collected on couples where the latent trait pertains to a quality of the couple. We might assume that the item parameters should be invariant across the reporter, and it might be reasonable to calibrate the parameters using a sample that includes both members of each couple. The responses within a dyad should not be independent, but these models could be employed to account for this LD. This would facilitate parameter calibration, as well as measure the difference between the members (e.g., men and women) on the latent trait and measure if there is any propensity to respond to each item in a certain manner that is not accounted for by the latent trait.

The models as parameterized here assume that the items function identically across time, which is an appropriate assumption when the goal is item parameter calibration for use in subsequent studies. In other words, it would be difficult to determine the set of item parameters to use when scoring a scale in future applications when the items on the scale function differently across time. However, researchers might want to use these models to examine if items function differently across time. It would be interesting for these models to be reparameterized so that measurement invariance can be examined.

Finally, it will be useful to apply the proposed models to data with different sources of LD beyond LD between responses to the same item across time. The bi-factor model may be an appropriate approach to measuring the amount of LD between two unique items on a scale. If LD between such items is considered a form of a second factor that explains variance between these items beyond what is explained by the first factor, then modeling this second factor using the bi-factor model may be a useful solution. Researchers may find that they can develop scales that have LD items and still reliably measure the outcome of interest if the LD is isolated using an LD factor.

The goal of this research was to develop a model that could calibrate item parameters using data from two time points by accounting for the LD between responses to the same items across time, and this goal was met with the LCA model. In addition to facilitating item parameter calibration with fewer examinees, this model includes estimates for the change in the latent trait over time, as well as including a metric for the dependence between item responses across time. The bi-factor model, while not ideal for the longitudinal item response data considered here, shows promise for modeling LD between items in other contexts. Together, the models introduced here open the door to future research on both longitudinal IRT and measuring LD within scales.

Table 1. Parameter recovery of data generated with the LCA model ($N = 5000$, $I = 4$).

LCA	**True Values**	Estimates	Estimate – True	**Bi–factor**	Estimates
a_1	**1.1**	1.10	0.00	a_{11}	0.81
b_1	**–1.0**	–1.02	–0.02	a_{12}	–0.41
κ_1	**0.1**	0.07	–0.03	d_1	1.07
a_2	**1.3**	1.28	–0.02	a_{21}	3.93
b_2	**1.0**	0.98	–0.02	a_{22}	1.10
κ_2	**0.4**	0.41	+0.01	d_2	–2.67
a_3	**1.5**	1.54	+0.04	a_{31}	1.15
b_3	**0.5**	0.51	+0.01	a_{32}	–0.55
κ_3	**0.2**	0.19	–0.01	d_3	–0.69
a_4	**1.7**	1.60	–0.10	a_{41}	1.69
b_4	**–0.5**	–0.54	–0.04	a_{42}	–1.49
κ_4	**0.3**	0.30	0.00	d_4	1.02
ρ	**0.5**	0.51	+0.01	ρ	0.82
μ_2	**0.2**	0.17	–0.03	μ_2	0.14
σ_2	**1.0**	0.98	–0.02	σ_2	0.85
$-ll$		22675.03		$-ll$	22832.44

Table 2. Parameter recovery of data generated with the LCA model ($N = 100$, $I = 5$).

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
a_1	**1.1**	1.00	–0.10	a_1	0.90	–0.20
b_1	**–1.0**	–1.17	–0.17	b_1	–1.22	–0.22
κ_1	**0.1**	0.06	–0.04			
a_2	**1.3**	1.29	–0.01	a_2	1.00	–0.30
b_2	**1.0**	1.00	0.00	b_2	1.02	+0.02
κ_2	**0.4**	0.29	–0.11			
a_3	**1.5**	1.46	–0.04	a_3	1.41	–0.09
b_3	**0.5**	0.68	+0.18	b_3	0.72	+0.22
κ_3	**0.2**	0.06	–0.14			
a_4	**1.7**	3.14	+1.44	a_4	4.02	+2.32
b_4	**–0.5**	–0.23	+0.27	b_4	–0.18	+0.32
κ_4	**0.3**	0.00	–0.03			
a_5	**1.9**	2.19	+0.29	a_5	2.18	+0.28
b_5	**0.0**	0.03	+0.03	b_5	0.09	+0.09
κ_5	**0.25**	0.26	+0.01			
ρ	**0.5**	0.67	+0.17	ρ	0.81	+0.31
μ_2	**0.2**	0.27	+0.07	μ_2	0.19	–0.01
σ_2	**1.0**	0.99	–0.01	σ_2	0.98	–0.02
$-ll$		552.68				

Table 3. Parameter recovery of data generated with the LCA model ($N = 250$, $I = 5$).

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
$\boldsymbol{a_1}$	**1.1**	1.20	+0.10	$\boldsymbol{a_1}$	1.19	+0.09
$\boldsymbol{b_1}$	**–1.0**	–0.80	+0.20	$\boldsymbol{b_1}$	–0.81	+0.19
$\boldsymbol{\kappa_1}$	**0.1**	0.04	–0.06			
$\boldsymbol{a_2}$	**1.3**	1.45	+0.15	$\boldsymbol{a_2}$	1.33	+0.03
$\boldsymbol{b_2}$	**1.0**	0.89	–0.11	$\boldsymbol{b_2}$	1.00	+0.00
$\boldsymbol{\kappa_2}$	**0.4**	0.55	+0.15			
$\boldsymbol{a_3}$	**1.5**	1.33	–0.17	$\boldsymbol{a_3}$	1.47	–0.03
$\boldsymbol{b_3}$	**0.5**	0.63	+0.13	$\boldsymbol{b_3}$	0.64	+0.14
$\boldsymbol{\kappa_3}$	**0.2**	0.21	+0.01			
$\boldsymbol{a_4}$	**1.7**	1.62	–0.08	$\boldsymbol{a_4}$	1.62	–0.08
$\boldsymbol{b_4}$	**–0.5**	–0.32	+0.18	$\boldsymbol{b_4}$	–0.32	+0.18
$\boldsymbol{\kappa_4}$	**0.3**	0.21	–0.09			
$\boldsymbol{a_5}$	**1.9**	2.54	+0.64	$\boldsymbol{a_5}$	2.49	+0.59
$\boldsymbol{b_5}$	**0.0**	0.11	+0.11	$\boldsymbol{b_5}$	0.08	+0.08
$\boldsymbol{\kappa_5}$	**0.25**	0.29	+0.04			
$\boldsymbol{\rho}$	**0.5**	0.55	+0.05	$\boldsymbol{\rho}$	0.89	+0.39
$\boldsymbol{\mu_2}$	**0.2**	0.25	+0.05	$\boldsymbol{\mu_2}$	0.16	–0.04
$\boldsymbol{\sigma_2}$	**1.0**	0.96	–0.04	$\boldsymbol{\sigma_2}$	0.93	–0.07
–ll		1410.33				

