Exploring Alternate Latent Trait Metrics with the Filtered Monotonic Polynomial IRT Model

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Leah Marie Feuerstahler

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Niels Waller

Thesis Advisor

July, 2016

© Leah Marie Feuerstahler 2016

ALL RIGHTS RESERVED

Abstract

Item response theory (IRT) is a broad modeling framework that makes precise predictions about item response behavior given individuals' locations on a latent (unobserved) variable. If the item-trait regressions, also known as item response functions (IRFs), are monotonically increasing and if assumptions about unidimensionality and local independence are satisfied, then examinees can be ordered uniquely on the latent trait. Scales that satisfy these three assumptions can be transformed monotonically without altering scale properties—that is, they define an ordinal-level scale (Stevens, 1946). When fitting an IRT model, however, the scale of the latent variable—that is, its location and interval spacing—must be identified by introducing extra assumptions. In practice, the scale is identified by specifying either the parametric form of the IRF (parametric IRT) or the distribution of the latent trait (nonparametric IRT).

Filtered monotonic polynomial IRT (FMP) has been proposed as a type of nonparametric IRT method (Liang & Browne, 2015), but shares important properties with parametric methods. In this dissertation, it is demonstrated that any IRF defined within the FMP framework can be re-expressed as another FMP IRF by taking linear or nonlinear transformations of the latent trait. A general form for these transformations is presented in terms of matrix algebra.

Finally, I propose a composite FMP IRT model in which nonlinear transformations of the latent trait are modeled explicitly by a monotonic composite function.

I argue that the composite model offers many advantages over existing methods. First, the composite FMP model narrows the methodological gap between parametric and nonparametric item response models, allowing for item banking and adaptive testing within a flexible modeling framework. Second, this composite model suggests a sequential NIRT curve-fitting method that allows users to explore both alternate (e.g., non-normal) latent densities and flexible IRF shapes. Finally, the composite FMP model allows users to explore and employ alternate scalings of the latent trait without sacrificing the methodological advantages of parametric models.

Contents

List of Tables

List of Figures

Chapter 1

Constructing Measurement Scales with Item Response Theory

1.1 Assumptions of Item Response Modeling

As with any statistical model, item response models draw inferences from data only after making a certain number of assumptions. In item response theory (IRT), the most commonly used models share three assumptions: unidimensionality, local independence, and monotonicity. The first assumption, unidimensionality, requires that individual differences in item response probabilities are wholly attributable to individual differences on exactly one latent variable. The second assumption, local independence, is a direct consequence of unidimensionality. Local independence states that, conditional on the latent trait, responses to a series of test items are independent of each other. The third assumption is that all item response functions (IRFs) are monotonically increasing functions of the

latent trait. Graphically, the monotonicity assumption states that all IRFs are monotonically increasing functions of the latent variable θ. Hereafter, the term "monotonic" will always signify monotonically *increasing* because a monotonically decreasing IRF can be made monotonically increasing by reverse keying the item. Any set of items that satisfies the unidimensionality, local independence, and monotonicity assumptions, by definition, satisfies Mokken's (1971) monotone homogeneity model (MHM). An important feature of the MHM is that it implies a unique ordering of individuals on the latent trait. In terms of Stevens' (1946) well-known taxonomy, scales that satisfy the three MHM assumptions measure examinees on an ordinal scale.

The three MHM assumptions are sufficient, but not necessary, conditions for ordinal-level measurement. Accordingly, many alternative models have been developed that violate these assumptions (Ip, 2002; Reckase, 2009; Roberts & Laughlin, 1996). Although it may be desirable to construct scales that measure exactly one trait, there is often a tradeoff between constructing unidimensional tests and constructing tests that capture the breadth of psychological constructs (see Ip & Chen, 2015). Instead of developing tests that measure narrowly defined constructs, it may be preferable to employ multidimensional IRT models that simultaneously measure several latent traits (see Reckase, 2009, for an overview of multidimensional IRT models). Alternatively, there may be reasons to use unidimensional IRT models even when data are multidimensional, particularly if the data are "unidimensional enough" (Bonifay, Reise, Scheines, & Meijer, 2015) or if item and person parameter estimates are not substantially distorted by multidimensionality (Reise, Cook, & Moore, 2015). Similarly, the assumption of local independence

may be too strong to hold exactly in real data. As such, locally dependent IRT models have been proposed to accommodate violations of this assumption (Ip, 2002; Ip, Wang, de Boeck, & Meulders, 2004). Note however that, because local independence is a consequence of unidimensionality, locally dependent models are closely related to multidimensional models (Ip, 2010). Finally, monotonic IRFs are commonly assumed in item response theory, but they are not necessarily appropriate for all item response data. Specifically, monotonicity implies a *dominance* response process wherein increased levels of the latent trait lead to higher response probabilities (Stark, Chernyshenko, Drasgow, & Williams, 2006). In contrast, *ideal-point* response processes (Coombs, 1964) imply bell-shaped response functions. Under an ideal-point model, the closer a person's latent trait score is to the peak of the IRF, the higher the probability of a keyed item response. Ideal-point models, also known as unfolding models (Davison, 1977; Roberts & Laughlin, 1996), have been found to provide comparable or superior fit and more interpretable scores than dominance response models when applied to some personality questionnaires (Stark et al., 2006) and vocational interest inventories (Tay, Drasgow, Rounds, & Williams, 2009). Despite their promise, ideal-point models are seldom used and not thoroughly understood (see Reise, 2010, for a critique of ideal-point models).

1.2 Scale Identification

An important property of ordinal-level scales, such as scales that satisfy the MHM, is that monotonic score transformations do not alter the properties of the scale.

Lord (1975) demonstrated that for item response models, monotonic scale transformations do not alter model predictions, so long as the item response functions are also transformed appropriately. In other words, for every item response model, a monotonic transformation of the latent trait parameter results in an equally admissible item response model. Put more formally, suppose an item response model follows the MHM such that the probability P of a keyed response to item i equals

$$P_i(\theta) = f(\theta), \tag{1.1}$$

where f is a monotonically increasing function bounded by 0 and 1. For any strictly monotonic transformation, h^{-1}, of the latent trait[1],

$$\theta^\star = h^{-1}(\theta), \tag{1.2}$$

an alternate item response model exists such that

$$P_i(\theta) = P_i^\star(\theta^\star) \tag{1.3}$$

where

$$P_i^\star(\theta^\star) = f[h(\theta^\star)]. \tag{1.4}$$

The relation shown in Equation 1.4 holds for any monotonically increasing (and therefore, invertible) item response function f. Without additional evidence, there

[1]Because h^{-1} is a strictly monotonic function, it is guaranteed to have an inverse, and thus the function h is also strictly monotonic and invertible. The inverse transformation, h^{-1}, is used in the current definition for notational consistency.

is no psychological reason to prefer the θ scale to the $\theta^{\star}$ scale (Lord, 1975, 1980, p. 84).

The relation shown in Equation 1.4 implies that, for any monotonic transformation of θ, an item response model exists that makes identical predictions as the original model. This fact suggests that the scaling of the latent variable—that is, the choice of monotonic transformation of examinee scores—is, to some extent, arbitrary. However, because item response models make point predictions about latent trait values, fitting an item response model requires identifying a particular scaling of the latent variable. In other words, scale identification is a necessary condition for IRT model identification.

In item response theory, the scale is usually determined in one of two ways. The researcher may specify in advance either the functional form of the IRF or the distribution of the latent trait (Thissen, 2009). Broadly speaking, this distinction corresponds to the difference between parametric item response theory (PIRT) and nonparametric item response theory (NIRT). Specifically, PIRT identifies the scale by specifying a particular mathematical form for the IRF whereas NIRT identifies the scale by specifying the latent trait distribution.

When fitting a PIRT model, the researcher assumes that the chosen parametric model fits all scale items. This assumption is made in addition to the MHM assumptions. As such, a data set that satisfies the MHM assumptions need not fit the chosen PIRT model. If, however, a particular PIRT model is appropriate for all scale items, the scale is identified up to linear transformations of the latent trait. In other words, for a given IRF functional form, only linear transformations of the latent trait are permissible without altering model predictions. Because of this

fact, it is often stated that PIRT models provide interval-level measurement (e.g., Sijtsma & Molenaar, 2002), an issue that will be revisited in a later section. The linear indeterminacy of the latent trait is a well-known property of PIRT models that poses a model identification problem. This identification problem is often resolved by standardizing the latent variable during item parameter estimation.

In contrast to PIRT models, NIRT models are identified by specifying the latent trait distribution.[2] The latent trait distribution is often specified by transforming initial trait estimates computed from the observed data. The simplest of these initial estimates is the observed sum score. The observed sum score is an estimate of the true score in classical test theory, and the classical true score is a permissible transformation of the θ metric under the assumptions of the MHM model. With this justification, many authors have fit nonparametric item response models by conditioning on observed sum scores. For example, Mokken scale analysis (1971) is usually based on the analysis of observed sum scores. Other authors have transformed the observed sum score distribution (or proportion-correct distribution) to follow a standard normal distribution (e.g., Ramsay, 1991). One problem with using observed sum scores is that, for a test with I items, only $I+1$ unique scores can be observed, even though true scores are defined on the continuous interval $[0, I]$. To break ties, Ramsay (1991) jittered the normalized sum scores (i.e., broke ties by adding a small amount of random noise to each score). Another problem with sum scores is that they treat every item equally. That is, sum

[2]Throughout, the term "NIRT" refers to methods that aim to flexibly fit IRFs. Here, the term encompasses quasi-parametric methods, as described by Ramsay (1991 pp. 612–613), in which estimated model parameters are not interpreted outside of the model and which are able to fit a wide variety of (possibly non-monotonic) IRFs.

scores do not differentiate between items of different quality (i.e., discriminability or difficulty). To circumvent the item quality and discreteness problems associated with sum scores, it has recently been recommended (Lathrop, 2015; Liang, 2007; Liang & Browne, 2015) to use the (normalized) first principal components scores as initial estimates of the latent trait. Importantly, both sum scores and principal components scores, along with their transformations, are only estimates of the latent trait.

1.3 Estimating Item Response Models

In IRT, the scale must be identified prior to parameter estimation. A thorough review of PIRT parameter estimation techniques is given by Baker and Kim (2004), and more detail about the technical aspects of item calibration methods is given in a later section. Below I focus on how various estimation methods identify the latent trait metric. In particular, I distinguish between fixed-effects and random-effects treatments of the latent trait. In IRT, fixed-effects estimation draws inferences about trait scores for particular examinees, and random-effects estimation considers examinees in the calibration sample to be a random sample from a population of examinees.

1.3.1 Fixed-effects

When using fixed-effects item calibration methods, estimates of both the item and person parameters are obtained. Fixed-effects estimation for NIRT models usually involves conditioning on initial estimates of the latent trait. The initial estimates

may be item total scores, first principal component scores, or functions of either quantity (e.g., normalized scores). Estimated curves are obtained by treating these initial estimates as known quantities. After the curves are estimated, updated latent trait estimates may be obtained by treating the estimated IRFs as known.

For PIRT models, perhaps the most obvious method for employing fixed-effects estimation is to maximize the complete-data likelihood. This approach is known as joint maximum likelihood (JML). Because person parameters are estimated along with item parameters, JML makes no assumptions about the shape of the latent trait distribution. In practice, JML is usually implemented using a two-stage approach known as the Birnbaum paradigm (Birnbaum, 1968, p. 420; see also Baker & Kim, 2004, p. 87). In the first stage, item parameters are updated by treating person parameter estimates as fixed, and in the second stage, person parameter estimates are updated by treating item parameter estimates as fixed. These two steps iterate until parameter estimates stabilize.

A fundamental problem with JML estimation is that the item parameter estimates have not been proven to be consistent (Neyman & Scott, 1948). This problem occurs because the total number of parameters to be estimated increases as sample size increases. In the Rasch model only, consistent estimates of the item parameters are available by conditioning on the number-correct score (Andersen, 1973). This method is available because the number-correct score is a sufficient statistic for the latent trait in the Rasch model. For non-Rasch models, sufficient statistics for the latent trait do not exist, and so the person parameters cannot be eliminated from the likelihood function. Thus, the person parameters are incidental when updating item parameter estimates (Harwell, Baker, & Zwarts, 1988).

These problems with JML estimation have been the primary impetus for the development of newer item calibration techniques, and Baker and Kim (2004, p. 108) predicted that over time JML will be replaced by random-effects methods.

1.3.2 Random-effects

Some researchers (e.g., Takane & de Leeuw, 1987) have argued that fixed-effects methods are inappropriate in the context of IRT. Specifically, when estimating item parameters, item characteristics are of greater interest than than the characteristics of individuals in the calibration sample. From this perspective, it is appropriate to consider *random-effects* estimation in which individuals are considered to be a random sample from a population. The random-effects approach overcomes many of the technical problems associated with JML. However, in random-effects estimation, the population distribution is specified by the researcher, which adds another model assumption that is not directly testable.

In random-effects item parameter estimation, the marginal likelihood of the item parameters is maximized. The marginal likelihood is obtained by integrating over a distribution of the latent trait (Bock & Lieberman, 1970). This approach, known as marginal maximum likelihood (MML) estimation, is usually implemented using an application of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to item response models, as described by Bock and Aitkin (1981). In MML, the latent trait distribution is specified as a Bayesian prior distribution that is integrated out of the likelihood. The choice of latent trait prior distribution, such as a standard normal distribution, is sufficient

to determine the scale location and unit. Additionally some NIRT models can be estimated via MML (e.g., Falk & Cai, 2016; Rossi, Wang, & Ramsay, 2002).

When fitting NIRT models via random-effects estimation, the user specifies the latent trait distribution and flexibly estimates the IRF shape. In contrast, fitting PIRT models via random-effects methods requires the user to specify both the latent trait distribution and the IRF functional form. When both sets of restrictions are put in place, such as with MML estimation of parametric models, the scale is over-determined (Thissen, 2009). In MML, the restriction on the latent trait distribution takes the form of a Bayesian prior that can, in theory, be overwhelmed by the data. The extent to which the prior is overwhelmed by data in real data estimation has been investigated by numerous authors with mixed results (cf. Woods, 2015). Importantly, when the prior distribution shape does not match the latent trait distribution shape (as determined by the parametric IRF functional form), there are competing identification restrictions that may negatively affect the accuracy of estimated item parameters. When the model is over-identified in this manner, there are two strategies for increasing the flexibility of the model: (a) use NIRT methods to flexibly estimate the IRF, or (b) flexibly estimate the latent density. Before describing these approaches, I will first overview common parametric item response models.

1.4 Parametric Models

When using parametric models, the latent trait variable θ is, by definition, the scale on which all IRFs have the mathematical forms that are specified by the researcher (Lord, 1975, p. 205). A number of parametric item response models for binary item responses have been proposed that meet the three MHM assumptions. Note that in the remainder of this work, attention will be focused on IRT models for dichotomously scored item responses. Of these models, the earliest are based on the normal cumulative distribution function. For example, the two-parameter normal ogive model (2PNO; Lawley, 1943, 1944), specifies the probability of responding in a keyed direction as

$$P_i(\theta) = \Phi[a_i(\theta - b_i)], \tag{1.5}$$

where Φ denotes the normal cumulative distribution function, a_i is the item discrimination parameter, and b_i is the item difficulty parameter for item i. Computationally, it is more convenient to work with a logistic approximation to the normal distribution (Birnbaum, 1968). The two-parameter logistic model (2PL) gives

$$P_i(\theta) = \{1 + \exp[-Da_i(\theta - b_i)]\}^{-1}, \tag{1.6}$$

where D is a scaling constant. When $D = 1.702$, the curves defined by Equations 1.5 and 1.6 trace highly similar curves for the same a_i and b_i parameters. Part of the appeal of the 2PNO and 2PL is that the a_i and b_i parameters are interpretable

in terms of item characteristics (see Lord & Novick, 1968, pp. 366–368). For example, on ability tests, items with higher b_i values (holding a_i constant) are more difficult for all examinees. Moreover, the 2PNO can be derived by making several assumptions about the item response process (Lord & Novick, 1968, pp. 370–371).[3]

The item response function defined in Equation 1.6 is sometimes referred to as the difficulty/discrimination parameterization of the 2PL. An alternative parameterization that will be important later in this work is the slope-intercept parameterization, written

$$P_i(\theta) = \{1 + \exp[-(b_{0i} + b_{1i}\theta)]\}^{-1}, \tag{1.7}$$

where b_{0i} is the item intercept and b_{1i} is the item slope for item i. Note that this parameterization is identical to the logistic regression of the binary item responses on θ. In other words, if θ values are known, estimating the item parameters for the slope-intercept parameterization of the 2PL is equivalent to fitting a logistic regression model with one predictor.