Table 4. Parameter recovery of data generated with the LCA model ($N = 500$, $I = 5$).

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
a_1	**1.1**	1.24	+0.14	a_1	1.25	+0.15
b_1	**–1.0**	–0.79	+0.21	b_1	–0.85	+0.15
κ_1	**0.1**	0.03	–0.07			
a_2	**1.3**	1.73	+0.43	a_2	1.87	+0.57
b_2	**1.0**	0.84	–0.16	b_2	0.82	–0.18
κ_2	**0.4**	0.37	–0.03			
a_3	**1.5**	1.52	+0.02	a_3	1.41	–0.09
b_3	**0.5**	0.51	+0.01	b_3	0.50	+0.00
κ_3	**0.2**	0.22	+0.02			
a_4	**1.7**	1.62	–0.08	a_4	1.53	–0.17
b_4	**–0.5**	–0.43	+0.07	b_4	–0.43	+0.07
κ_4	**0.3**	0.20	–0.10			
a_5	**1.9**	1.86	–0.04	a_5	1.94	+0.04
b_5	**0.0**	–0.01	–0.01	b_5	0.03	+0.03
κ_5	**0.25**	0.24	–0.01			
ρ	**0.5**	0.55	+0.05	ρ	0.82	+0.32
μ_2	**0.2**	0.25	+0.05	μ_2	0.15	–0.05
σ_2	**1.0**	1.05	+0.05	σ_2	0.97	–0.03
$-ll$		2839.44				

Table 5. Parameter recovery of data generated with the LCA model (N = 100, I = 10).

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
$\boldsymbol{a_1}$	**1.1**	1.36	+0.26	$\boldsymbol{a_1}$	1.69	+0.59
$\boldsymbol{b_1}$	**–1.0**	–0.71	+0.29	$\boldsymbol{b_1}$	–0.51	+0.49
$\boldsymbol{\kappa_1}$	**0.1**	0.12	+0.02			
$\boldsymbol{a_2}$	**1.3**	1.59	+0.29	$\boldsymbol{a_2}$	1.70	+0.40
$\boldsymbol{b_2}$	**1.0**	0.65	–0.35	$\boldsymbol{b_2}$	0.74	–0.26
$\boldsymbol{\kappa_2}$	**0.4**	0.38	–0.02			
$\boldsymbol{a_3}$	**1.5**	1.42	–0.08	$\boldsymbol{a_3}$	1.36	–0.14
$\boldsymbol{b_3}$	**0.5**	0.39	–0.11	$\boldsymbol{b_3}$	0.44	–0.06
$\boldsymbol{\kappa_3}$	**0.2**	0.08	–0.12			
$\boldsymbol{a_4}$	**1.7**	2.45	+0.75	$\boldsymbol{a_4}$	2.46	+0.76
$\boldsymbol{b_4}$	**–0.5**	–0.27	+0.23	$\boldsymbol{b_4}$	–0.22	+0.28
$\boldsymbol{\kappa_4}$	**0.3**	0.32	+0.02			
$\boldsymbol{a_5}$	**1.9**	1.90	0.00	$\boldsymbol{a_5}$	1.67	–0.23
$\boldsymbol{b_5}$	**0.0**	0.03	+0.03	$\boldsymbol{b_5}$	0.00	0.00
$\boldsymbol{\kappa_5}$	**0.25**	0.43	+0.18			
$\boldsymbol{a_6}$	**1.2**	1.41	+0.21	$\boldsymbol{a_6}$	1.39	+0.19
$\boldsymbol{b_6}$	**0.0**	–0.01	–0.01	$\boldsymbol{b_6}$	0.08	+0.08
$\boldsymbol{\kappa_6}$	**0.2**	0.16	–0.04			
$\boldsymbol{a_7}$	**1.4**	1.55	+0.15	$\boldsymbol{a_7}$	1.47	+0.07
$\boldsymbol{b_7}$	**–0.5**	–0.31	+0.19	$\boldsymbol{b_7}$	–0.59	–0.09

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
$\boldsymbol{\kappa_7}$	**0.1**	0.02	–0.08			
$\boldsymbol{a_8}$	**1.6**	2.26	+0.66	$\boldsymbol{a_8}$	1.85	+0.25
$\boldsymbol{b_8}$	**0.5**	0.27	–0.23	$\boldsymbol{b_8}$	0.34	–0.16
$\boldsymbol{\kappa_8}$	**0.25**	0.33	+0.08			
$\boldsymbol{a_9}$	**1.8**	2.04	+0.24	$\boldsymbol{a_9}$	2.11	+0.31
$\boldsymbol{b_9}$	**1.0**	0.83	–0.17	$\boldsymbol{b_9}$	0.79	–0.21
$\boldsymbol{\kappa_9}$	**0.3**	0.37	+0.07			
$\boldsymbol{a_{10}}$	**2.0**	1.33	–0.67	$\boldsymbol{a_{10}}$	1.99	–0.01
$\boldsymbol{b_{10}}$	**–1.0**	–1.15	–0.15	$\boldsymbol{b_{10}}$	–1.06	–0.06
$\boldsymbol{\kappa_{10}}$	**0.4**	0.35	–0.05			
$\boldsymbol{\rho}$	**0.5**	0.45	–0.05	$\boldsymbol{\rho}$	0.75	+0.25
$\boldsymbol{\mu_2}$	**0.2**	0.20	0.00	$\boldsymbol{\mu_2}$	0.15	–0.05
$\boldsymbol{\sigma_2}$	**1.0**	0.79	–0.21	$\boldsymbol{\sigma_2}$	0.83	–0.17
–ll		1095.76				

Table 6. Parameter recovery of data generated with the LCA model ($N = 250$, $I = 10$).

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
a_1	**1.1**	1.37	+0.27	a_1	1.40	+0.30
b_1	**–1.0**	–0.77	+0.23	b_1	–0.88	+0.12
κ_1	**0.1**	0.18	+0.08			
a_2	**1.3**	1.29	–0.01	a_2	1.36	+0.06
b_2	**1.0**	1.15	+0.15	b_2	1.07	+0.07
κ_2	**0.4**	0.41	+0.01			
a_3	**1.5**	1.76	+0.26	a_3	1.82	+0.32
b_3	**0.5**	0.54	+0.04	b_3	0.54	+0.04
κ_3	**0.2**	0.22	+0.02			
a_4	**1.7**	1.44	–0.26	a_4	1.28	–0.42
b_4	**–0.5**	–0.47	+0.03	b_4	–0.46	+0.04
κ_4	**0.3**	0.29	–0.01			
a_5	**1.9**	2.24	+0.34	a_5	2.46	+0.56
b_5	**0.0**	0.13	+0.03	b_5	0.16	+0.16
κ_5	**0.25**	0.21	–0.04			
a_6	**1.2**	1.45	+0.25	a_6	1.43	+0.23
b_6	**0.0**	–0.01	–0.01	b_6	0.05	+0.05
κ_6	**0.2**	0.34	+0.14			
a_7	**1.4**	1.31	–0.09	a_7	1.42	+0.02
b_7	**–0.5**	–0.48	+0.02	b_7	–0.53	–0.03

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
κ_7	**0.1**	0.08	–0.02			
a_8	**1.6**	1.53	–0.07	a_8	1.51	–0.09
b_8	**0.5**	0.62	+0.12	b_8	0.67	+0.17
κ_8	**0.25**	0.24	–0.01			
a_9	**1.8**	1.90	+0.10	a_9	2.22	+0.42
b_9	**1.0**	1.06	+0.06	b_9	1.03	+0.03
κ_9	**0.3**	0.28	–0.02			
a_{10}	**2.0**	1.83	–0.17	a_{10}	1.46	–0.54
b_{10}	**–1.0**	–1.05	–0.05	b_{10}	–1.15	–0.15
κ_{10}	**0.4**	0.41	+0.01			
ρ	**0.5**	0.56	+0.06	ρ	0.82	+0.32
μ_2	**0.2**	0.18	–0.02	μ_2	0.11	–0.09
σ_2	**1.0**	0.93	–0.07	σ_2	0.91	–0.09
$-ll$		2631.79				

Table 7. Parameter recovery of data generated with the LCA model ($N = 500$, $I = 10$).

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
$\boldsymbol{a_1}$	**1.1**	1.05	–0.05	$\boldsymbol{a_1}$	1.08	–0.02
$\boldsymbol{b_1}$	**–1.0**	–1.20	–0.20	$\boldsymbol{b_1}$	–1.15	–0.15
$\boldsymbol{\kappa_1}$	**0.1**	0.13	+0.03			
$\boldsymbol{a_2}$	**1.3**	1.14	–0.16	$\boldsymbol{a_2}$	1.05	–0.25
$\boldsymbol{b_2}$	**1.0**	0.84	–0.16	$\boldsymbol{b_2}$	0.95	–0.05
$\boldsymbol{\kappa_2}$	**0.4**	0.36	–0.04			
$\boldsymbol{a_3}$	**1.5**	1.14	–0.36	$\boldsymbol{a_3}$	1.20	–0.30
$\boldsymbol{b_3}$	**0.5**	0.52	+0.02	$\boldsymbol{b_3}$	0.42	–0.08
$\boldsymbol{\kappa_3}$	**0.2**	0.21	+0.01			
$\boldsymbol{a_4}$	**1.7**	1.40	–0.30	$\boldsymbol{a_4}$	1.26	–0.44
$\boldsymbol{b_4}$	**–0.5**	–0.76	–0.26	$\boldsymbol{b_4}$	–0.76	–0.26
$\boldsymbol{\kappa_4}$	**0.3**	0.27	–0.03			
$\boldsymbol{a_5}$	**1.9**	1.81	–0.09	$\boldsymbol{a_5}$	2.06	+0.16
$\boldsymbol{b_5}$	**0.0**	–0.25	–0.25	$\boldsymbol{b_5}$	–0.31	–0.31
$\boldsymbol{\kappa_5}$	**0.25**	0.26	+0.01			
$\boldsymbol{a_6}$	**1.2**	1.14	–0.06	$\boldsymbol{a_6}$	1.06	–0.14
$\boldsymbol{b_6}$	**0.0**	–0.09	–0.09	$\boldsymbol{b_6}$	–0.18	–0.18
$\boldsymbol{\kappa_6}$	**0.2**	0.26	+0.06			
$\boldsymbol{a_7}$	**1.4**	1.30	–0.10	$\boldsymbol{a_7}$	1.45	+0.05
$\boldsymbol{b_7}$	**–0.5**	–0.62	–0.12	$\boldsymbol{b_7}$	–0.50	+0.00

LCA	**True Values**	Estimates	Estimate – True	**2PL**	Estimates	Estimate – True
κ_7	**0.1**	0.10	0.00			
a_8	**1.6**	1.28	–0.32	a_8	1.22	–0.38
b_8	**0.5**	0.26	–0.24	b_8	0.25	–0.25
κ_8	**0.25**	0.28	+0.03			
a_9	**1.8**	1.72	–0.08	a_9	1.58	–0.22
b_9	**1.0**	0.77	–0.23	b_9	0.86	–0.14
κ_9	**0.3**	0.29	–0.01			
a_{10}	**2.0**	2.09	+0.09	a_{10}	2.31	+0.31
b_{10}	**–1.0**	–1.27	–0.27	b_{10}	–1.19	–0.19
κ_{10}	**0.4**	0.36	–0.04			
ρ	**0.5**	0.48	–0.02	ρ	0.76	+0.26
μ_2	**0.2**	0.16	–0.04	μ_2	0.10	–0.10
σ_2	**1.0**	1.10	+0.10	σ_2	0.96	–0.04
$-ll$		5440.36				

Table 8. Parameter recovery of data generated with bi–factor model (N = 5000, I = 4).