Many argue that the 2PL cannot adequately describe item response behavior

[3]Lord and Novick's (1968, pp. 370–371) derivation is as follows. Suppose that observed binary item responses $\boldsymbol{y}_i$ are a strict dichotomization of an unobserved continuous variable $\tilde{\boldsymbol{y}}_i$. If

1. The regression of $\tilde{\boldsymbol{y}}_i$ on θ is linear,
2. the conditional distribution of $\tilde{\boldsymbol{y}}_i$ given θ is normal with constant variance $\sigma^2_{\tilde{\boldsymbol{y}}_i|\theta}$, and
3. $\sigma^2_{\tilde{\boldsymbol{y}}_i|\theta}$ is independent of θ,

then the nonlinear regression of $\boldsymbol{y}_i$ on θ follows the 2PNO. This derivation makes the same base assumptions as the MHM, but does not assume population distributions for θ or $\tilde{\boldsymbol{y}}_i$. A limitation of this derivation is that it makes assumptions about unobservable quantities, and Lord and Novick admit that this derivation is one that "some theorists find interesting and others do not" (p. 370).

on ability and achievement tests. Specifically, it has been argued that on multiple-choice tests, examinees might guess the correct answer regardless of their ability level. To model this phenomenon, the three-parameter logistic model (3PL; Birnbaum, 1968) has been proposed. The 3PL IRF is written

$$P_i(\theta) = c_i + (1 - c_i) \left\{1 + \exp[-a_i(\theta - b_i)]\right\}^{-1}, \tag{1.8}$$

where the c_i parameter determines the lower asymptote value of the item response function and other quantities are as previously defined. The 2PL and 3PL are among the most commonly used item response models for binary item responses. Despite the popularity of these item response models, they need not provide the most plausible parametric models for item response behavior in all content domains. For example, the four-parameter logistic model (4PL; Barton & Lord, 1981), written

$$P_i(\theta) = c_i + (d_i - c_i) \left\{1 + \exp[-a_i(\theta - b_i)]\right\}^{-1}, \tag{1.9}$$

modifies the 3PL by adding a flexible upper asymptote parameter, d_i. By allowing IRF upper asymptotes to be less than one, the model can accommodate response "slips" by high ability examinees (Rulison & Loken, 2009) on cognitive tests and better characterize response behavior on psychopathology items (Reise & Waller, 2003; Waller & Feuerstahler, 2016). Yet another alternative IRF shape was proposed by Samejima (2000), who argued that point-symmetric models such as the 2PL are inappropriate when the psychological response process is more complex at some θ levels than at other θ levels. Instead, she proposed the logical positive

exponent (LPE) model, written

$$P_i(\theta) = [\tilde{P}_i(\theta)]^{\xi_i}, \tag{1.10}$$

where ξ_i is an acceleration, or item complexity, parameter and $\tilde{P}_i(\theta)$ is the 2PL or 2PNO. It has been argued (Samejima, 2000) that the LPE model is appropriate for tasks or items that involve applying a complex combinations of skills or strategies. The complexity of the task or item is accounted for by ξ_i, wherein increasing ξ_i increases the θ value at which the IRF slope is maximum. It has been suggested (Bolt, Deng, & Lee, 2014) that ξ_i is positively related to the number of subprocesses that must work in conjunction to elicit a keyed response. For example, suppose that solving a math problem involves multiple steps, and thus multiple opportunities to err. A correct answer to the math problem occurs only when all steps are completed correctly. In this scenario, it has been argued (Bolt et al., 2014) that an LPE model with $\xi_i > 1$ will better represent the probability of success on this math problem than the 2PL (i.e., an LPE model with $\xi_i = 1$).

Another parametric model worthy of brief mention is the Rasch model (Rasch, 1960/1980). This model, sometimes also called the one-parameter logistic model (1PL), contains one item-level parameter that is interpreted as item difficulty. The 1PL is written

$$P_i(\theta) = \{1 + \exp[-(\theta - b_i)]\}^{-1} \tag{1.11}$$

and also satisfies the assumptions of the MHM. The Rasch model boasts several advantages over higher parameterized models. For example, the Rasch model

satisfies invariant item ordering—that is, items can be uniquely ordered by difficulty such that this ordering holds for an examinee at any θ level. Also in this model, observed sum scores are a sufficient statistic for the latent trait (Rasch, 1960/1980), a fact that simplifies model estimation (Andersen, 1973). Moreover, the Rasch model can be considered a probabilistic form of additive conjoint measurement (Perline, Wright, & Wainer, 1979), which places the Rasch model in the context of representational measurement (Luce & Tukey, 1964). However, the Rasch model belongs to a different school of thought than the one employed in this paper. Broadly speaking, advocates of the Rasch model prefer to construct tests by carefully selecting items that fit the model. Advocates of more complex IRT models prefer to alter the model to fit characteristics of the data. In this paper, the latter attitude is taken; for items satisfying the MHM assumptions, an IRT model that fits a given set of items is preferred over a model that follows a particular parametric form.

1.5 Nonparametric Models

In contrast to parametric IRT models, nonparametric IRT models freely estimate IRFs conditional on the latent trait. To identify a NIRT model, therefore, the latent trait must be identified prior to model fitting. This is accomplished by specifying the latent trait distribution. Importantly, any univariate continuous distribution of latent trait scores can be transformed to follow any other univariate continuous distribution (see Duncan & MacEachern, 2008, pp. 46–47). For this reason, any scale that satisfies the MHM can be modeled using NIRT methods,

but not all scales that satisfy the MHM fit PIRT models. Although NIRT models are more general than PIRT models, NIRT is not always used as an alternative to PIRT. One characteristic feature of the NIRT literature is a strong tradition of, and readily available methods for, checking model assumptions. The use of NIRT as an assumption-checking tool is illustrated by Meijer and Baneke (2004), and Stout (2001) argued that NIRT-type analyses are an important precursor to determine whether PIRT methods are appropriate for a given data set. Sijtsma and Meijer (2007) distinguish this exploratory use of NIRT from confirmatory uses of NIRT that are used to draw inferences about data structure, items, and persons. In this paper, I am primarily concerned with this latter use of NIRT. In other words, this work is more concerned with the use of NIRT as a viable alternative to parametric models. From this perspective, the main advantage of using NIRT is one of model-data fit. To justify the use of NIRT from this perspective, it must be demonstrated than NIRT models fit significantly better than PIRT methods. Moreover, it is not sufficient to demonstrate that parametric models provide poor fit to item response data—it must be shown that there does not exist a scale on which all items follow the same (e.g., logistic) form, as is assumed in PIRT models.

To date, few studies have empirically compared NIRT and PIRT models in terms of fit. One such study was conducted by Chernyshenko, Stark, Chan, Drasgow, and Williams (2001). These authors compared the fit of the 2PL and 3PL to nonparametrically estimated curves on several personality measures. These researchers found that parametric models fit some scales well, but not others. One caveat to this type of research is that, in relatively small samples, there may be

little power to determine whether the parametric model fits. Nevertheless, it is plausible that NIRT provides superior fit in moderate to large data sets and may lead to less biased estimates of the IRFs (cf. Xu & Douglas, 2006).

Several NIRT models have been proposed. Perhaps the most widely used NIRT method is Ramsay's (1991) kernel smoothing. This method estimates a smooth IRF by taking a weighted average of responses to a given item, where weights are assigned according to a pre-specified kernel function and bandwidth parameter. With this method, estimated IRFs are not constrained to be monotonically increasing. Thus, kernel smoothing may be used to investigate severe violations of monotonicity. Applications of kernel smoothing have primarily been exploratory. A notable exception is the work of Xu and Douglas (2006), who applied kernel smoothing in a confirmatory manner in computerized adaptive testing. One reason why kernel smoothing is less suited for confirmatory data analysis is that fitted IRFs are not functions of estimated parameters. This means that the estimated IRFs cannot be described by a small number of parameters, for example, when scoring examinees that are not in the calibration sample. Other NIRT methods, termed "quasi-parametric" methods (Ramsay, 1991), are based on parameter estimation. In contrast to PIRT models, quasi-parametric models allow for highly flexible estimated IRFs and include parameters that are not meant to be interpreted outside the model. That is, quasi-parametric models include item parameters that are not interpretable in terms of psychological processes but that serve to fit a wide variety of IRF shapes. Quasi-parametric item response models include monotone splines (Ramsay & Abrahamowicz, 1989; Ramsay & Winsberg, 1991), penalized maximum likelihood (Rossi et al., 2002), and multilinear formula

scoring (Drasgow, Levine, Williams, McLaughlin, & Candell, 1989; Levine, 1984). Each of these methods characterizes the item response function as a linear combination of basis functions. Specifically, multilinear formula scoring models the IRF as the weighted sum of orthogonal functions such as orthogonal polynomials (see Chernyshenko, et al., 2001). Monotone splines, as the name suggests, are constructed from piecewise polynomial functions that are constrained to be monotonically increasing. In the penalized maximum likelihood approach, monotone splines are constructed to model the logit probability of a keyed response (the name "penalized maximum likelihood" comes from the estimation method recommended by these authors who employ an EM algorithm-based estimation method with a smoothness penalty).

More recently, at least two classes of NIRT models have been proposed that begin to bridge the gap between parametric and nonparametric IRT. First, Duncan and MacEachern (2008, 2013) proposed a nonparametric Bayesian approach to IRT. Specifically, these authors proposed two nonparametric Bayesian models, one that allows flexible estimation of the latent density and the other that allows flexible estimation of the IRFs. These authors view their work as a compromise between flexibly estimating the latent trait distribution and flexibly estimating the IRF shape. This work begins to narrow the methodological gap between the two disparate sets of methods, but it requires a choice between whether to explore alternate latent densities or alternate IRFs.

A second recently developed NIRT model is the filtered monotonic polynomial IRT model (FMP; Liang, 2007; Liang & Browne, 2015). The FMP model is similar to (polynomial) spline-based quasi-parametric methods, but instead of

using splines, for each IRF, FMP applies the same polynomial function across the θ continuum. At first blush, FMP may appear to be a cruder, less flexible approach to flexible item curve fitting. However, FMP curves can fit any monotonically increasing IRF to an arbitrary degree of precision (using a polynomial function of a sufficiently high degree). Moreover, the FMP model reduces exactly to the 2PL in its simplest form. In this case, item parameters may be interpreted as 2PL item parameters and item banking, adaptive testing, and other parametric methods may be applied in the usual manner. The FMP model is the major focus of this work, and more information on model properties will be given in a later section.

1.6 Flexible Latent Densities

Recall that when parametric IRT models are estimated using MML, the scale is identified by specifying both the parametric form of the IRF and the latent trait distribution. Although both identification constraints may be necessary for model convergence in small samples, these specifications may be too restrictive in moderate to large samples. There are two distinct ways to relax these restrictions. First, it is possible to specify the latent trait distribution and freely estimate the IRFs using NIRT. Alternatively, the latent trait density may be estimated conditional on a chosen parametric model. It has been argued (e.g., Woods & Thissen, 2006) that it is preferable to flexibly estimate the latent trait density rather than flexibly estimate the IRF. Specifically, if the latent trait density is estimated, then item

calibration produces item parameter estimates that follow the usual interpretation, and well-established PIRT methods such as trait estimation, item banking, and adaptive testing, may be applied in a straightforward manner. As argued by Woods and Thissen (2006, p. 282), "there is sufficient value in classical IRT to justify the alternative [to NIRT] of estimating $g(\theta)$ with logistic IRFs." Moreover, it is often unlikely that latent trait scores are normally distributed in a given sample. For example, if the calibration sample is a mixture of populations or if relevant symptoms are likely to occur in small numbers of subjects, normality is not a viable assumption. Although it is possible to specify a distribution other than standard normal (e.g., beta$(10, 2)$), usually not enough is known about the sample to confidently specify an alternate latent distribution.

When the true latent trait distribution (conditional on the model) does not match the prior distribution in MML estimation, the item and person parameters may be biased. There is mixed evidence as to how robust item and person parameter estimates are to misspecification of the prior trait distribution. Several researchers have explored the effects of non-normality of θ on item and person parameter estimates for various IRT models. Many researchers (e.g., Kirisci, Hsu, & Yu, 2001; Roberts, Donoghue, & Laughlin, 2002) have concluded that latent density misspecification has little to no effect on item or person parameter recovery. Others have found biased item and person parameter estimates particularly when the latent density is skewed (Reise & Yu, 1990; Seong, 1990; Stone, 1992), when items have low discriminations (Reise & Yu, 1990), when too few quadrature points are used (Seong, 1990), and for short tests (Stone, 1992). Moreover, bias in estimated trait scores tends to be greatest at the distribution extremes.

For skewed distributions, if there are more examinees with trait scores in regions where the test does not discriminate, there will be decreased variability in response patterns, exacerbating bias (Roberts et al., 2002). Woods (2015), based on a thorough review of the literature, concluded that biased item parameter estimates occur primarily when latent distributions are skewed and when items have extreme difficulties. Similarly, estimates of extreme latent trait scores tend to be biased when the latent distribution is skewed. Thus, when the latent distribution is misspecified, methods that flexibly estimate the latent trait distribution may provide superior model-data fit.

Early methods to estimate the latent trait distribution were proposed by Anderson and Madsen (1977) for the Rasch model, Bock and Aitkin (1981), and Mislevy (1984). These methods are not ideal because they require estimating a large number of parameters in addition to the computationally intensive problem of item parameter estimation. Nevertheless, these methods can produce more accurate item and person parameter estimates when the latent distribution is incorrectly specified as normal (e.g., Woods, 2007a). Modern approaches to flexible latent density estimation include Ramsay curves (Woods & Thissen, 2006), Davidian curves (Woods & Lin, 2009), and Johnson curves (van den Oord, 2005). These methods have in common that they introduce only a small number of additional model parameters, produce smooth estimates of the latent density, and reduce to the standard normal density when appropriate. Numerous studies have shown that these modern methods also provide more accurate item and person parameter estimates than incorrectly specifying a normal latent density (van den Oord, 2005; Woods, 2007b; Woods, 2008b), and tend to perform better than older

methods for estimating the latent density (Woods, 2008b; Woods & Lin, 2009). In many cases, Ramsay curves perform similarly to Davidian curves (Woods & Lin, 2009), although Woods (2015) recommends Davidian curves because they have been extended to multidimensional models and are simpler to implement.

Perhaps the greatest advantage of latent density estimation over NIRT is that it does not require learning a new IRT model. Instead, commonly used parametric item response techniques apply in a straightforward manner to tests calibrated with methods that estimate the latent trait distribution. When applying item parameter estimates to examinees not in the calibration sample, there is no difference in implementation between the purely parametric and the flexible latent density approaches. However, the density estimation methods assume correct model specification, and non-normality can be incorrectly found when the model is not correctly specified (Woods, 2008a). Another disadvantage to the flexible latent density methods is that there is no guarantee of fit. If not all items follow the specified parametric shape for some distribution of θ, the model will not fit the data. Whereas NIRT methods are guaranteed to model any IRF that satisfies the MHM, there is no such guarantee for parametric item response theory, even when the latent densities are estimated.

1.7 Interval or Ordinal?

At this juncture, it is necessary to address a contentious issue in latent trait theory: the measurement level of the latent trait. In Stevens's (1946) well-known taxonomy, measurement scales are classified as nominal, ordinal, interval, or ratio-level

scales. Neither the nominal nor the ratio levels of measurement accurately characterize (model-fitting) latent trait scores; this point is not controversial. However, there is much disagreement and confusion over whether latent trait scores comprise interval scales or ordinal scales. It is usually acknowledged that NIRT methods produce only ordinal-level scales (Sijtsma & Molenaar, 2002, pp. 15–16). However, many (e.g., Sijtsma & Molenaar, 2002; Yen, 1986) contend that parametric item response models produce equal-interval scales. In fact, Sijtsma and Molenaar (2002) consider the distinction between ordinal and interval scales to be a major difference between NIRT and PIRT: "[w]e definitely do *not* [emphasis in original] want to argue for the overall replacement of parametric by nonparametric IRT models. Parametric IRT models lead to point estimates of θ and to interval scales for measuring respondents" (p. 16). Part of the confusion over the status of θ stems from the idea of admissible transformations. According to Stevens (1946), ordinal-level scales are invariant to order-preserving (i.e., monotonic) transformations, whereas interval-level scales are invariant only up to linear transformations. From this fact, it is argued (Yen, 1986, p. 309) that θ is an equal-interval scale, because predicted item response probabilities are preserved only under linear transformations.

Although only linear transformations of θ are permissible when working with PIRT models (i.e., without altering the parametric IRF family), it is important to remember how the scaling of θ is determined. Namely, the scaling of θ is the result of model identification restrictions. For PIRT models, the scale is identified by the choice of IRF functional form. Once a particular IRF shape is chosen, θ is simply the scale on which all items have that mathematical form (Lord, 1975). As

such, any monotonic transformation of θ will result in another scale that defines an equally admissible IRT model, but with a different functional form. In the words of Yen (1986, p. 309), "convention plays a role in the definition of the IRT scale. For IRT, the convention is in the assumption of a logistic (or normal ogive) item characteristic function; the assumption of a different function will produce a different scale".

Perhaps the most defensible way to construct equal-interval scales lies in the theory of representational measurement. Several authors have suggested that the Rasch model (or 1PL) is intimately related to the representational theory of additive conjoint measurement (Perline et al., 1979). Specifically, it has been claimed that the Rasch model is a probabilistic formulation of the theory of additive conjoint measurement. This is because latent abilities and item difficulties are related only by a separable, additive function. Specifically,

$$P_i(\theta) = H(\theta + b_i), \tag{1.12}$$

where the function H—the filter function in FMP terminology—is the logistic cumulative distribution function. However, demonstrating the fit of the Rasch model to real data is not the same as demonstrating that the data satisfy the axioms of additive conjoint measurement (Karabatsos, 2001). Alternative methods to test these axioms in real data have been proposed (e.g., Domingue, 2014), and empirical evidence that the axioms hold is needed to justify claims about representational measurement. All things considered, we are compelled to agree with Mislevy (1987, p. 248) who concluded "[t]hat a particular IRT model fits a

dataset, therefore, is not sufficient grounds to claim scale properties stronger than ordinal."

The above arguments imply that the θ scale, obtained by specifying and fitting an item response model, is not interval-level measurement. However, one may still hesitate to relegate the θ scale to ordinal-level measurement. For instance, ordinal measurements are typically discrete variables such as ranks, whereas θ is widely considered to be a continuous variable (e.g., Stout, 2007). From this observation, it could be argued that θ does not cleanly fit into the familiar nominal-ordinal-interval-ratio categories. Many authors have noted that psychological measures do not cleanly fit into these four categories. For instance, Baker, Hardyck, and Petrinovich (1966, p. 294) claim that "it can be safely asserted that most measurements in psychology yield scales which are somewhere between ordinal and interval scales", and Mosteller and Tukey (1977, Chapter 5) suggest that Stevens' categories do not adequately characterize variables such as counts or grades. From the author's perspective, it is more important to understand the properties of a scale than to agree on classification. In the context of this paper, it is most important to understand that θ is continuous and that nonlinear transformations of θ need not alter model predictions. Nevertheless, throughout this paper I refer to θ as an ordinal-level variable to emphasize that monotonic transformations of the latent trait are permissible.

From the perspective of representational measurement, fitting parametric item response models does not necessarily produce interval-level measurement, even if the parametric IRF form determines the metric up to a linear transformation. It follows that there is nothing special about the logistic (or normal ogive) shape

of the response function that makes the θ scale preferable to other scales that are nonlinear monotonic transformations of θ (Lord, 1975). In other words, the parametric form of a response function is arbitrary; there is no psychological reason why a normal ogive (or logistic) IRF should closely correspond to the psychological response process. However, one reason why parametric models such as the 2PL are popular is that they are, mathematically, simpler to work with than other functional forms. Despite this fact, it is not clear that the θ scale resulting from the 2PL, for example, is the most useful scaling. For example, it may be desirable to choose a scaling such that θ is linearly related to external variables of interest (Yen, 1986). However, a particular scaling may be linearly related to one variable of interest yet nonlinearly related to another variable of interest. For example, Yen (1986, p. 314) suggested that an academic achievement measure may be linearly related to college GPA yet nonlinearly related to per-pupil expenditures. This is one reason why the θ scale resulting from a given parametric form could be deemed unsatisfactory. Another reason why θ may be undesirable was suggested by Lord (1975), who noticed that scales fit with the 3PL tended to have correlated difficulty and discrimination parameters. He argued that a transformed scale with uncorrelated item locations and slopes is more psychologically plausible. Absent external information about the response process, the most useful scaling of θ is debatable. Unfortunately, researchers do not routinely consider transformations of the θ metric, and in PIRT, θ is usually (explicitly or implicitly) interpreted as an equal-interval measure of the construct of interest. Moreover, no methods have been developed to systematically investigate transformations of the θ metric (and their corresponding IRFs). In a

later section, a broad framework based on the FMP model is developed that allows for such investigations of alternate latent trait metrics.

Chapter 2

Filtered Monotonic Polynomials

2.1 Model Form and Model History

Filtered monotonic polynomial IRT (FMP) is a recent addition to the family of NIRT models. The FMP model was first proposed for use in IRT by Liang (2007). However, the theory of filtered monotonic polynomials can be traced back to Elphinstone (1983, 1985). Elphinstone proposed filtered monotonic polynomials as part of a method to nonparametrically estimate cumulative distribution functions from sampled data. Cumulative distribution functions are similar to item response functions in that they must be monotonically increasing and are bounded between zero and one. Noting these similarities, Liang (2007) suggested that FMP could be a useful extension of NIRT methodology.

In its most general form, the FMP IRF is specified using the composite function,

$$P_i(\theta) = H[m_i(\theta)], \tag{2.1}$$

where H is a monotonic filter function bounded between zero and one, and for item i, $m_i(\theta)$ is an unbounded and monotonically increasing polynomial function of the latent trait θ. Specifically, let

$$m_i(\theta) = b_{0i} + b_{1i}\theta + b_{2i}\theta^2 + \cdots + b_{2k_i+1,i}\theta^{2k_i+1}, \tag{2.2}$$

where $2k_i + 1$ equals the order of the polynomial for item i, k_i is a nonnegative integer, and $\boldsymbol{b}_i = (b_{0i}, b_{1i}, \ldots, b_{2k_i+1,i})'$ are item parameters that define the location and shape of the IRF. For convenience, k_i will hereafter be called the *item complexity* parameter[1] such that an item with complexity k_i implies a polynomial with order $2k_i + 1$. Notice that k_i is allowed to vary across items; therefore, items need not be modeled with a more complex IRF than is necessary. Notice also that $m_i(\theta)$ will always be of an odd degree in the FMP model. As will be demonstrated in a later subsection, the order of $m_i(\theta)$ must be odd as a necessary (but not sufficient) condition to ensure monotonicity.

For sufficiently large k_i, any continuous monotone function can be approximated to arbitrary precision by the polynomial function $m_i(\theta)$. Further, for

[1]Throughout this paper, the phrase *item complexity* will be used to refer to the value of k_i, where higher values of k_i indicate higher-order polynomials. This use of the term is not to be confused with the multifaceted nature of item content or traits (as used, for example, by Yen, 1986) or with the acceleration parameter in Samejima's (2000) logical positive exponent model.

sufficiently large k_i, any cumulative distribution function can be approximated to arbitrary precision by the composite function $H[m_i(\theta)]$ (Elphinstone, 1983). This latter fact is true regardless of the choice of filter function H, although the k_i value needed for a given degree of precision will vary across filters. Elphinstone (1983) explored exponential, gamma, and normal distribution functions as potential filter functions. When applying filtered monotonic polynomials to IRT, it is convenient to specify H as the cumulative distribution of the logistic density function (i.e., the inverse logit function). With this choice, the probability of responding in the keyed direction to item i equals

$$\begin{aligned} P_i(\theta) &= H[m_i(\theta)] && (2.3) \\ &= \{1 + \exp[-m_i(\theta)]\}^{-1}. && (2.4) \end{aligned}$$

When $k_i = 0$, the FMP IRF equals

$$P_i(\theta) = \{1 + \exp[-(b_{0i} + b_{1i}\theta)]\}^{-1}, \quad (2.5)$$

and is equivalent to the slope-threshold parameterization of the 2PL. This fact emphasizes a major advantage of FMP IRT over other quasi-parametric NIRT methods. Namely, the FMP model can be considered an extension of, rather than an alternative to, parametric item response models. That is, if the 2PL is an appropriate shape for modeling some set of item responses, then item parameters estimated under FMP IRT correspond directly to the item parameters estimated

from the 2PL. Because k_i can vary across items, this feature of the FMP model allows the user to fit the 2PL to items when appropriate but to fit more complex IRFs to items that fit the 2PL poorly.

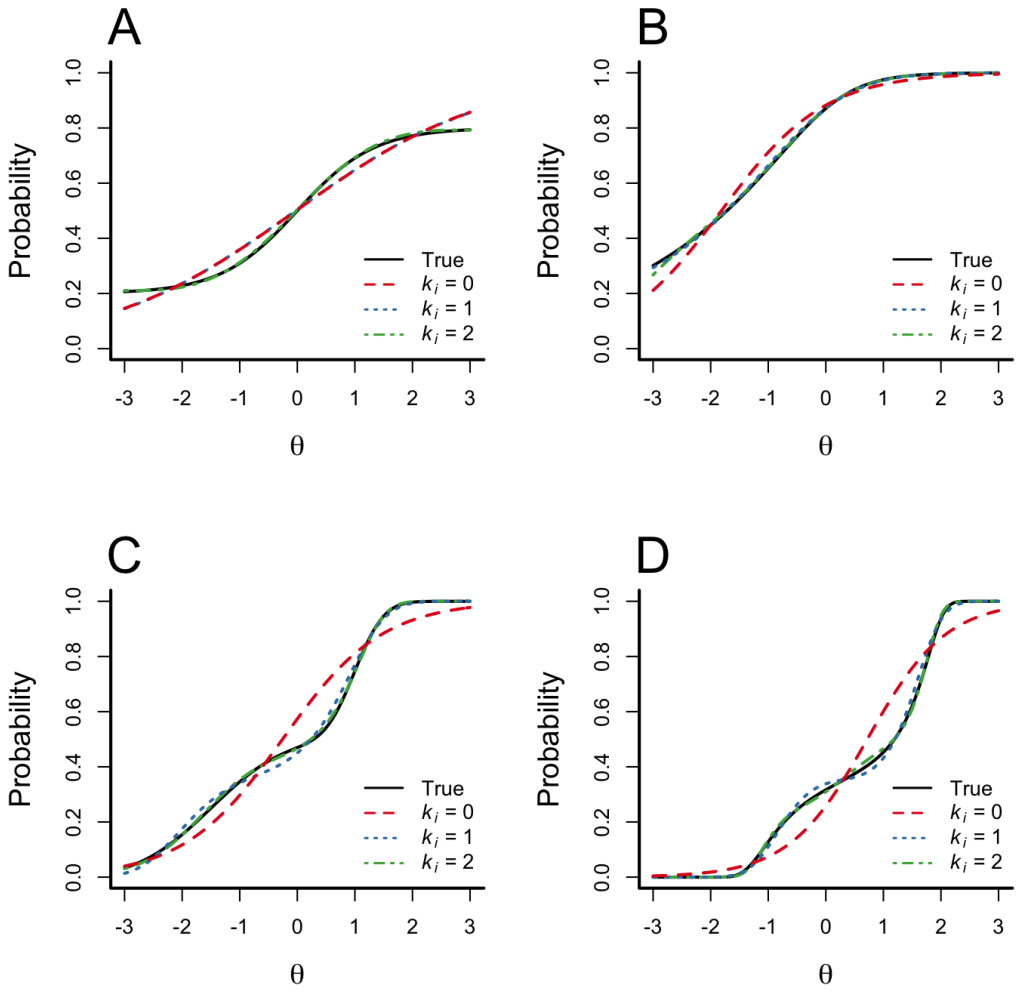

Figure 2.1. Example item response functions recovered with the FMP model. In Panel A, the true curve follows the 4PL; in Panel B, the true curve follows the LPE; in Panel C, the true curve follows a mixture normal distribution; in Panel D, the true curve follows the FMP model with $k_i = 8$.

The ability of FMP curves to approximate nonstandard IRFs is illustrated in Figure 2.1. This figure contains four panels. In each panel, a data-generating curve is displayed along with FMP estimates of that curve at $k_i = \{0, 1, 2\}$. In each panel, data were generated according to a nonstandard IRF (as described below) using 50,000 θ values drawn from a unif$(-3, 3)$ distribution. Curves were estimated using the fixed-effects estimation method described later and by conditioning on the true θ values.

In this figure, Panel A displays an IRF generated from the 4PL as defined in Equation 1.9, with $a_i = 1.5$, $b_i = 0$, $c_i = .2$, and $d_i = .8$. Panel B displays an IRF generated from Samejima's logical positive exponent (LPE) model, as shown in Equation 1.10, with $a_i = 2$, $b_i = 0$, and $\xi_i = .2$. Panel C displays an IRF generated from a mixture of normal distributions such that

$$P_i(\theta) = \pi_i \Phi(\theta|\mu_{1i}, \sigma_{1i}) + (1 - \pi_i)\Phi(\theta|\mu_{2i}, \sigma_{2i}) \tag{2.6}$$

where $\Phi(\theta|\mu, \sigma)$ indicates the normal cumulative distribution function with mean μ and standard deviation σ. In this example, $\pi_i = 0.5$, $\mu_{1i} = -1.5$, $\sigma_{1i} = 1.0$, $\mu_{2i} = 1.0$, and $\sigma_{2i} = 0.4$. These values equal the average parameters for the data-generating curves of Simulation 2 by Liang and Browne (2015). Finally, Panel D displays an IRF generated from an FMP model with $k_i = 8$ and the $\boldsymbol{b}_i$ coefficients reported in Table 2.1.

Table 2.1

True FMP item parameters for Figure 1, Panel D.

Parameter	Value	Parameter	Value
b_{0i}	-.77	b_{9i}	2×10^{-3}
b_{1i}	.55	b_{10i}	-5×10^{-4}
b_{2i}	-.17	b_{11i}	2×10^{-4}
b_{3i}	.22	b_{12i}	-4×10^{-5}
b_{4i}	-.08	b_{13i}	1×10^{-5}
b_{5i}	.07	b_{14i}	-2×10^{-6}
b_{6i}	-.02	b_{15i}	5×10^{-7}
b_{7i}	.01	b_{16i}	-3×10^{-8}
b_{8i}	-4×10^{-3}	b_{17i}	9×10^{-9}

In each panel of Figure 2.1, notice that increasing model complexity improved the fit of the FMP curve to the true curve. One way to characterize the accuracy of fitted IRFs is with the root integrated mean squared error (RIMSE; Ramsay, 1991). For item i, the RIMSE_i equals

$$\text{RIMSE}_i = \sqrt{\int [\hat{P}_i(\theta) - P_i(\theta)]^2 g(\theta) d\theta}, \tag{2.7}$$

where $P_i(\theta)$ denotes a true IRF, $\hat{P}_i(\theta)$ denotes an estimated IRF, and $g(\theta)$ indicates a target latent trait distribution. Unless otherwise indicated, $g(\theta)$ will

be a standard normal distribution. Values of the RIMSE_i statistic can be interpreted on the probability metric. For each of the four data-generating curves, higher k_i values were associated with lower RIMSE_i values. For Panel A, the RIMSE_i values for $k_i = \{0, 1, 2\}$ equal $\{.034, .034, .004\}$. The RIMSE_i values, along with a visual inspection of the fitted curves, suggest that $k_i = 2$ is needed to trace the original curve with high fidelity. Note that the FMP model, as presented above, always has asymptotes at zero and one. Thus, the 3PL and 4PL are not special cases of the FMP model. However, as Panel A of Figure 2.1 suggests, FMP models with $k_i = 2$ may fit 4PL curves well in the θ regions that contain the vast majority of examinees (assuming an examinee trait distribution with few extreme scores, such as a normal distribution). In Panel B of the figure, a curve with $k_i = 1$ approximates the true LPE curve well. For this item, $\text{RIMSE}_i = \{.032, .006, .004\}$ for $k_i = \{0, 1, 2\}$. Notice in both Panel A and Panel B, small increases in item complexity offer substantial flexibility in tracing curves that predict response probabilities greater than zero for extremely low-ability examinees or less than one for extremely high-ability examinees. For the curves in Panel C, $\text{RIMSE}_i = \{.091, .025, .007\}$ for $k_i = \{0, 1, 2\}$, and for the curves in Panel D, $\text{RIMSE}_i = \{.085, .018, .010\}$ for $k_i = \{0, 1, 2\}$. For both Panel C and Panel D, there is a large improvement in fit from $k_i = 0$ to $k_i = 1$, but smaller and possibly trivial improvement in fit from $k_i = 1$ to $k_i = 2$. Note that, because a large number of evenly spaced θ values were used to estimate the FMP curves, sampling error is minimized. Thus, these RIMSE_i values reflect the optimal performance (i.e., the lowest RIMSE_i value) of the FMP approximation for a fixed k_i value. Note that the RIMSE_i values for a fixed k_i value differs across panels (i.e., across

data-generating curves). Also notice that in all cases, FMP curves with $k_i = 2$ approximate the true curves with $\text{RIMSE}_i \leq .01$, that is, the root mean square predicted probabilities are within .01 of the true curve. This result suggests that item complexities need not be large to approximate many monotonic curves to a high degree of precision.