Bi–factor	**True Values**	Estimates	Estimate – True	**LCA**	Estimates
$\boldsymbol{a_{11}}$	**1.1**	1.09	–0.01	$\boldsymbol{a_{1}}$	1.25
$\boldsymbol{a_{12}}$	**0.6**	0.65	+0.05	$\boldsymbol{b_{1}}$	–1.34
$\boldsymbol{d_{1}}$	**1.7**	1.67	–0.03	$\boldsymbol{\kappa_{1}}$	0.01
$\boldsymbol{a_{21}}$	**1.3**	1.19	–0.11	$\boldsymbol{a_{2}}$	1.76
$\boldsymbol{a_{22}}$	**1.0**	1.21	+0.21	$\boldsymbol{b_{2}}$	1.35
$\boldsymbol{d_{2}}$	**–2.3**	–2.32	–0.02	$\boldsymbol{\kappa_{2}}$	0.02
$\boldsymbol{a_{31}}$	**1.5**	1.44	–0.06	$\boldsymbol{a_{3}}$	2.22
$\boldsymbol{a_{32}}$	**1.5**	1.67	+0.17	$\boldsymbol{b_{3}}$	0.72
$\boldsymbol{d_{3}}$	**–1.5**	–1.58	–0.08	$\boldsymbol{\kappa_{3}}$	0.05
$\boldsymbol{a_{41}}$	**1.7**	1.60	–0.10	$\boldsymbol{a_{4}}$	1.67
$\boldsymbol{a_{42}}$	**0.7**	0.89	+0.19	$\boldsymbol{b_{4}}$	–0.67
$\boldsymbol{d_{4}}$	**1.2**	1.14	–0.06	$\boldsymbol{\kappa_{4}}$	0.01
$\boldsymbol{\rho}$	**0.5**	0.43	–0.07	$\boldsymbol{\rho}$	0.65
$\boldsymbol{\mu_{2}}$	**0.2**	0.23	+0.03	$\boldsymbol{\mu_{2}}$	0.18
$\boldsymbol{\sigma_{2}}$	**1.0**	0.95	–0.05	$\boldsymbol{\sigma_{2}}$	0.96
–ll		19977.01		***–ll***	19988.41

Table 9. Parameter recovery of data generated with the bi–factor model (N = 100, I = 5).

Bi–factor	**True Values**	Estimates	Estimate – True	**CCFA**	Estimates	Translation	Translation – True
a_{11}	**1.1**	1.12	+0.02	λ_1	0.66	1.53	+0.43
a_{12}	**0.6**	0.33	–0.27	δ_1	–0.03	0.40	–0.20
d_1	**1.7**	1.86	+0.16	τ_1	–1.22	1.67	–0.03
a_{21}	**1.3**	1.84	+0.54	λ_2	0.83	3.52	+2.22
a_{22}	**1.0**	1.64	+0.64	δ_2	–0.15	1.64	+0.64
d_2	**–2.3**	–2.98	–0.68	τ_2	1.73	–4.31	–2.01
a_{31}	**1.5**	1.62	+0.12	λ_3	0.85	4.41	+2.91
a_{32}	**1.5**	1.45	–0.05	δ_3	–0.17	2.14	+0.64
d_3	**–1.5**	–1.39	+0.11	τ_3	0.93	–2.84	–1.34
a_{41}	**1.7**	4.69	+2.99	λ_4	0.70	1.72	+0.02
a_{42}	**0.7**	0.27	–0.43	δ_4	–0.03	0.43	–0.28
d_4	**1.2**	1.94	+0.74	τ_4	–0.60	0.87	–0.33
a_{51}	**1.9**	1.52	–0.38	λ_5	0.72	2.23	+0.33
a_{52}	**1.1**	1.58	+0.48	δ_5	0.18	1.31	+0.21
d_5	**0.0**	–0.24	–0.24	τ_5	0.11	–0.20	–0.20
ρ	**0.5**	0.39	–0.11	Σ_{12}	0.67	0.69	+0.19
μ_2	**0.2**	0.08	–0.12	μ_2	0.04	0.04	–0.16
σ_2	**1.0**	0.92	–0.08	Σ_{22}	0.94	0.97	–0.03
–ll		498.35					

Table 10. Parameter recovery of data generated with the bi–factor model (*N* = 250, *I* = 5).

Bi–	**True**	Estimates	Estimate	**CCFA**	Estimates	Translation	Translation
factor	**Values**		– True				– True
a_{11}	**1.1**	0.59	–0.51	λ_1	0.73	2.36	+1.26
a_{12}	**0.6**	1.21	+0.61	δ_1	–0.19	1.41	+0.81
d_1	**1.7**	1.75	+0.05	τ_1	–1.12	2.13	+0.43
a_{21}	**1.3**	1.21	–0.09	λ_2	0.59	1.28	–0.02
a_{22}	**1.0**	0.41	–0.59	δ_2	–0.04	0.43	–0.57
d_2	**–2.3**	–2.14	+0.16	τ_2	1.16	–1.48	+0.82
a_{31}	**1.5**	0.98	–0.52	λ_3	0.78	2.50	+1.00
a_{32}	**1.5**	1.58	+0.08	δ_3	0.11	1.06	–0.44
d_3	**–1.5**	–1.47	+0.03	τ_3	0.81	–1.53	–0.03
a_{41}	**1.7**	1.12	–0.58	λ_4	0.70	1.72	+0.02
a_{42}	**0.7**	1.26	+0.56	δ_4	0.03	0.43	–0.28
d_4	**1.2**	1.24	+0.04	τ_4	–0.72	1.04	–0.16
a_{51}	**1.9**	3.69	+1.79	λ_5	0.77	2.32	+0.42
a_{52}	**1.1**	1.74	+0.64	δ_5	0.09	0.91	–0.19
d_5	**0.0**	–0.43	–0.43	τ_5	0.08	–0.14	–0.14
ρ	**0.5**	0.24	–0.26	Σ_{12}	0.56	0.57	+0.07
μ_2	**0.2**	0.40	+0.20	μ_2	0.21	0.21	+0.01
σ_2	**1.0**	1.03	+0.03	Σ_{22}	0.96	0.98	–0.02
–ll		367.16					