2.2 Model Estimation

2.2.1 Ensuring monotonicity

Not all IRFs that satisfy Equation 2.4 are monotonically increasing functions of θ. An FMP item response function $P_i(\theta)$ is monotonic (increasing) if and only if the polynomial function $m_i(\theta)$ is also monotonic. Liang (2007) and Liang and Browne (2015) suggest ensuring monotonicity by means of a parameter transformation. The parameter transformation described below closely follows the development presented by Liang and Browne (2015).

For $m_i(\theta)$ to be a monotonic function, a necessary and sufficient condition is that its first derivative,

$$\frac{\partial m_i(\theta)}{\partial \theta} = p_i(\theta) = a_{0i} + a_{1i}\theta + \cdots + a_{2k_i,i}\theta^{2k_i}, \tag{2.8}$$

is nonnegative at all θ. Here, let

$$b_{0i} = \xi_i \tag{2.9}$$

be the constant of integration and

$$b_{si} = \frac{a_{s-1,i}}{s} \tag{2.10}$$

for $s = 1, 2, \ldots, 2k_i + 1$.

Notice that $p_i(\theta)$ is a polynomial function of degree $2k_i$. A nonnegative polynomial of an even degree can be re-expressed as the product of k_i quadratic functions (Elphinstone, 1983, p. 173),

$$\begin{aligned} p_i(\theta) &= \lambda_i \prod_{s=1}^{k_i} \left[1 - 2\alpha_{si}\theta + \left(\alpha_{si}^2 + \beta_{si}\right)\theta^2\right] \text{ if } k_i \geq 1 && (2.11) \\ &= \lambda_i \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } k_i = 0, && (2.12) \end{aligned}$$

where $\lambda_i \geq 0$ and $\beta_{si} \geq 0$, $s = 1, \ldots k_i$. Following the suggestion by Falk and Cai (2015, 2016), let

$$\omega_i = \ln(\lambda_i) \tag{2.13}$$

and

$$\tau_{si} = \ln(\beta_{si}) \tag{2.14}$$

such that ω_i and τ_{si} are defined on the entire real line. Using these transformed parameters, the vector $\boldsymbol{a}_i = (a_{0i}a_{1i}, a_{2i}, \ldots, a_{2k_i,i})'$ can be expressed using matrix

notation. Specifically,

$$\boldsymbol{a}_i = \boldsymbol{T}_i^{(k_i)} \boldsymbol{T}_i^{(k_i-1)} \cdots \boldsymbol{T}_i^{(2)} \boldsymbol{T}_i^{(1)} \exp(\omega_i) \tag{2.15}$$

where $\boldsymbol{T}^{(s)}$ is a $(2s+1) \times (2s-1)$ matrix with elements

$$\begin{aligned} \left[\boldsymbol{T}^{(s)}\right]_{rc} &= 1 \text{ if } r = c, \\ &= -2\alpha_{si} \text{ if } r = c+1, \\ &= \alpha_{si}^2 + \exp(\tau_{si}) \text{ if } r = c+2, \text{ and} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.16}$$

For example,

$$\boldsymbol{T}_i^{(1)} = \begin{bmatrix} 1 \\ -2\alpha_{1i} \\ \alpha_{1i}^2 + \exp\left(\tau_{1i}\right) \end{bmatrix}, \tag{2.17}$$

$$\boldsymbol{T}_i^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ -2\alpha_{2i} & 1 & 0 \\ \alpha_{2i}^2 + \exp\left(\tau_{2i}\right) & -2\alpha_{2i} & 1 \\ 0 & \alpha_{2i}^2 + \exp\left(\tau_{2i}\right) & -2\alpha_{2i} \\ 0 & 0 & \alpha_{2i}^2 + \exp\left(\tau_{2i}\right) \end{bmatrix}, \text{ and} \tag{2.18}$$

$$\boldsymbol{T}_i^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2\alpha_{3i} & 1 & 0 & 0 & 0 \\ \alpha_{3i}^2 + \exp(\tau_{3i}) & -2\alpha_{3i} & 1 & 0 & 0 \\ 0 & \alpha_{3i}^2 + \exp(\tau_{3i}) & -2\alpha_{3i} & 1 & 0 \\ 0 & 0 & \alpha_{3i}^2 + \exp(\tau_{3i}) & -2\alpha_{3i} & 1 \\ 0 & 0 & 0 & \alpha_{3i}^2 + \exp(\tau_{3i}) & -2\alpha_{3i} \\ 0 & 0 & 0 & 0 & \alpha_{3i}^2 + \exp(\tau_{3i}) \end{bmatrix}. \tag{2.19}$$

Under this parameterization, the coefficients

$$\boldsymbol{b}_i = (b_{0i}, b_{1i}, b_{2i}, \ldots, b_{2k_i+1,i})' \tag{2.20}$$

can be estimated using the transformed parameter vector

$$\boldsymbol{\gamma}_i = (\xi_i, \omega_i, \alpha_{1i}, \tau_{1i}, \alpha_{2i}, \tau_{2i}, \cdots, \alpha_{2k_i,i}, \tau_{2k_i,i})'. \tag{2.21}$$

Notice that both $\boldsymbol{b}_i$ and $\boldsymbol{\gamma}_i$ contain $2k_i+2$ coefficients, so there is no increase in the number of model parameters when using the transformed parameters. Further, because λ_i and β_{si} are further transformed to ω_i and τ_{si}, there is no need to impose box constraints on the estimated parameters as described by Liang and Browne (2015).

2.2.2 Fixed-effects estimation with theta surrogates

The earliest attempts at FMP model estimation (Liang, 2007; Liang & Browne, 2015) used a fixed-effects approach conditional on surrogate θ values. Specifically,

these authors treated normalized first principal component scores as fixed values when estimating FMP IRFs. The fixed-effects approach described by these authors is similar to JML estimation, although it does not iteratively estimate the latent trait while estimating curves, and is comparable to other methods for estimating NIRT curves using fixed-effects (e.g., Ramsay, 1991).

For examinee n on item i, the log likelihood of a keyed response $y_{in} = 1$ equals

$$F_n(\boldsymbol{\gamma}_i) = \ln L(\boldsymbol{\gamma}_i | y_{in}, \theta_n) = y_{in} \ln P_{in} + (1 - y_{in}) \ln(1 - P_{in}), \tag{2.22}$$

where $P_{in} = P(\theta_n | \boldsymbol{\gamma}_i)$. The maximum likelihood estimate of $\boldsymbol{\gamma}_i$ is found by minimizing the negative log likelihood for item i,

$$F(\boldsymbol{\gamma}_i) = -\ln L(\boldsymbol{\gamma}_i | \boldsymbol{Y}_i, \boldsymbol{\theta}) = -\sum_{n=1}^{N} F_n(\boldsymbol{\gamma}_i), \tag{2.23}$$

where N denotes sample size. Using fixed-effects estimation, maximum likelihood estimates may be obtained separately for each item. However, if there are across-item constraints (as will be needed in a later section), we may wish to simultaneously estimate all item parameters for a test. The maximum likelihood estimate of

$$\boldsymbol{\zeta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_I)' \tag{2.24}$$

is found by minimizing the negative full-data log likelihood function

$$F(\boldsymbol{\zeta}) = -\ln L(\boldsymbol{\zeta} | \boldsymbol{Y}, \boldsymbol{\theta}) = -\sum_{i=1}^{I} \sum_{n=1}^{N} F_n(\boldsymbol{\gamma}_i), \tag{2.25}$$

where I denotes test length. When no across-item constraints are needed, it is simpler to minimize the likelihood separately for each item.

Following the development in Liang and Browne (2015, online appendix B), I next find the gradient $\boldsymbol{g}_i$ of $F(\boldsymbol{\gamma}_i)$ for item i. If there are no across-item constraints, the likelihood given by Equation 2.25 has gradient $\boldsymbol{g} = (\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_I)'$. The elements of $\boldsymbol{g}_i$ are found by an application of the chain rule. For element l of $\boldsymbol{\gamma}_i$,

$$\begin{aligned} g_{li} = \frac{\partial F_i}{\partial \gamma_{li}} &= \sum_{n=1}^{N} \frac{\partial F_{in}}{\partial P_{in}} \frac{\partial P_{in}}{\partial m_{in}} \frac{\partial m_{in}}{\partial \gamma_{li}} \\ &= \sum_{n=1}^{N} (y_{in} - P_{in}) \frac{\partial m_{in}}{\partial \gamma_{li}}, \end{aligned} \tag{2.26}$$

where F_{in} and m_{in} are shorthand notation for $F_n(\gamma_i)$ and $m_i(\theta_n)$. The equality given in Equation 2.26 holds because

$$\frac{\partial F_{in}}{\partial P_{in}} = -\frac{y_{in} - P_{in}}{P_{in}(1 - P_{in})} \tag{2.27}$$

and

$$\frac{\partial P_{in}}{\partial m_{in}} = P_{in}(1 - P_{in}). \tag{2.28}$$

Now,

$$\frac{\partial m_{in}}{\partial \gamma_l} = \boldsymbol{\nu}'_{in} \frac{\partial \boldsymbol{a}_i}{\partial \gamma_{li}} \tag{2.29}$$

where a typical element s of vector $\boldsymbol{\nu}_{in}$ equals

$$\nu_{s,in} = \frac{\partial m_{in}}{\partial a_{si}} = \frac{1}{s+1}\theta_{in}^{s+1},\ s = 0, 1, \ldots, 2k, \tag{2.30}$$

and

$$\begin{aligned}
\frac{\partial m_{in}}{\partial \xi_i} &= 1, \\
\frac{\partial m_{in}}{\partial \omega_i} &= \boldsymbol{\nu}_{in}' \boldsymbol{T}^{(k)} \boldsymbol{T}^{(k-1)} \cdots \boldsymbol{T}^{(2)} \boldsymbol{T}^{(1)} \exp(\omega_i), \\
\frac{\partial m_{in}}{\partial \alpha_{si}} &= \boldsymbol{\nu}_{in}' \boldsymbol{T}^{(k)} \boldsymbol{T}^{(k-1)} \cdots \frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}} \cdots \boldsymbol{T}^{(2)} \boldsymbol{T}^{(1)} \exp(\omega_i), \text{ and} \\
\frac{\partial m_{in}}{\partial \tau_{si}} &= \boldsymbol{\nu}_{in}' \boldsymbol{T}^{(k)} \boldsymbol{T}^{(k-1)} \cdots \frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}} \cdots \boldsymbol{T}^{(2)} \boldsymbol{T}^{(1)} \exp(\omega_i).
\end{aligned} \tag{2.31}$$

The derivative matrices $\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}$ and $\frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}}$ are each of dimension $(2s+1) \times (2s-1)$. All elements of these derivative matrices are zero with the following exceptions:

$$\left[\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}\right]_{rc} = -2 \quad \text{if} \quad r = c+1, \tag{2.32}$$

$$\left[\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}\right]_{rc} = 2\alpha_{si} \quad \text{if} \quad r = c+2, \text{ and} \tag{2.33}$$

$$\left[\frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}}\right]_{rc} = \exp(\tau_{si}) \quad \text{if} \quad r = c+2. \tag{2.34}$$

The Hessian matrix $\boldsymbol{H}$ of $F(\boldsymbol{\zeta})$ is block diagonal with item-specific Hessians

$\boldsymbol{H}_i$ of $F(\boldsymbol{\gamma}_i)$ on the diagonals. A typical element h_{lri} of $\boldsymbol{H}_i$ has the form

$$h_{lri} = \frac{\partial^2 F_i}{\partial \gamma_{li}\gamma_{ri}} = \sum_{n=1}^{N}\left[P_{in}(1-P_{in})\frac{\partial m_{in}}{\partial \gamma_{li}}\frac{\partial m_{in}}{\partial \gamma_{ri}} - (y_{si}-P_{si})\frac{\partial^2 m_{in}}{\partial \gamma_{li}\partial\gamma_{ri}}\right], \qquad (2.35)$$

where the quantities $\frac{\partial m_{in}}{\partial \gamma_{li}}$ and $\frac{\partial m_{in}}{\partial \gamma_{ri}}$ are given in Equation 2.31. The quantities $\frac{\partial^2 m_{in}}{\partial \gamma_{li}\partial\gamma_{ri}}$ are found by taking partial derivatives of the quantities in Equation 2.31. Following this scheme,

$$\begin{aligned}
\frac{\partial^2 m_{in}}{\partial \xi_i \partial \gamma_{ri}} = \frac{\partial^2 m_{in}}{\partial \gamma_{li}\partial \xi_i} &= 0 \text{ for all } l \text{ and } r, && (2.36)\\
\frac{\partial^2 m_{in}}{\partial \omega_i^2} &= \frac{\partial m_{in}}{\partial \omega_i}, && (2.37)\\
\frac{\partial^2 m_{in}}{\partial \omega_i \partial \alpha_{si}} = \frac{\partial^2 m_{in}}{\partial \alpha_{si}\partial \omega_i} &= \frac{\partial m_{in}}{\partial \alpha_{si}}, && (2.38)\\
\frac{\partial^2 m_{in}}{\partial \omega_i \partial \tau_{si}} = \frac{\partial^2 m_{in}}{\partial \tau_{si}\partial \omega_i} &= \frac{\partial m_{in}}{\partial \tau_{si}}, && (2.39)\\
\frac{\partial^2 m_{in}}{\partial \alpha_{si}^2} &= \boldsymbol{\nu}'_{in}\boldsymbol{T}^{(k)}\boldsymbol{T}^{(k-1)}\cdots\frac{\partial^2 \boldsymbol{T}^{(s)}}{\partial \alpha_{si}^2}\cdots\boldsymbol{T}^{(1)}\exp(\omega_i), && (2.40)\\
\frac{\partial^2 m_{in}}{\partial \alpha_{si}\tau_{si}} &= 0, && (2.41)\\
\frac{\partial^2 m_{in}}{\partial \alpha_{si}\alpha_{ti}} = \frac{\partial^2 m_{in}}{\partial \alpha_{ti}\alpha_{si}} &= \boldsymbol{\nu}'_{in}\boldsymbol{T}^{(k)}\boldsymbol{T}^{(k-1)}\cdots\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}\cdots\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}\cdots\boldsymbol{T}^{(1)}\exp(\omega_i),\ s\neq t, && (2.42)\\
\frac{\partial^2 m_{in}}{\partial \alpha_{si}\tau_{ti}} = \frac{\partial^2 m_{in}}{\partial \tau_{ti}\alpha_{si}} &= \boldsymbol{\nu}'_{in}\boldsymbol{T}^{(k)}\boldsymbol{T}^{(k-1)}\cdots\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}\cdots\frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}}\cdots\boldsymbol{T}^{(1)}\exp(\omega_i),\ s\neq t, && (2.43)\\
\frac{\partial^2 m_{in}}{\partial \tau_{si}^2} &= \boldsymbol{\nu}'_{in}\boldsymbol{T}^{(k)}\boldsymbol{T}^{(k-1)}\cdots\frac{\partial^2 \boldsymbol{T}^{(s)}}{\partial \tau_{si}^2}\cdots\boldsymbol{T}^{(1)}\exp(\omega_i),\ \text{ and} && (2.44)\\
\frac{\partial^2 m_{in}}{\partial \tau_{si}\tau_{ti}} = \frac{\partial^2 m_{in}}{\partial \tau_{ti}\tau_{si}} &= \boldsymbol{\nu}'_{in}\boldsymbol{T}^{(k)}\boldsymbol{T}^{(k-1)}\cdots\frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}}\cdots\frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}}\cdots\boldsymbol{T}^{(1)}\exp(\omega_i),\ s\neq t. && (2.45)
\end{aligned}$$

The matrix $\frac{\partial^2 \boldsymbol{T}^{(s)}}{\partial \alpha_{si}^2}$ is of dimension $(2s+1)\times(2s-1)$ and all elements are zero except $\left[\frac{\partial \boldsymbol{T}^{(s)}}{\partial \alpha_{si}}\right]_{rc} = 2\alpha_{si}$ if $r = c+2$, and $\frac{\partial^2 \boldsymbol{T}^{(s)}}{\partial \tau_{si}^2} = \frac{\partial \boldsymbol{T}^{(s)}}{\partial \tau_{si}}$ for all s. Item parameter

estimates may be obtained using a Newton-Raphson algorithm. Specifically, at iteration $t + 1$,

$$\hat{\boldsymbol{\zeta}}^{(t+1)} = \hat{\boldsymbol{\zeta}}^{(t)} - \boldsymbol{H}^{(t)-1}\boldsymbol{g}^{(t)} \tag{2.46}$$

where $\boldsymbol{H}^{(t)}$ and $\boldsymbol{g}^{(t)}$ are the Hessian and gradient functions shown in Equations 2.35 and 2.26 evaluated at $\hat{\boldsymbol{\zeta}}^{(t)}$. Iterations continue until the change in likelihood across two successive iterations becomes sufficiently small.

2.2.3 Random-effects estimation with the EM algorithm

The random-effects specification of FMP models may be achieved using marginal maximum likelihood (MML; Bock & Aitkin, 1981; Bock & Lieberman, 1970) via an application of the EM algorithm (Dempster et al., 1977). Random-effects estimation via MML has already been applied to two extensions of the FMP model: an FMP model with a guessing parameter (Falk & Cai, 2015) and a polytomous-item multiple-group FMP model (Falk & Cai, 2016). Additionally, these authors describe the use of Bayesian prior distributions on the item parameters for both extended FMP models.

In the MML approach to item parameter estimation, the marginal likelihood of the data is obtained by integrating over the prior distribution of latent trait. Under the assumption of local independence, the complete-data likelihood equals

$$L(\boldsymbol{\zeta}|\boldsymbol{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{n=1}^{N} P(y_{in}|\boldsymbol{\gamma}_i, \theta_n)^{y_{in}} \left[1 - P(y_{in}|\boldsymbol{\gamma}_i, \theta_n)\right]^{1-y_{in}}, \tag{2.47}$$

and the contribution of person n to the complete data likelihood equals

$$L(\boldsymbol{\zeta}|\boldsymbol{y}_n, \theta_n) = \prod_{i=1}^{I} P(y_{in}|\boldsymbol{\gamma}_i, \theta_n)^{y_{in}} \left[1 - P(y_{in}|\boldsymbol{\gamma}_i, \theta_n)\right]^{1-y_{in}}. \tag{2.48}$$

In random-effects estimation, individuals are treated as a random sample from a population. Thus, instead of maximizing the full-data likelihood as was done in random-effects estimation, we maximize the marginal likelihood, which is found by integrating out the latent trait. Let $g(\theta)$ denote the prior distribution of the latent trait. Then, the marginal likelihood for person n equals

$$L(\boldsymbol{\zeta}|\boldsymbol{y}_n) = \int L(\boldsymbol{\zeta}|\boldsymbol{y}_n, \theta_n) g(\theta_n) d\theta_n, \tag{2.49}$$

and the complete-data marginal likelihood equals

$$L(\boldsymbol{\zeta}|\boldsymbol{Y}) = \prod_{n=1}^{N} L(\boldsymbol{\zeta}|\boldsymbol{y}_n). \tag{2.50}$$

To perform the integration in Equation 2.49, Q quadrature points can be used to approximate $g(\theta)$. The quadrature is characterized by a set of quadrature nodes $\boldsymbol{X} = (X_1, X_2, \ldots, X_Q)'$ and their associated weights $A(X_1), A(X_2), \ldots A(X_Q)$. Now, let

$$L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q) = \prod_{i=1}^{I} P_i(X_q)^{y_{in}} \left[1 - P_i(X_q)\right]^{1-y_{in}} \tag{2.51}$$

represent the quadrature expression of Equation 2.48 at point X_q. Using this quadrature scheme, we can re-express the marginal likelihood in Equation 2.49 by replacing the integral with a weighted sum over the quadrature nodes. An

individual's contribution to the marginal likelihood equals

$$L(\boldsymbol{\zeta}|\boldsymbol{y}_n, \boldsymbol{X}) = \sum_{q=1}^{Q} L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q), \tag{2.52}$$

and the full-data marginal likelihood in quadrature form equals

$$L(\boldsymbol{\zeta}|\boldsymbol{Y}, \boldsymbol{X}) = \prod_{n=1}^{N} L(\boldsymbol{\zeta}|\boldsymbol{y}_n, \boldsymbol{X}). \tag{2.53}$$

Next recall that $g(\theta)$ is the prior distribution of ability. After observing the data, the posterior distribution of the latent trait for individual n equals

$$P(\theta_n|\boldsymbol{y}_n, \boldsymbol{\zeta}) = \frac{L(\boldsymbol{\zeta}|\boldsymbol{y}_n, \theta_n)g(\theta_n)}{\int L(\boldsymbol{\zeta}|\boldsymbol{y}_n, \theta_n)g(\theta_n)d\theta}, \tag{2.54}$$

where $L(\boldsymbol{y}_n|\theta_n, \boldsymbol{\zeta})$ is defined as in Equation 2.48. In quadrature form, the expression in Equation 2.54 becomes

$$P(X_q|\boldsymbol{y}_n, \boldsymbol{\zeta}) = \frac{L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q)A(X_q)}{\sum_{q=1}^{Q} L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q)A(X_q)}. \tag{2.55}$$

The expected number of persons at quadrature point q, denoted $\bar{n}_q$, equals

$$\bar{n}_q = \sum_{n=1}^{N} \left[\frac{L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q)A(X_q)}{\sum_{q=1}^{Q} L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q)A(X_q)} \right], \tag{2.56}$$

and the number of correct responses to item i at quadrature point q, denoted $\bar{r}_{iq}$, equals

$$\bar{r}_{iq} = \sum_{n=1}^{N} \left[\frac{y_{in}L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q)A(X_q)}{\sum_{q=1}^{Q} L(\boldsymbol{\zeta}|\boldsymbol{y}_n, X_q)A(X_q)} \right]. \tag{2.57}$$

The quantities $\bar{n}_q$ and $\bar{r}_{iq}$ can then be treated as artificial data to maximize the quadrature expression of the marginal data likelihood, which equals

$$L(\boldsymbol{r}, \boldsymbol{n}|\boldsymbol{X}, \boldsymbol{\zeta}) = \prod_{i=1}^{I} \prod_{q=1}^{Q} P(X_q|\boldsymbol{\gamma}_i)^{\bar{r}_{iq}} \left[1 - P(X_q|\boldsymbol{\gamma}_i)\right]^{\bar{n}_q - \bar{r}_{iq}}, \tag{2.58}$$

where $\boldsymbol{r} = (\bar{r}_{11}, \bar{r}_{12}, \ldots, \bar{r}_{IQ})'$ and $\boldsymbol{n} = (\bar{n}_1, \ldots, \bar{n}_Q)'$. The natural log of the marginal likelihood equals

$$\ln L(\boldsymbol{r}, \boldsymbol{n}|\boldsymbol{X}, \boldsymbol{\zeta}) = \sum_{i=1}^{I} \sum_{q=1}^{Q} \left\{\bar{r}_{iq} \ln P(X_q|\boldsymbol{\gamma}_i) + (\bar{n}_q - \bar{r}_{iq}) \ln[1 - P(X_q|\boldsymbol{\gamma}_i)]\right\}. \tag{2.59}$$

The quantities described above are used in the EM algorithm (Dempster et al., 1977) to estimate item parameters. The EM algorithm iterates between two steps, the expectation (E) step and the maximization (M) step. In the E step, provisional estimates of $\boldsymbol{\zeta}$ (e.g., parameter estimates found by fixed-effects estimation) are used to find the expected number of persons at each quadrature point (using Equation 2.56) and the expected number of correct responses to each item at each quadrature point (using Equation 2.57). In the M step, $\bar{n}_q$ and $\bar{r}_{iq}$ are treated as artificial data and are used to maximize the quadrature form of the complete-data log likelihood given in Equation 2.59. The updated parameter estimates computed during the M step are then treated as fixed in the next iteration of the E step. Put another way, $\bar{n}_q$ and $\bar{r}_{iq}$ are updated in the E step, and $\hat{\boldsymbol{\gamma}}_i$, $i = 1, \ldots, I$, is updated in the M step. The E step and the M step are repeated iteratively until $\hat{\boldsymbol{\zeta}}$ stabilizes. For all MML analyses reported in this paper, the EM algorithm will

terminate when all elements of $\hat{\boldsymbol{\zeta}}$ change by no more than .0001 across consecutive iterations.

2.3 Model Selection

When estimating parameters for the FMP model using fixed-effects or random-effects methods, item complexities (k_i) must be specified in advance of model fitting. Because the optimal choices of item complexities are unknown in advance, previous applications of FMP IRT model have compared the fit of several models in order to select the final k_i values. Importantly, the most appropriate k_i value for a given item depends not only on the data-generating IRF, but on sample size. In smaller samples, attempts to fit high polynomial degrees will likely capitalize on chance and may not improve the proximity of the estimated IRF to the true IRF. Conversely, in large samples, higher polynomial degrees may lead to closer approximation of the population IRF and less biased predicted probabilities. Thus, *ceteris paribus*, it is desirable to estimate fewer parameters in smaller samples and more parameters for larger samples (see Browne, 2000; Cudeck & Henly, 1991, Liang & Browne, 2015). Put another way, the guiding philosophy of FMP is not to unearth the correct data-generating model but rather to fit a model that leads to the most stable and reliable inferences possible given the data. This philosophy is similar to the sentiment expressed by Molenaar (2001): "[m]odern technology allows detailed features of single IRFs to be inspected. However, this should not lead to over-reporting about small details. At that level, all models are somewhat incorrect. The main question should be whether a model discrepancy seriously

influences major conclusions" (p. 296).

Several criteria have been proposed for FMP model selection. Liang (2007) fit items using a fixed-effects approach and compared three model selection criteria: the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), and the likelihood ratio test. The AIC value equals

$$\text{AIC} = -2F + 2p \tag{2.60}$$

where F is the maximum log likelihood value given by either Equation 2.23 for a single item, Equation 2.25 or Equation 2.59 for a set of I items, and p is the number of estimated parameters. Alternatively,

$$\text{BIC} = -2F + p \ln N \tag{2.61}$$

applies a stronger penalty for model complexity than does the AIC when $N \geq 8$, where N represents sample size. Finally, a likelihood ratio statistic may be used to compare the fit of two nested models. Let F_1 be the log likelihood of the nested model, and let F_2 be the log likelihood of the more general model. The likelihood ratio statistic comparing these models equals

$$G^2(\text{dif}) = -2\left(F_1 - F_2\right) \tag{2.62}$$

and follows an approximate χ^2 distribution with degrees of freedom equal to the difference in number of estimated parameters across the two models.

When using fixed-effects methods, items can be fit one-at-a-time, and k_i values

can be determined on an item-by-item basis. For a single item i, Liang and Browne (2015) recommended fitting the $k_i = 0$ and $k_i = 1$ models first. If the model selection method (AIC, BIC, or G^2(dif)) selects the $k_i = 1$ model, then the k_2 model is also fit. The $k_i = 1$ and $k_i = 2$ models are then compared. This process continues until the model selection method does not select the model with the larger k_i value. This method allows k_i to differ across items and aims to prevent overfitting IRFs. Liang (2007) evaluated this model selection method by looking at the RIMSE$_i$ values produced by fitting items to several k_i values rather than evaluating whether the data-generating model was selected. This evaluation criterion reflects the FMP philosophy that it is more important to fit an approximate good-fitting model rather than finding the data-generating model *per se*. Using the AIC, in simulation studies, Liang (2007) found that the smallest RIMSE$_i$ values were selected by the BIC when the 2PL ($k_i = 0$) was the correct model. However, when the true $k_i > 0$, models with the smallest RIMSE$_i$ values were selected by the AIC. In general, she found only trivial differences in the RIMSE$_i$ values for different model selection criteria, and concluded that "the choice of criterion in practice is not very critical" (p. 73). Accordingly, Liang and Browne (2015) used only AIC for model selection.

Falk and Cai (2016) also suggested a step-wise AIC-based approach for sequential model fitting and model selection. However, because they used the random-effects approach via MML, all items were estimated simultaneously. In their approach, they began by fitting a model with $k_i = 0$ for all items. They then looped over items, setting $k_i = 1$ for exactly one item and all other items were fit with $k_i = 0$. The item that led to the greatest improvement in AIC was then fixed

at $k_i = 1$, and the remaining items were again looped over by setting $k_i = 1$ for exactly one item in addition to the item previously retained at $k_i = 1$. This process was repeated until AIC did not improve by increasing the complexity of more items. Falk and Cai (2015)[2] also explored $S - X^2$ (Orlando & Thissen, 2000, 2003) as a criterion for sequential model-fitting. Using adjusted p-values for the $S - X^2$ statistic, all items that are flagged as poorly fitting were then fit using $k_i = 1$. In a real data example ($N = 10,000$) using the $S - X^2$ criterion, they found that up to 25% of the items on a large-scale math assessment could be better modeled using $k_i > 0$ (using a guessing-added version of the FMP model). However, in a simulation study, they found fairly low power of this method to detect true non-standard IRFs in samples as large as $N = 5,000$.

2.4 Latent Trait Estimation

2.4.1 Maximum likelihood solution

After calibrating FMP item parameters using either the fixed-effects or random-effects solution, individual trait score estimates can be computed by treating the estimated item parameters as fixed values. Under the assumption of local independence, the probability of the response vector $\boldsymbol{y}_n$ for person n equals

$$P(\boldsymbol{y}_n|\theta_n) = \prod_{i=1}^{I} P_i(\theta_n)^{y_{in}} \left[1 - P_i(\theta_n)\right]^{1-y_{in}}, \tag{2.63}$$

[2]Although the Falk and Cai (2016) document is dated after the Falk and Cai (2015) document, the 2016 paper actually precedes the 2015 paper.

where $P_i(\theta_n)$ is the IRF for item i evaluated at θ_n. The log of the likelihood function equals

$$L = y_{in} \ln[P_i(\theta_n)] + (1 - y_{in}) \ln[1 - P_i(\theta_n)], \tag{2.64}$$

and the maximum likelihood ability estimate equals the θ value that maximizes Equation 2.64. One disadvantage of the maximum likelihood trait estimate is that, for nonmixed response patterns (i.e., all 0's or all 1's), the latent trait estimate is infinite. To circumvent this problem, maximum likelihood estimates may be truncated (e.g., at ± 4). Alternatively, another trait estimation method such as *expected a posteriori* (EAP) estimation (Bock & Aitkin, 1981; Bock & Mislevy, 1982) may be used.

2.4.2 Item and test information

The accuracy of latent trait estimates is gauged using the item information function. In unidimensional IRT models for binary data, the item information function (IIF) equals

$$\begin{aligned} \mathcal{I}_i(\theta) &= -E\left(\frac{\partial^2 \ln L_i}{\partial \theta^2}\right) \\ &= \frac{\left(\frac{\partial P_i}{\partial \theta}\right)^2}{P_i(1 - P_i)} \end{aligned} \tag{2.65}$$

where E is the expectation operator, and for item i, $\ln L_i$ is the quantity in Equation 2.23, P_i is shorthand for $P(\theta|\boldsymbol{b}_i)$, and $\boldsymbol{b}_i$ is a vector of item parameters.