Table 11. Parameter recovery of data generated with the bi–factor model (N = 500, I = 5).

Bi– factor	**True Values**	Estimates	Estimate – True	**CCFA**	Estimates	Translation	Translation – True
a_{11}	**1.1**	1.31	+0.21	λ_1	0.61	1.36	+0.26
a_{12}	**0.6**	0.22	–0.38	δ_1	–0.05	0.50	–0.10
d_1	**1.7**	1.79	+0.09	τ_1	–1.04	1.37	–0.33
a_{21}	**1.3**	1.43	+0.13	λ_2	0.67	1.58	+0.28
a_{22}	**1.0**	0.73	–0.27	δ_2	–0.03	0.41	–0.59
d_2	**–2.3**	–2.20	+0.10	τ_2	1.27	–1.76	+0.54
a_{31}	**1.5**	1.93	+0.43	λ_3	0.78	2.34	+0.84
a_{32}	**1.5**	0.96	–0.54	δ_3	0.07	0.79	–0.71
d_3	**–1.5**	–1.50	0.00	τ_3	0.89	–1.57	–0.07
a_{41}	**1.7**	2.16	+0.46	λ_4	0.78	2.18	+0.48
a_{42}	**0.7**	0.58	–0.12	δ_4	0.02	0.39	–0.31
d_4	**1.2**	1.21	+0.01	τ_4	–0.67	1.10	–0.10
a_{51}	**1.9**	3.58	+1.68	λ_5	0.81	2.68	+0.78
a_{52}	**1.1**	–0.24	–1.34	δ_5	0.08	0.94	–0.16
d_5	**0.0**	0.06	+0.06	τ_5	–0.01	0.02	+0.02
ρ	**0.5**	0.61	+0.11	Σ_{12}	0.55	0.64	+0.14
μ_2	**0.2**	0.09	–0.11	μ_2	0.12	0.12	–0.08
σ_2	**1.0**	0.85	–0.15	Σ_{22}	0.74	0.86	–0.14
$-ll$		2518.22					

Table 12. Parameter recovery of data generated with the bi–factor model (N = 100, I = 10).

Bi–	**True**		Estimate				Translation
factor	**Values**	Estimates	– True	**CCFA**	Estimates	Translation	– True
a_{11}	**1.1**	1.12	+0.02	λ_1	0.58	1.35	+0.25
a_{12}	**0.6**	0.48	–0.12	δ_1	–0.13	0.84	+0.24
d_1	**1.7**	1.69	–0.01	τ_1	–1.00	1.37	–0.33
a_{21}	**1.3**	0.97	–0.33	λ_2	0.68	1.64	+0.34
a_{22}	**1.0**	1.22	+0.22	δ_2	0.04	0.48	–0.52
d_2	**–2.3**	–2.07	+0.23	τ_2	1.19	-1.69	+0.61
a_{31}	**1.5**	1.46	–0.04	λ_3	0.86	4.21	+2.71
a_{32}	**1.5**	2.64	+1.14	δ_3	0.14	1.83	+0.33
d_3	**–1.5**	–1.99	–0.49	τ_3	1.10	–3.17	–1.67
a_{41}	**1.7**	2.56	+0.86	λ_4	0.79	2.83	+1.13
a_{42}	**0.7**	0.60	–0.10	δ_4	–0.15	1.39	+0.69
d_4	**1.2**	1.53	+0.33	τ_4	–0.77	1.62	+0.42
a_{51}	**1.9**	1.55	–0.35	λ_5	0.83	2.66	+0.76
a_{52}	**1.1**	1.21	+0.11	δ_5	–0.03	0.56	–0.54
d_5	**0.0**	0.01	+0.01	τ_5	–0.02	0.04	+0.04
a_{61}	**1.2**	1.08	–0.12	λ_6	0.73	1.88	+0.68
a_{62}	**1.4**	1.59	+0.19	δ_6	0.03	0.45	–0.95
d_6	**0.0**	0.32	+0.32	τ_6	–0.18	0.27	+0.27
a_{71}	**1.4**	2.35	+0.95	λ_7	0.75	2.02	+0.62
a_{72}	**0.8**	0.53	–0.27	δ_7	–0.04	0.54	–0.26

Bi–	**True**		Estimate				Translation
factor	**Values**	Estimates	– True	**CCFA**	Estimates	Translation	– True
d_7	**1.1**	2.00	+0.90	τ_7	–0.99	1.57	+0.47
a_{81}	**1.6**	1.27	–0.33	λ_8	0.80	2.37	+0.77
a_{82}	**1.2**	1.93	+0.73	δ_8	0.03	0.51	–0.69
d_8	**–1.4**	–1.78	–0.38	τ_8	1.00	–1.74	–0.34
a_{91}	**1.8**	2.40	+0.60	λ_9	0.82	5.36	+3.56
a_{92}	**1.3**	1.52	+0.22	δ_9	–0.26	3.33	+2.03
d_9	**–3.1**	–3.94	–0.84	τ_9	1.93	–7.42	–4.32
$a_{10,1}$	**2.0**	1.70	–0.30	λ_{10}	0.76	2.01	+0.01
$a_{10,2}$	**0.9**	1.40	+0.50	δ_{10}	0.01	0.26	–0.64
d_{10}	**2.9**	3.44	+0.54	τ_{10}	–1.78	2.77	–0.13
ρ	**0.5**	0.48	–0.02	Σ_{12}	0.67	0.70	+0.20
μ_2	**0.2**	0.12	–0.08	μ_2	0.07	0.07	–0.13
σ_2	**1.0**	1.12	+0.12	Σ_{22}	0.91	0.95	–0.05
$-ll$		881.96					

Table 13. Parameter recovery of data generated with the bi–factor model (N = 250, I = 10).