It can be shown that

$$\frac{\partial P_i}{\partial \theta} = \frac{\partial m_i}{\partial \theta} P_i(1 - P_i), \tag{2.66}$$

which means that

$$\mathcal{I}_i(\theta) = \left(\frac{\partial m_i}{\partial \theta} \right)^2 P_i(1 - P_i) \tag{2.67}$$

is an equivalent expression for the IIF. Note that $\frac{dm_i(\theta)}{d\theta}$ has already been defined in Equations 2.8 and 2.10. With these definitions, the item information can be written,

$$\mathcal{I}_i(\theta) = \left[\sum_{s=0}^{2k_i} (s+1) b_{si} \theta^s \right]^2 P_i(1 - P_i), \tag{2.68}$$

and the test information function (TIF) equals the sum of item response functions:

$$\mathcal{I}(\theta) = \sum_{i=1}^{I} \mathcal{I}_i(\theta). \tag{2.69}$$

Finally, the inverse square root of the test information, evaluated at $\hat{\theta}_n$, is the expected standard error for the maximum likelihood estimate $\hat{\theta}_n$.

2.4.3 Expected a posteriori solution

For the fixed-effects approach conditioning on normalized θ surrogates, Liang (2007) recommended Bayesian EAP estimation (Bock & Aitkin, 1981; Bock & Mislevy, 1982). After item calibration, EAP scores are found by integrating over

an assumed distribution of ability scores, denoted $g(\theta)$. For individual n, the expected value of his *a posteriori* distribution of ability equals

$$E(\theta|\boldsymbol{y}_n, \boldsymbol{\zeta}) = \frac{\int \prod_{i=1}^{I} \theta P_i(\theta)^{y_{in}} \left[1 - P_i(\theta)\right]^{1-y_{in}} g(\theta) d\theta}{\int \prod_{i=1}^{I} P_i(\theta)^{y_{in}} \left[1 - P_i(\theta)\right]^{1-y_{in}} g(\theta) d\theta}. \tag{2.70}$$

The integrals may be solved numerically by summing over rectangular quadrature. Let X_q denote the quadrature nodes, and let $A(X_q)$ denote the quadrature weights found by taking the height of the prior density at each quadrature node. Note that, although the notation is the same for the delineated above for the EM algorithm, it is not necessary to use the same quadrature scheme in both places. Whereas choosing a large number of quadrature may lead to very slow performance of the EM algorithm, EAP trait estimation is a simpler problem and can be computed quickly even with a large number of quadrature points. Nor is it necessary to use the same prior distribution in both places. In quadrature form, the estimated latent trait then equals

$$\hat{\theta}_n = \frac{\sum_{q=1}^{Q} \prod_{i=1}^{I} X_q P_i(X_q)^{y_{in}} \left[1 - P_i(X_q)\right]^{1-y_{in}} A(X_q)}{\sum_{q=1}^{Q} \prod_{i=1}^{I} P_i(X_q)^{y_{in}} \left[1 - P_i(X_q)\right]^{1-y_{in}} A(X_q)}, \quad n = 1, \ldots, N. \tag{2.71}$$

The posterior variance of the EAP estimator can also be found using quadrature,

$$\text{var}(\hat{\theta}_n) = \frac{\sum_{q=1}^{Q} \prod_{i=1}^{I} \left(X_q - \hat{\theta}_n\right)^2 P_i(X_q)^{y_{in}} \left[1 - P_i(X_q)\right]^{1-y_{in}} A(X_q)}{\sum_{q=1}^{Q} \prod_{i=1}^{I} P_i(X_q)^{y_{in}} \left[1 - P_i(X_q)\right]^{1-y_{in}} A(X_q)}, \quad n = 1, \ldots, N, \tag{2.72}$$

where $\hat{\theta}_n$ is the estimate obtained from Equation 2.71 and all other terms are as previously defined.

Chapter 3

Simulation Study

3.1 Design

An `R` package that fits the FMP model using the methods described in the previous section has been written by the author. In this section, I describe a simulation study designed to determine the feasibility of FMP model estimation using fixed-effects and random-effects approaches. The goals of this simulation are (a) to establish the data conditions under which accurate FMP IRFs and latent trait scores can be estimated, (b) to compare the relative accuracy of the fixed-effects and random-effects methods, and (c) to establish general guidelines for sample size requirements for the FMP model.

Data were generated according to 5 sample size and 3 test length conditions. The studied sample sizes were $N \in \{200, 500, 1000, 2000, 5000\}$ subjects, all trait values were drawn from from standard normal distributions, and the studied test lengths were $I \in \{20, 40, 60\}$ items. Additionally, data sets were generated under

three k_i values: $k_i \in \{0, 1, 2\}$. Within a test, all items were associated with the same k_i value. One hundred data sets were generated at each sample size, test length, and k_i combination.

The data-generating item parameters were drawn from distributions originally described by Liang (2007, p. 49):

$$\xi_i \sim \text{unif}(-1, 1), \tag{3.1}$$

$$\lambda_i \sim \text{unif}(0.3, 2.5), \tag{3.2}$$

$$\alpha_{1i}, \alpha_{2i} \sim \text{unif}(-1, 1), \tag{3.3}$$

and

$$\beta_{1i}, \beta_{2i} \sim \text{unif}(0, 1). \tag{3.4}$$

Liang arrived at these distributions after experimenting to find data-generating parameters that would yield commonly seen IRFs. For the data simulated in this study, these data-generating parameters yielded classical item difficulties (i.e., observed item-level proportions correct) ranging from .16 to .84. Histograms of the observed proportion correct values for the three k_i values are shown in Figure 3.1. These plots indicate that items generated according to these parameters yield classical item difficulties that are not too extreme and that are roughly symmetric about .5. Thus, the IRFs generated for this study lead to a fair amount

of variability in examinee responses.

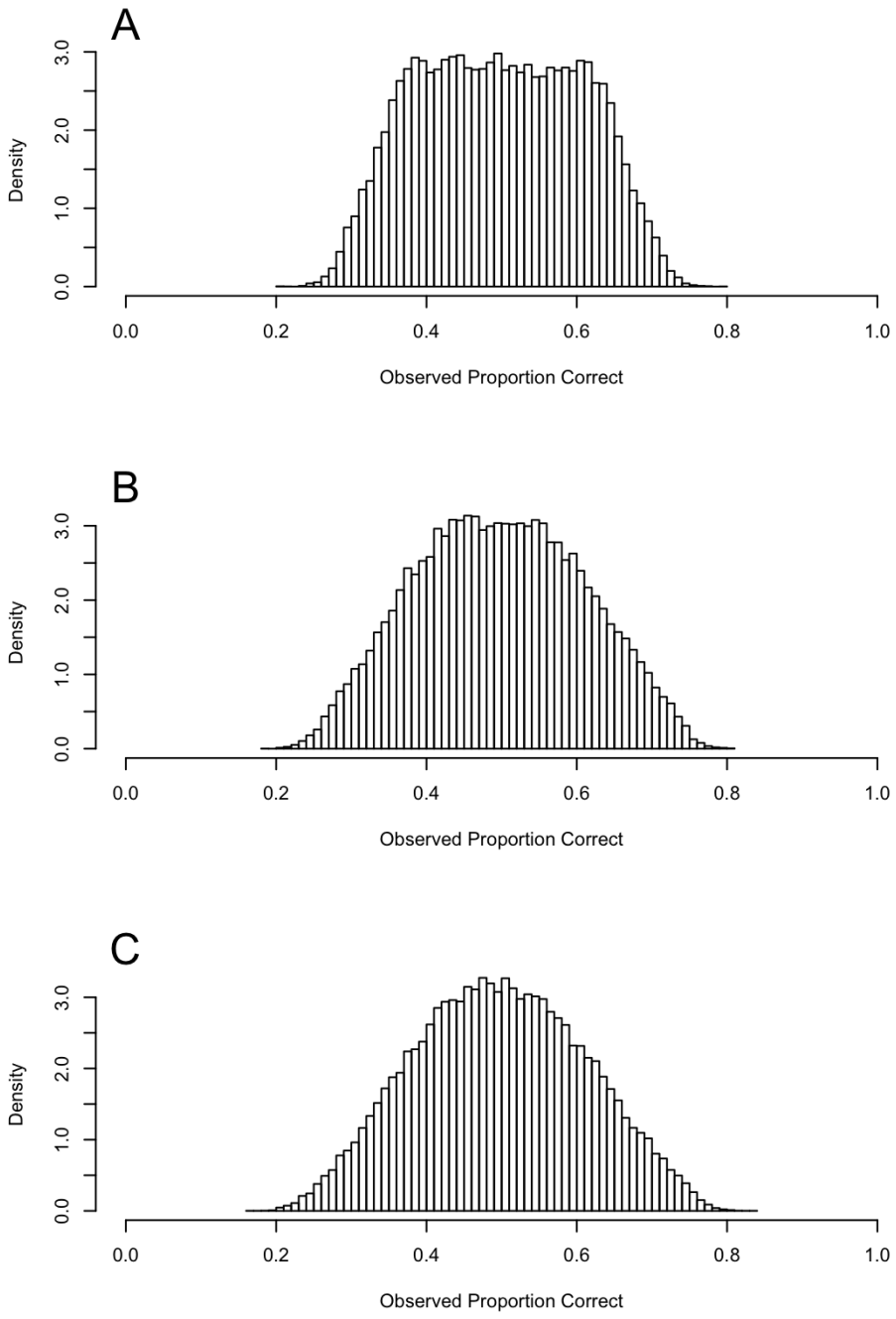

Figure 3.1. Histogram of classical item difficulties for $k_i = 0$ (Panel A), $k_i = 1$ (Panel B), and $k_i = 2$ (Panel C).

New population IRFs were randomly generated for each test length and sample size condition. Each data set was fit using both the fixed-effects and the random-effects estimation methods. For each method, data sets were fit three times by setting the complexity of the *estimated* IRFs, denoted $\tilde{k}_i$, equal to either 0, 1, or 2 for all test items. IRFs for each data set were also estimated using kernel smoothing (Ramsay, 1991) via the `kernSmoothIRT` package (Mazza, Punzo, & McGuire, 2014) in `R` (R Core Team, 2015). Kernel smoothed IRF estimates were obtained using a Gaussian kernel with bandwidths selected by Silverman's (1968) rule of thumb, 51 evaluation points, and conditioning on normalized ranked sum scores. In aggregate, there were seven sets of estimates for each simulated data set—fixed-effects and random-effects crossed with $\tilde{k} \in \{0, 1, 2\}$, plus kernel smoothing. Trait estimates were computed conditional on the estimated IRFs using EAP estimation with a standard normal prior for the FMP conditions, and using maximum likelihood for the kernel smoothed estimates.

Model convergence for the FMP conditions was assessed in several ways. For the fixed-effects simulations, parameter estimates were obtained by maximizing the log likelihood in Equation 2.23 using the BFGS optimization method (Broyden, 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970). In the BFGS optimization routine, the maximum number of iterations was set to 500 so that all sets of estimated parameters converged. For random-effects estimation, 25 quadrature points were used to approximate a standard normal distribution of ability. In the M step of the random-effects method, the BFGS method was used to maximize the marginal likelihood. Again, the maximum number of iterations was set to 500 so that all sets of estimated parameters converged at each iteration of the M step.

In the random-effects conditions, the E and M steps iterated until no parameter estimate on the $\boldsymbol{\gamma}$ metric changed by more than .0001 on successive iterations. All FMP models that were estimated in this study met these convergence criteria.

To assess model estimation accuracy, I considered item parameter recovery, item function recovery, and latent trait score recovery. Item parameter recovery was evaluated by looking at the accuracy of the $\hat{b}_{si}$ parameters, $s = 1, \ldots, 2k_i + 1$ in conditions for which $k_i = \tilde{k}_i$. Item function recovery was assessed using RIMSE_i with a standard normal target trait distribution. Latent trait recovery was assessed by Pearson product-moment, Spearman's ρ, and Kendall's τ correlation coefficients. Finally, I recorded the $\tilde{k}_i$ values associated with the items (fixed-effects conditions) or tests (random-effects conditions) selected by the AIC and BIC model selection criteria. The performance of these model selection criteria was assessed by comparing the AIC- and BIC-selected $\tilde{k}_i$ values to the data-generating k_i values, and by comparing the distribution of RIMSE_i values for the AIC- and and BIC-selected IRFs.

3.2 Results

3.2.1 Item parameter recovery

Recovery was first assessed by looking at the accuracy of estimated b_{si} parameters, $s = 1, \ldots, 2k_i+1$. Although most $\hat{b}_{si}$ were relatively close to 0 (all data-generating b_{si} values were less than 6 in absolute value, and 83% of data-generating b_{si} values were less than 1 in absolute value), in some cases, fixed-effects and random-effects

estimation yielded extreme $\hat{b}_{si}$ values. For instance, in the smallest data condition ($N = 200$ subjects and $I = 20$ items) and for cases in which $\tilde{k}_i = k_i = 2$, $\hat{b}_{4i}$ values ranged from -105 to +67 for fixed-effects estimation, and from -79 to +156 for random-effects estimation, even though data-generating b_{4i} values only ranged from -3.6 to $+3.6$. Note that all studied conditions met the previously outlined convergence criteria. Thus, these extreme parameter estimates are not the result of model non-convergence. Moreover, high variability in b_{si} values was observed even in the largest studied data conditions. Figure 3.2 shows the distribution of simple errors $(\hat{b}_{si} - b_{si})$, $s = 0, \ldots 2k_i + 1$ for the subset of conditions in which $\tilde{k}_i = k_i$, $N = 5,000$, and $I = 60$. In this figure, Panel A displays results for fixed-effects estimation, and Panel B displays results for random-effects estimation. Recall that the b_{2i}, b_{3i}, b_{4i}, and b_{5i} parameters are not estimated when $\tilde{k}_i = 0$ and that the b_{4i} and b_{5i} parameters are not estimated when $\tilde{k}_i = 1$, and so there are no errors in $\hat{b}_{si}$ associated with these conditions. Figure 3.2 shows that $\hat{b}_{0i}$ and $\hat{b}_{1i}$ are recovered well when $\tilde{k}_i = k_i = 0$. This is reassuring because the $k_i = 0$ FMP model is equivalent to the familiar 2PL, a model in which the estimated item parameters are routinely interpreted. For the $\tilde{k}_i = 1$ and $\tilde{k}_i = 2$ conditions, the $\hat{b}_{si}$ are only slightly biased, but there exists a fair amount of variability, especially in the $\tilde{k}_i = k_i = 2$ conditions. The biases of parameter estimates are similar for the fixed-effects and random-effects conditions, although random-effects estimation leads to slightly more variable parameter estimates than fixed-effects estimation. One way to decrease the variability of these estimates would be to adopt a marginal Bayesian estimation approach (see Falk & Cai, 2015, 2106). However, accuracy of $\hat{b}_{si}$ values is not necessarily the most relevant index of model recovery. That

is, even if individual $\hat{b}_{si}$ parameters are not estimated accurately, other model predictions may still be reliable. Recall that the FMP item response model is a quasi-parametric item response model (Liang & Browne, 2015; Ramsay, 1991). As such, the item parameters b_{si} are included to provide flexibility to the model, and are not intended for interpretation. Thus it may be the case that, even if individual item parameters are estimated inaccurately, other relevant quantities are estimated accuracy. In the following subsections, recovery is assessed in terms of item response function recovery and latent trait score recovery.