Bi– factor	**True Values**	Estimates	Estimate – True	**CCFA**	Estimates	Translation	Translation – True
$\boldsymbol{a_{11}}$	**1.1**	0.93	–0.17	$\boldsymbol{\lambda_1}$	0.54	1.30	+0.20
$\boldsymbol{a_{12}}$	**0.6**	0.57	–0.03	$\boldsymbol{\delta_1}$	–0.21	1.10	+0.50
$\boldsymbol{d_1}$	**1.7**	1.83	+0.13	$\boldsymbol{\tau_1}$	–1.10	1.56	–0.14
$\boldsymbol{a_{21}}$	**1.3**	1.50	+0.20	$\boldsymbol{\lambda_2}$	0.68	1.71	+0.41
$\boldsymbol{a_{22}}$	**1.0**	1.08	+0.08	$\boldsymbol{\delta_2}$	–0.08	0.71	–0.29
$\boldsymbol{d_2}$	**–2.3**	–2.25	+0.05	$\boldsymbol{\tau_2}$	1.25	–1.85	+0.45
$\boldsymbol{a_{31}}$	**1.5**	1.20	–0.30	$\boldsymbol{\lambda_3}$	0.76	2.09	+0.59
$\boldsymbol{a_{32}}$	**1.5**	1.84	+0.34	$\boldsymbol{\delta_3}$	–0.04	0.55	–0.95
$\boldsymbol{d_3}$	**–1.5**	–1.34	+0.16	$\boldsymbol{\tau_3}$	0.76	–1.23	+0.27
$\boldsymbol{a_{41}}$	**1.7**	1.46	–0.24	$\boldsymbol{\lambda_4}$	0.65	1.48	–0.22
$\boldsymbol{a_{42}}$	**0.7**	0.82	+0.12	$\boldsymbol{\delta_4}$	0.02	0.32	–0.38
$\boldsymbol{d_4}$	**1.2**	1.51	+0.31	$\boldsymbol{\tau_4}$	–0.85	1.14	–0.06
$\boldsymbol{a_{51}}$	**1.9**	2.12	+0.22	$\boldsymbol{\lambda_5}$	0.78	2.21	+0.31
$\boldsymbol{a_{52}}$	**1.1**	1.40	+0.30	$\boldsymbol{\delta_5}$	–0.03	0.49	–0.61
$\boldsymbol{d_5}$	**0.0**	0.11	+0.11	$\boldsymbol{\tau_5}$	–0.07	0.12	+0.12
$\boldsymbol{a_{61}}$	**1.2**	0.82	–0.38	$\boldsymbol{\lambda_6}$	0.68	1.71	+0.51
$\boldsymbol{a_{62}}$	**1.4**	1.65	+0.25	$\boldsymbol{\delta_6}$	0.08	0.71	–0.69
$\boldsymbol{d_6}$	**0.0**	0.22	+0.22	$\boldsymbol{\tau_6}$	–0.12	0.18	+0.18
$\boldsymbol{a_{71}}$	**1.4**	1.65	+0.25	$\boldsymbol{\lambda_7}$	0.69	1.89	+0.49
$\boldsymbol{a_{72}}$	**0.8**	1.03	+0.23	$\boldsymbol{\delta_7}$	–0.14	1.03	+0.23

Bi–factor	**True Values**	Estimates	Estimate – True	**CCFA**	Estimates	Translation	Translation – True
d_7	**1.1**	1.26	+0.16	τ_7	–0.70	1.13	+0.03
a_{81}	**1.6**	1.94	+0.34	λ_8	0.81	2.38	+0.78
a_{82}	**1.2**	1.95	+0.75	δ_8	0.01	0.29	–0.91
d_8	**–1.4**	–1.40	0.00	τ_8	0.76	–1.32	+0.08
a_{91}	**1.8**	1.22	–0.58	λ_9	0.75	2.08	+0.28
a_{92}	**1.3**	1.46	+0.16	δ_9	0.06	0.68	–0.62
d_9	**–3.1**	–2.74	+0.36	τ_9	1.68	–2.73	+0.37
$a_{10,1}$	**2.0**	1.95	–0.05	λ_{10}	0.73	3.04	+1.04
$a_{10,2}$	**0.9**	1.13	+0.23	δ_{10}	–0.30	2.28	+1.38
d_{10}	**2.9**	3.18	+0.28	τ_{10}	–1.66	4.06	+1.16
ρ	**0.5**	0.37	–0.13	Σ_{12}	0.77	0.71	+0.21
μ_2	**0.2**	0.04	–0.16	μ_2	0.04	0.04	–0.16
σ_2	**1.0**	0.96	–0.04	Σ_{22}	1.19	1.09	+0.09
$-ll$		2299.08					

Table 14. Parameter recovery of data generated with the bi–factor model ($N = 500$, $I = 10$).