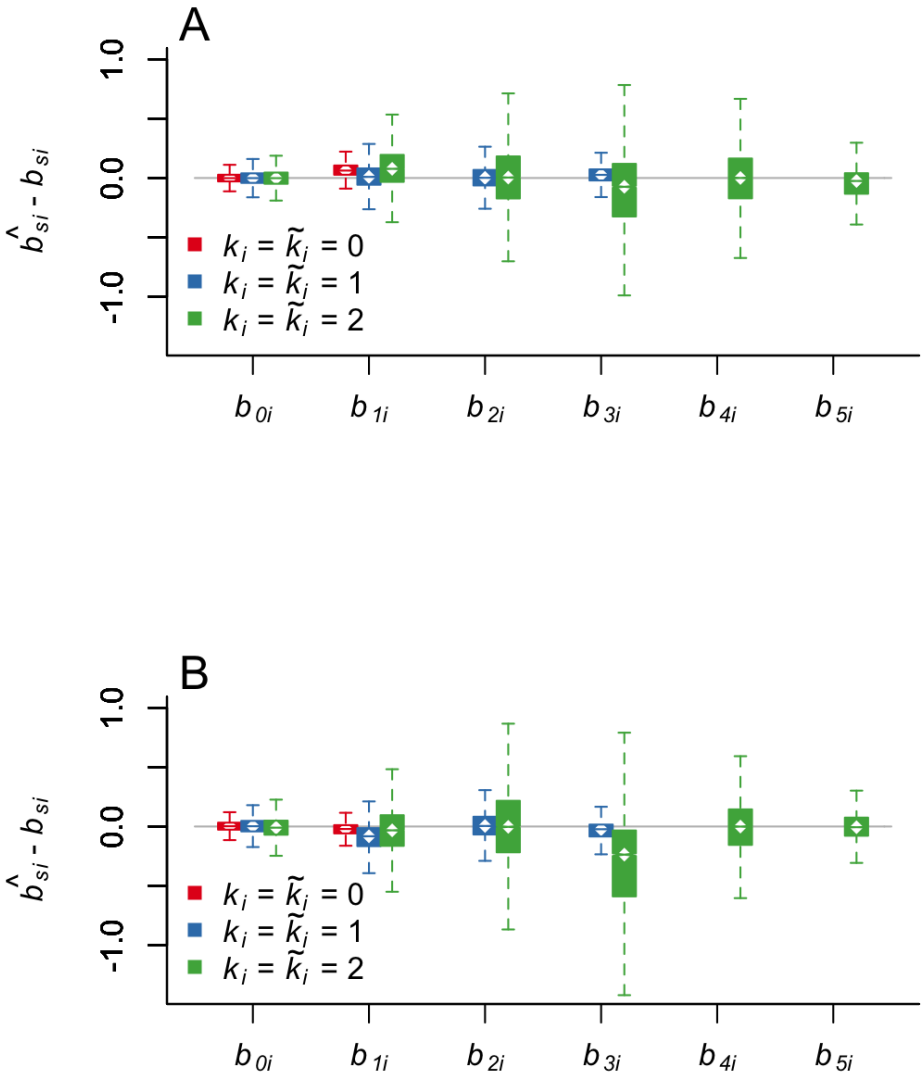

Figure 3.2. Distribution of errors in estimating FMP item parameters. Errors are defined as $\hat{b}_{si} - b_{si}$ for $s \in \{0, 1, 2, 3, 4, 5\}$, and only conditions with $N = 5,000$ subjects, $I = 60$ items, and $k_i = \tilde{k}_i$ are included. Panel A shows results for fixed-effects estimation, and Panel B shows results for random-effects estimation. Box plot whiskers extend from $1Q - 1.5IQR$ to $3Q + 1.5IQR$, and outliers are not displayed. Note that b_{4i} and b_{5i} are not estimated when $\tilde{k}_i = 1$, and b_{2i}, b_{3i}, b_{4i}, and b_{5i} are not estimated when $\tilde{k}_i = 0$.

3.2.2 Item response function recovery

The accuracy of an estimated item response function for item i can be quantified by the root integrated mean squared error (RIMSE_i; see Equation 2.7). Recall that RIMSE_i are on a probability metric, and that smaller RIMSE_i values indicate that the estimated IRF is a better approximation of the data-generating IRF. The across-item and across-replication RIMSE_i means and standard deviations are reported in Tables 3.1, 3.2, and 3.3 conditional on sample size, test length, and $\tilde{k}_i$ values. Tables 3.1, 3.2, and 3.3 report results for data generated under $k_i = 0$, $k_i = 1$, and $k_i = 2$, respectively. Recall that in all cases, both the data-generating θ distribution and the θ distribution assumed during item parameter estimation are standard normal.

Table 3.1

RIMSE$_i$ *means (standard deviations) for seven estimation methods,* $k_i = 0$

		Fixed-Effects			Random-Effects			Kernel
I	N	$\tilde{k}_i = 0$	$\tilde{k}_i = 1$	$\tilde{k}_i = 2$	$\tilde{k}_i = 0$	$\tilde{k}_i = 1$	$\tilde{k}_i = 2$	Smoothing
20	200	.048 (.025)	.061 (.023)	.068 (.021)	.043 (.023)	.066 (.027)	.078 (.027)	.061 (.026)
	500	.035 (.018)	.043 (.016)	.048 (.015)	.028 (.015)	.042 (.018)	.055 (.018)	.050 (.029)
	1,000	.028 (.013)	.034 (.011)	.038 (.012)	.019 (.010)	.029 (.012)	.042 (.014)	.045 (.034)
	2,000	.025 (.010)	.029 (.009)	.031 (.010)	.014 (.007)	.020 (.009)	.032 (.012)	.043 (.039)
	5,000	.023 (.007)	.025 (.007)	.027 (.009)	.009 (.005)	.013 (.006)	.021 (.008)	.041 (.043)
40	200	.044 (.024)	.056 (.023)	.064 (.021)	.043 (.023)	.063 (.025)	.075 (.024)	.062 (.025)
	500	.029 (.015)	.037 (.015)	.043 (.014)	.027 (.015)	.039 (.017)	.051 (.016)	.047 (.023)
	1,000	.022 (.011)	.027 (.011)	.031 (.012)	.019 (.010)	.026 (.011)	.036 (.012)	.040 (.023)
	2,000	.017 (.008)	.021 (.008)	.023 (.009)	.013 (.007)	.019 (.008)	.026 (.009)	.036 (.026)
	5,000	.013 (.006)	.016 (.005)	.017 (.007)	.009 (.005)	.012 (.005)	.016 (.006)	.032 (.030)
60	200	.043 (.023)	.055 (.023)	.062 (.020)	.044 (.025)	.064 (.026)	.079 (.026)	.063 (.026)
	500	.027 (.015)	.035 (.014)	.041 (.014)	.028 (.015)	.039 (.016)	.052 (.017)	.048 (.022)
	1,000	.020 (.011)	.026 (.011)	.030 (.011)	.021 (.012)	.028 (.012)	.036 (.013)	.040 (.022)
	2,000	.015 (.008)	.019 (.008)	.021 (.009)	.015 (.008)	.020 (.009)	.025 (.009)	.035 (.023)
	5,000	.011 (.006)	.013 (.006)	.015 (.007)	.010 (.007)	.013 (.006)	.015 (.006)	.030 (.025)

Table 3.2

RIMSE$_i$ *means (standard deviations) for seven estimation methods,* $k_i = 1$

		Fixed-Effects			Random-Effects			Kernel
I	N	$\tilde{k}_i = 0$	$\tilde{k}_i = 1$	$\tilde{k}_i = 2$	$\tilde{k}_i = 0$	$\tilde{k}_i = 1$	$\tilde{k}_i = 2$	Smoothing
20	200	.070 (.024)	.063 (.024)	.070 (.023)	.066 (.023)	.066 (.026)	.079 (.027)	.060 (.023)
	500	.062 (.020)	.046 (.017)	.050 (.017)	.057 (.020)	.044 (.017)	.054 (.018)	.045 (.024)
	1,000	.057 (.018)	.037 (.013)	.040 (.012)	.052 (.019)	.030 (.012)	.041 (.014)	.039 (.026)
	2,000	.055 (.017)	.032 (.011)	.034 (.011)	.049 (.019)	.022 (.009)	.031 (.011)	.036 (.029)
	5,000	.054 (.017)	.029 (.009)	.030 (.010)	.047 (.019)	.014 (.006)	.020 (.008)	.034 (.031)
40	200	.067 (.023)	.058 (.023)	.064 (.022)	.066 (.024)	.064 (.025)	.077 (.026)	.059 (.023)
	500	.057 (.019)	.039 (.015)	.044 (.015)	.056 (.020)	.041 (.017)	.052 (.018)	.042 (.018)
	1,000	.053 (.018)	.029 (.011)	.032 (.011)	.051 (.019)	.029 (.012)	.037 (.013)	.033 (.017)
	2,000	.051 (.018)	.022 (.009)	.025 (.009)	.049 (.019)	.021 (.008)	.027 (.009)	.027 (.018)
	5,000	.049 (.018)	.018 (.006)	.019 (.006)	.047 (.019)	.014 (.006)	.017 (.006)	.023 (.020)
60	200	.066 (.023)	.057 (.023)	.064 (.022)	.069 (.025)	.065 (.026)	.078 (.027)	.060 (.025)
	500	.055 (.019)	.037 (.015)	.042 (.014)	.057 (.020)	.043 (.018)	.052 (.018)	.043 (.018)
	1,000	.051 (.018)	.027 (.011)	.031 (.011)	.053 (.019)	.033 (.014)	.039 (.014)	.033 (.016)
	2,000	.049 (.018)	.020 (.008)	.022 (.009)	.049 (.019)	.024 (.011)	.028 (.011)	.026 (.016)
	5,000	.048 (.018)	.015 (.006)	.016 (.006)	.047 (.019)	.017 (.008)	.020 (.007)	.021 (.017)

Table 3.3

RIMSE$_i$ *means (standard deviations) for seven estimation methods,* $k_i = 2$

		Fixed-Effects			Random-Effects			Kernel
I	N	$\tilde{k}_i = 0$	$\tilde{k}_i = 1$	$\tilde{k}_i = 2$	$\tilde{k}_i = 0$	$\tilde{k}_i = 1$	$\tilde{k}_i = 2$	Smoothing
20	200	.080 (.027)	.065 (.025)	.071 (.024)	.077 (.026)	.067 (.025)	.079 (.027)	.060 (.023)
	500	.073 (.023)	.048 (.018)	.052 (.018)	.068 (.023)	.046 (.018)	.055 (.019)	.044 (.022)
	1,000	.069 (.022)	.040 (.014)	.042 (.014)	.064 (.023)	.032 (.012)	.041 (.013)	.037 (.022)
	2,000	.067 (.022)	.035 (.012)	.036 (.013)	.062 (.023)	.024 (.009)	.031 (.011)	.033 (.024)
	5,000	.066 (.022)	.031 (.011)	.032 (.011)	.061 (.023)	.017 (.006)	.020 (.008)	.030 (.025)
40	200	.077 (.025)	.059 (.023)	.066 (.022)	.080 (.027)	.069 (.028)	.079 (.028)	.060 (.023)
	500	.069 (.023)	.041 (.015)	.045 (.015)	.069 (.023)	.044 (.018)	.052 (.018)	.041 (.016)
	1,000	.065 (.022)	.031 (.012)	.034 (.011)	.065 (.023)	.033 (.014)	.040 (.014)	.032 (.015)
	2,000	.064 (.022)	.025 (.009)	.026 (.009)	.063 (.023)	.026 (.010)	.030 (.011)	.026 (.016)
	5,000	.062 (.022)	.020 (.007)	.020 (.007)	.061 (.023)	.018 (.006)	.020 (.007)	.021 (.017)
60	200	.076 (.025)	.058 (.023)	.064 (.022)	.080 (.028)	.068 (.029)	.079 (.029)	.061 (.023)
	500	.067 (.023)	.038 (.014)	.042 (.014)	.071 (.024)	.048 (.021)	.054 (.020)	.042 (.016)
	1,000	.064 (.022)	.029 (.011)	.032 (.011)	.066 (.023)	.036 (.016)	.040 (.015)	.032 (.013)
	2,000	.062 (.022)	.023 (.008)	.023 (.008)	.062 (.023)	.032 (.013)	.034 (.014)	.025 (.012)
	5,000	.061 (.023)	.018 (.006)	.017 (.006)	.061 (.023)	.025 (.009)	.026 (.009)	.018 (.012)

For all estimation methods and each k_i condition, increasing sample size and test length improves IRF recovery. The exception to this trend occurs for 60-item tests and random-effects estimation. Specifically, conditional on sample size, the average RIMSE_i values for 60-item tests are not always smaller than the average RIMSE_i values for the 40- or 20-item tests. One possible explanation for this phenomenon is that random-effects estimation requires all item parameters to be estimated simultaneously. Thus for random-effects estimation, increasing test length increases the number of model parameters that must be estimated at once. In contrast, for the fixed-effects and kernel smoothing estimation methods, IRFs are estimated one-at-a-time conditional on surrogate θ values. For fixed-effects estimation and kernel smoothing, test length affects the surrogate θ values, but does not affect the number of parameters that are estimated simultaneously. Perhaps for this reason, these data demonstrate an interaction between test length and estimation method in terms of IRF recovery. Namely, random-effects estimation leads to smaller RIMSE_i values than fixed-effects estimation in 20-item tests, except for a small number of conditions for which sample size is small and $\tilde{k}_i = 1$ or $\tilde{k}_i = 2$. In contrast, in 60-item tests, fixed-effects estimation often performs as well or better than random-effects estimation.

With sufficient model flexibility (i.e., large $\tilde{k}_i$ values) and large data sets, all estimation methods should lead to excellent IRF recovery. However, although larger $\tilde{k}_i$ values permit more flexibility in the estimated IRFs, it is not necessarily the case that this increased flexibility leads to better IRF recovery. For instance, kernel smoothing allows for more flexible IRFs than any of the studied FMP models, yet in these simulations it often has the worst IRF recovery. This could

be attributed partly to the crude way of specifying the surrogate θ values. As implemented in this study, kernel smoothing conditions on normalized ranked sum scores whereas fixed-effects FMP conditions on normalized first principal component scores. These results show that kernel smoothing often leads to higher average RIMSE values than fixed-effects or random-effects FMP, particularly when $k_i = 0$.

Returning to the FMP results, it is not necessarily the case that estimating IRFs with $\tilde{k}_i = 2$ leads to better IRF recovery than $\tilde{k}_i = 1$ or $\tilde{k}_i = 0$, even though $\tilde{k}_i = 2$ permits greater flexibility. Moreover, it is not necessarily the case that estimating IRFs using the data-generating k_i value (i.e., $\tilde{k}_i = k_i$) leads to better IRF recovery than other $\tilde{k}_i$ values. In fact, data sets generated under $k_i = 2$ often are associated with smaller RIMSE_i values when $\tilde{k}_i < 2$, that is, under a model that cannot recover the data-generating IRF exactly. Closer inspection of Tables 3.1, 3.2, and 3.3 emphasizes the importance of choosing an appropriate $\tilde{k}_i$ value. When data are generated according to the 2PL ($k_i = 0$, Table 3.1), the lowest RIMSE_i values occur when $\tilde{k}_i = 0$, for both fixed-effects and random-effects estimation at all studied sample sizes. This is true even though the $\tilde{k}_i = 0$ model is nested within the $\tilde{k}_i = 1$ and $\tilde{k}_i = 2$ models. This is likely because the more complex models are more difficult to estimate and more likely to reflect random variation in a dataset, especially at small sample sizes. As sample size increases, the influence of random variation decreases, and we find that the RIMSE_i values in the $\tilde{k}_i = 0$ conditions become increasingly similar to the RIMSEs in the $\tilde{k}_i > 0$ conditions. When $k_i = 1$ (Table 3.2) the lowest RIMSE_i values occur when $\tilde{k}_i = 1$ for both fixed-effects and random-effects estimation. For the $k_i = 1$

conditions, the improvement in average RIMSE_i when using $\tilde{k}_i = 1$ versus $\tilde{k}_i = 0$ is most prominent at *large* sample sizes. In contrast, the improvement in average RIMSE_i when using $\tilde{k}_i = 1$ versus $\tilde{k}_i = 2$ is usually minimized in larger samples. Additionally, when $k_i = 1$ and sample size is large, random-effects estimation tends to lead to better IRF recovery than fixed-effects estimation in 20-item and 40-item tests. When $k_i = 2$ (Table 3.3), $\tilde{k}_i = 1$ tends to lead to equal or smaller average RIMSE_i values at the studied sample sizes. It may be the case that if larger sample sizes were studied, $\tilde{k}_i = 2$ would outperform $\tilde{k}_i = 1$ in the $k_i = 2$ conditions.