Bi–factor	**True Values**	Estimates	Estimate – True	**CCFA**	Estimates	Translation	Translation – True
$\boldsymbol{a_{11}}$	**1.1**	1.27	+0.17	$\boldsymbol{\lambda_1}$	0.62	1.56	+0.46
$\boldsymbol{a_{12}}$	**0.6**	0.55	–0.05	$\boldsymbol{\delta_1}$	–0.16	1.01	+0.41
$\boldsymbol{d_1}$	**1.7**	1.61	–0.09	$\boldsymbol{\tau_1}$	–0.95	1.41	–0.29
$\boldsymbol{a_{21}}$	**1.3**	1.79	+0.49	$\boldsymbol{\lambda_2}$	0.72	1.86	+0.56
$\boldsymbol{a_{22}}$	**1.0**	0.69	–0.31	$\boldsymbol{\delta_2}$	0.05	0.58	–0.42
$\boldsymbol{d_2}$	**–2.3**	–2.50	–0.20	$\boldsymbol{\tau_2}$	1.40	–2.13	+0.17
$\boldsymbol{a_{31}}$	**1.5**	1.59	+0.09	$\boldsymbol{\lambda_3}$	0.75	2.10	+0.60
$\boldsymbol{a_{32}}$	**1.5**	1.27	–0.23	$\boldsymbol{\delta_3}$	–0.07	0.74	–0.76
$\boldsymbol{d_3}$	**–1.5**	–1.44	+0.06	$\boldsymbol{\tau_3}$	0.83	–1.37	+0.13
$\boldsymbol{a_{41}}$	**1.7**	1.91	+0.21	$\boldsymbol{\lambda_4}$	0.69	1.76	+0.06
$\boldsymbol{a_{42}}$	**0.7**	0.32	–0.38	$\boldsymbol{\delta_4}$	–0.08	0.72	+0.02
$\boldsymbol{d_4}$	**1.2**	1.11	–0.09	$\boldsymbol{\tau_4}$	–0.62	0.93	–0.27
$\boldsymbol{a_{51}}$	**1.9**	1.74	–0.16	$\boldsymbol{\lambda_5}$	0.74	2.01	+0.11
$\boldsymbol{a_{52}}$	**1.1**	0.77	–0.33	$\boldsymbol{\delta_5}$	–0.06	0.66	–0.44
$\boldsymbol{d_5}$	**0.0**	–0.23	–0.23	$\boldsymbol{\tau_5}$	0.13	–0.21	–0.21
$\boldsymbol{a_{61}}$	**1.2**	1.44	+0.24	$\boldsymbol{\lambda_6}$	0.72	1.78	+0.58
$\boldsymbol{a_{62}}$	**1.4**	1.21	–0.19	$\boldsymbol{\delta_6}$	0.01	0.25	–1.15
$\boldsymbol{d_6}$	**0.0**	–0.08	–0.08	$\boldsymbol{\tau_6}$	0.06	–0.09	–0.09
$\boldsymbol{a_{71}}$	**1.4**	1.34	–0.06	$\boldsymbol{\lambda_7}$	0.66	1.60	+0.20
$\boldsymbol{a_{72}}$	**0.8**	0.67	–0.13	$\boldsymbol{\delta_7}$	–0.07	0.64	–0.16

Bi–factor	**True Values**	Estimates	Estimate – True	**CCFA**	Estimates	Translation	Translation – True
d_7	**1.1**	0.95	–0.15	τ_7	–0.56	0.80	–0.30
a_{81}	**1.6**	1.72	+0.12	λ_8	0.77	2.08	+0.48
a_{82}	**1.2**	1.08	–0.12	δ_8	0.01	0.27	–0.93
d_8	**–1.4**	–1.43	–0.03	τ_8	0.85	–1.35	+0.05
a_{91}	**1.8**	2.28	+0.48	λ_9	0.80	2.37	+0.57
a_{92}	**1.3**	1.01	–0.29	δ_9	–0.03	0.51	–0.79
d_9	**–3.1**	–3.15	–0.05	τ_9	1.76	–3.06	+0.04
$a_{10,1}$	**2.0**	2.96	+0.96	λ_{10}	0.79	2.60	+0.60
$a_{10,2}$	**0.9**	0.46	–0.44	δ_{10}	0.11	1.09	+0.19
d_{10}	**2.9**	3.28	+0.38	τ_{10}	–1.60	3.10	+0.20
ρ	**0.5**	0.64	+0.14	Σ_{12}	0.69	0.72	+0.22
μ_2	**0.2**	0.21	+0.01	μ_2	0.19	0.19	–0.01
σ_2	**1.0**	0.90	–0.10	Σ_{22}	0.93	0.96	–0.04
–ll		4726.68					

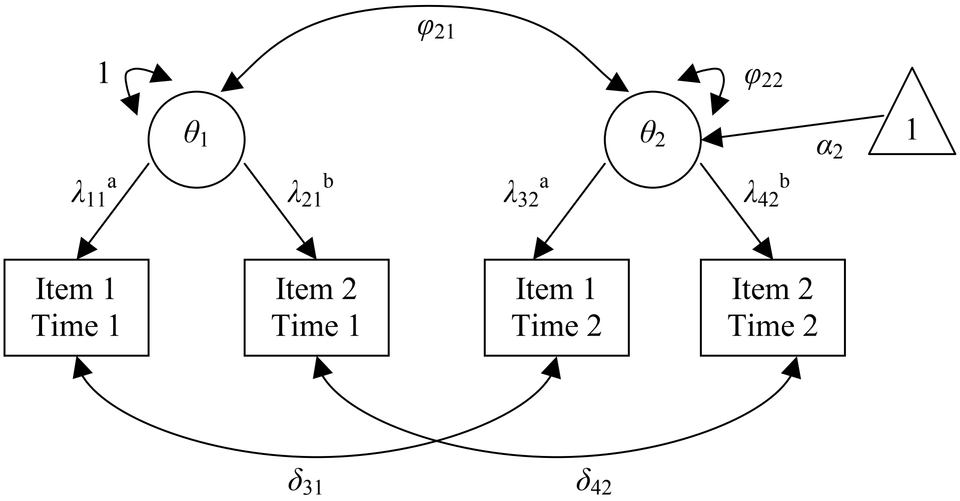
Table 15. Parameter estimates for the LCA model and the 2PL model for items from a psychological distress scale in the Context Study ($N = 3788$, $I = 4$).

LCA	Estimates (SE)	**2PL**	Wave 2 Estimates (SE)	Sum Score Statistics
a_1	2.19 (0.09)	a_1	2.14 (0.10)	
b_1	0.71 (0.03)	b_1	0.71 (0.03)	
κ_1	0.11 (0.02)			
a_2	2.00 (0.09)	a_2	2.00 (0.08)	
b_2	–0.44 (0.03)	b_2	–0.42 (0.03)	
κ_2	0.21 (0.02)			
a_7	2.46 (0.11)	a_7	2.39 (0.09)	
b_7	0.19 (0.02)	b_7	0.16 (0.02)	
κ_7	0.16 (0.02)			
a_9	3.28 (0.18)	a_9	3.22 (0.15)	
b_9	0.67 (0.03)	b_9	0.67 (0.03)	
κ_9	0.11 (0.02)			
ρ	0.50 (0.02)	ρ		0.60
μ_2	0.04 (0.03)	μ_2		0.08
σ_2	1.39 (0.05)	σ_2		1.06
$-ll$	15703.43			

Table 16. Parameter estimates for the bi–factor model and the CCFA model for items from a psychological distress scale in the Context Study ($N = 3788$, $I = 4$).

Bi–factor	Estimates (SE)	**CCFA**	Estimates (SE)	Translation (SE)
$\boldsymbol{a_{11}}$	2.13 (0.09)	$\boldsymbol{\lambda_1}$	0.80 (0.03)	2.84 (0.62)
$\boldsymbol{a_{21}}$	0.47 (0.18)	$\boldsymbol{\delta_1}$	0.13 (0.02)	1.28 (0.15)
$\boldsymbol{d_1}$	–1.53 (0.06)	$\boldsymbol{\tau_1}$	0.91 (0.04)	–1.90 (0.08)
$\boldsymbol{a_{12}}$	2.52 (0.22)	$\boldsymbol{\lambda_2}$	0.75 (0.03)	2.94 (0.47)
$\boldsymbol{a_{22}}$	1.47 (0.20)	$\boldsymbol{\delta_2}$	0.25 (0.02)	1.96 (0.18)
$\boldsymbol{d_2}$	1.09 (0.08)	$\boldsymbol{\tau_2}$	–0.49 (0.03)	1.13 (0.07)
$\boldsymbol{a_{17}}$	3.76 (0.38)	$\boldsymbol{\lambda_7}$	0.81 (0.03)	3.40 (0.63)
$\boldsymbol{a_{27}}$	–0.82 (0.26)	$\boldsymbol{\delta_7}$	0.18 (0.02)	1.78 (0.21)
$\boldsymbol{d_7}$	–0.67 (0.10)	$\boldsymbol{\tau_7}$	0.26 (0.03)	–0.64 (0.07)
$\boldsymbol{a_{19}}$	2.85 (0.13)	$\boldsymbol{\lambda_9}$	0.88 (0.05)	4.22 (1.72)
$\boldsymbol{a_{29}}$	0.39 (0.25)	$\boldsymbol{\delta_9}$	0.10 (0.02)	1.52 (0.27)
$\boldsymbol{d_9}$	–1.92 (0.09)	$\boldsymbol{\tau_9}$	1.21 (0.06)	–3.41 (0.17)
$\boldsymbol{\rho}$	0.61 (0.02)	$\boldsymbol{\Sigma_{12}}$	0.65 (0.03)	0.55 (0.03)
$\boldsymbol{\mu_2}$	0.07 (0.02)	$\boldsymbol{\mu_2}$	0.07 (0.02)	0.07 (0.02)
$\boldsymbol{\sigma_2}$	1.17 (0.04)	$\boldsymbol{\Sigma_{22}}$	1.40 (0.08)	1.18 (0.03)
–ll	15725.64			

Figure 1. Longitudinal IRT model using a bi-factor analysis approach (two items administered twice).



Note. Factor loadings with the same superscript are constrained to be equal.

Figure 2. Psychological distress items from the Understanding Adolescent Health Risk Behaviors survey.

How strongly do you agree or disagree with the following statements in describing how you have felt in the past 3 months?

	Strongly agree	**Agree somewhat**	**Neither**	**Disagree somewhat**	**Strongly disagree**
1. I had trouble getting my breath.	○	○	○	○	○
2. I got mad easily.	○	○	○	○	○
3. I felt sick in my stomach.	○	○	○	○	○
4. I was tired a lot.	○	○	○	○	○
5. I worried about what was going to happen.	○	○	○	○	○
6. I worried when I went to bed at night.	○	○	○	○	○
7. I often worried about bad things happening to me.	○	○	○	○	○
8. I hated myself.	○	○	○	○	○
9. I was a bad person.	○	○	○	○	○
10. I did everything wrong.	○	○	○	○	○

REFERENCES

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis, 95,* 1-22.

Bock, R. D. (1985). *Multivariate Statistical Methods in Behavioral Research.* Scientific Software International: Chicago.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society* (Series B)*, 39,* 1-38.

Douglas, J. A. (1999). Item response models for longitudinal quality of life data in clinical trials. *Statistics in Medicine, 18*, 2917-2931.

Ferrando, P. J., Lorenzo, U., & Molina, G. (2001). An item response theory analysis of response stability in personality measurement. *Applied Psychological Measurement, 25,* 3-17.

Fox, J. P., & Glas, C. A. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169-191.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5,* 299-314.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165-172.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen, & H. Wainer (Eds). *Test scoring*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM Algorithm. *Journal of the American Statistical Association, 86*, 899-909.

Mooijaart, A., & van der Heijden, P. G. M. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika, 57,* 261-269.

Muthén, L. K., & Muthén, B. O. (2003). *Mplus: statistical analysis with latent variables (version 3.13)* [Computer software]. Los Angeles, CA: Muthén & Muthén.

Oranje, A. (2003, April). *Comparison of estimation methods in factor analysis with categorized variables: Applications to NAEP data*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement, 30*, 100-120.

Raudenbush, S. W., & Sampson, R. J. (1999). Econometric: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

SAS Institute, Inc. (2005). *SAS Proprietary Software, Version 9.1* [Computer software]. Cary NC: SAS Institute, Inc.

Thissen, D., Chen, W-H, & Bock, R. D. (2003). *Multilog (version 7.0)* [Computer software]. Lincolnwood, IL: Scientific Software International.