In summary, IRF recovery depends on sample size, test length, the complexity of the data-generating curve, and the estimation method. No one estimation method works best for all types of data. For short tests, random-effects estimation tends to outperform fixed-effects FMP or kernel smoothing estimation, although random-effects and fixed-effects perform similarly when a short test is given to a large sample. Kernel smoothing often leads to higher and more variable RIMSE_i values than the FMP methods, and is not recommended when IRF accuracy is a primary concern. Overall, it is recommended to use FMP methods over kernel smoothing with the smoothing options used in this study.

3.2.3 Latent trait score recovery

For some applications, IRF recovery may be less important than trait score recovery. In this study, latent trait score recovery was assessed using three types of correlation coefficients—the Pearson product-moment correlation, Spearman's

rank-order correlation, and Kendall's τ correlation—computed between the true and estimated latent trait scores. Note that in all cases the estimated latent trait scores were computed after IRF estimation. As such, the surrogate θ values used for fixed-effects FMP and using kernel smoothing estimation are not included in these correlations. Latent trait score recovery is summarized in Figures 3.3–3.11. Figures 3.3, 3.4, and 3.5 display results for Pearson product-moment correlations, Figures 3.6, 3.7, and 3.8 display results for Spearman rank correlations, and Figures 3.9, 3.9, and 3.11 display results for Kendall's τ correlations. Pearson product-moment correlations index the degree of linear association between two variables, Spearman rank correlations index the degree of linear association between the *ranks* of two variables, and Kendall's τ correlation equals the proportion of concordant rank pairs minus the proportion of discordant rank pairs between two variables. For all types of correlation coefficients, values closer to 1 indicate a stronger monotonic relationship between the true and estimated latent trait scores.

Each panel in Figures 3.3–3.11 displays box plots of the distribution of correlations between the data-generating θ values and the $\hat{\theta}$ values obtained for each estimation method. Overall, there are only small differences among the different estimation methods in terms of the distribution of correlation coefficients. Aggregating over sample size and test length conditions, fixed-effects estimation with $\tilde{k}_i = 1$ has higher median Pearson product-moment and Spearman correlations than the other six estimation methods. Fixed-effects estimation with $\tilde{k}_i = 1$ has a median Pearson correlation equal to .9722 and a median Spearman correlation equal to .9786. For the other estimation methods, median Pearson correlations

equal .9678–.9716 and median Spearman correlations equal .9769–.9783. In contrast, Kendall's τ correlations are smaller in magnitude than Pearson and Spearman correlations, and are consistently higher for kernel smoothing-based latent trait estimates than for FMP-based latent trait estimates. For kernel smoothing, median Kendall correlations equal .8790, and for the other estimation methods, median Kendall correlations equal .8665–.8717. Thus, although there are consistent differences among estimation methods, the magnitudes of the median recovered correlations are highly similar across methods.

Taking a closer look at Pearson product-moment correlations (Figures 3.3, 3.4, and 3.5), kernel smoothing estimation leads to consistently lower correlations than the FMP estimation methods, especially for large samples. At small samples, random-effects estimation leads to slightly lower correlations than fixed-effects estimation, especially when $\tilde{k}_i \neq k_i$. However, correlations do not improve much as sample size increases and there are few differences in the distribution of correlation coefficients from $N = 1,000$ to $N = 5,000$. In some conditions, the distribution of Pearson correlations stabilizes in samples as low as $N = 500$. Moreover, longer tests lead to higher correlations. Across conditions (excluding the kernel smoothing results), the median Pearson correlations for 20-, 40-, and 60-item tests equal .940, .968, and .978. Given these results, it is recommended that at least $I = 40$ items and $N = 1,000$ subjects are needed for accurate latent trait score recovery as indexed by Pearson correlations.

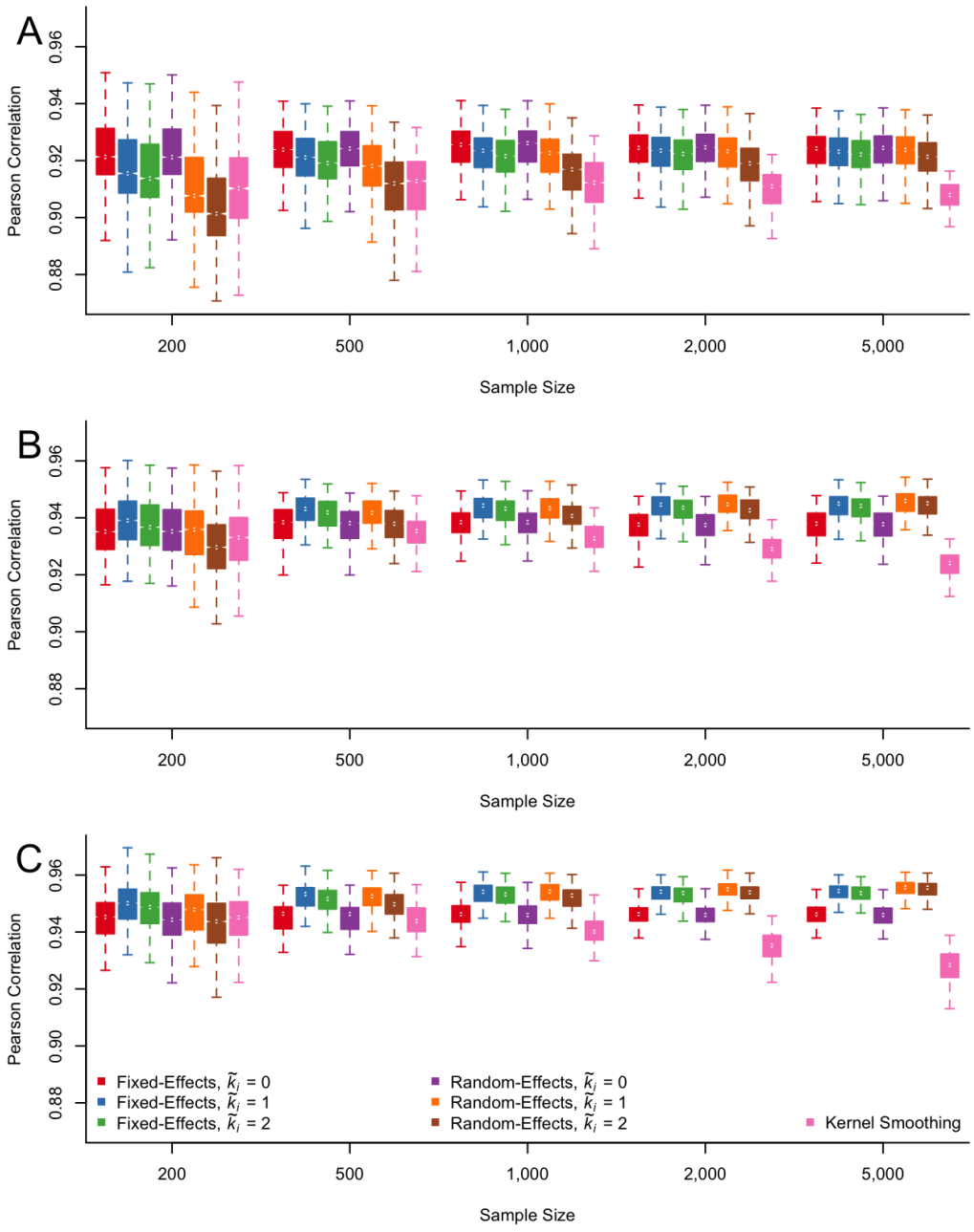

Figure 3.3. Distribution of Pearson correlations between θ and $\hat{\theta}$ for 20-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

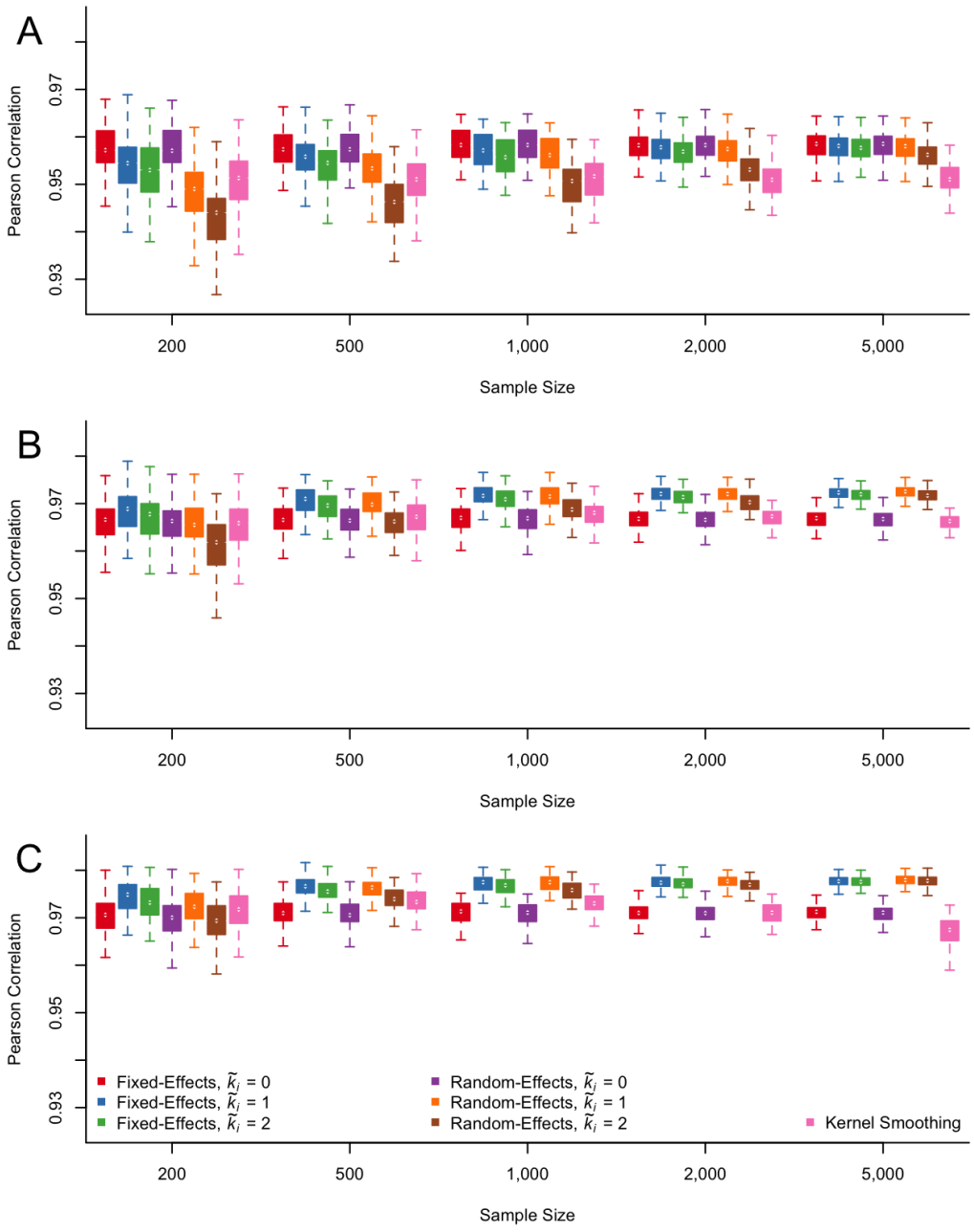

Figure 3.4. Distribution of Pearson correlations between θ and $\hat{\theta}$ for 40-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

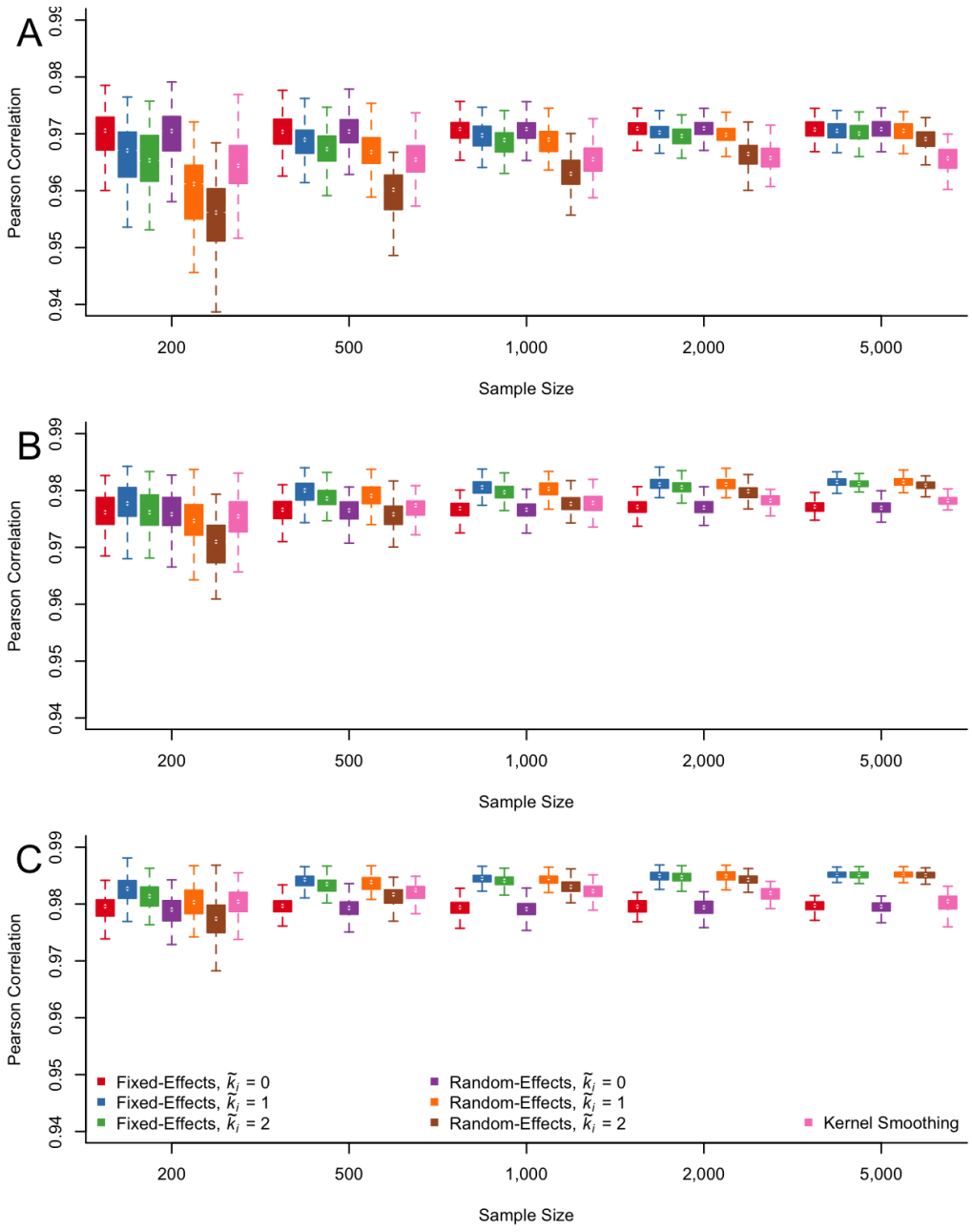

Figure 3.5. Distribution of Pearson correlations between θ and $\hat{\theta}$ for 60-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

Spearman rank-order correlations are shown in Figures 3.6, 3.7, and 3.8. Similar to Pearson correlations, longer tests lead to higher Spearman correlation coefficients. The median Spearman correlation for 20-, 40-, and 60-item tests (excluding kernel smoothing estimates) equal .949, .974, and .982. These results suggest that tests with at least 40 items are needed so that Spearman correlations are consistently high (e.g., above .95). Moreover, there are relatively few differences in the distribution of Spearman correlations across estimation methods. However, in small samples, random-effects estimation with $\tilde{k}_i > k_i$ leads to consistently lower correlations than the other estimation methods. For both fixed-effects and random-effects estimation, cases in which $\tilde{k}_i = 0$ and $k_i > 0$ lead to smaller correlations than the other estimation methods. Notably, the distribution of correlations resulting from kernel smoothing is comparable to the other methods. This is in contrast to Pearson correlations for which kernel smoothing led to lower correlations than other methods. Overall, as with Pearson correlations, data sets with at least $I = 40$ items and $N = 1,000$ subjects seems to result in Spearman correlations that are consistently large.

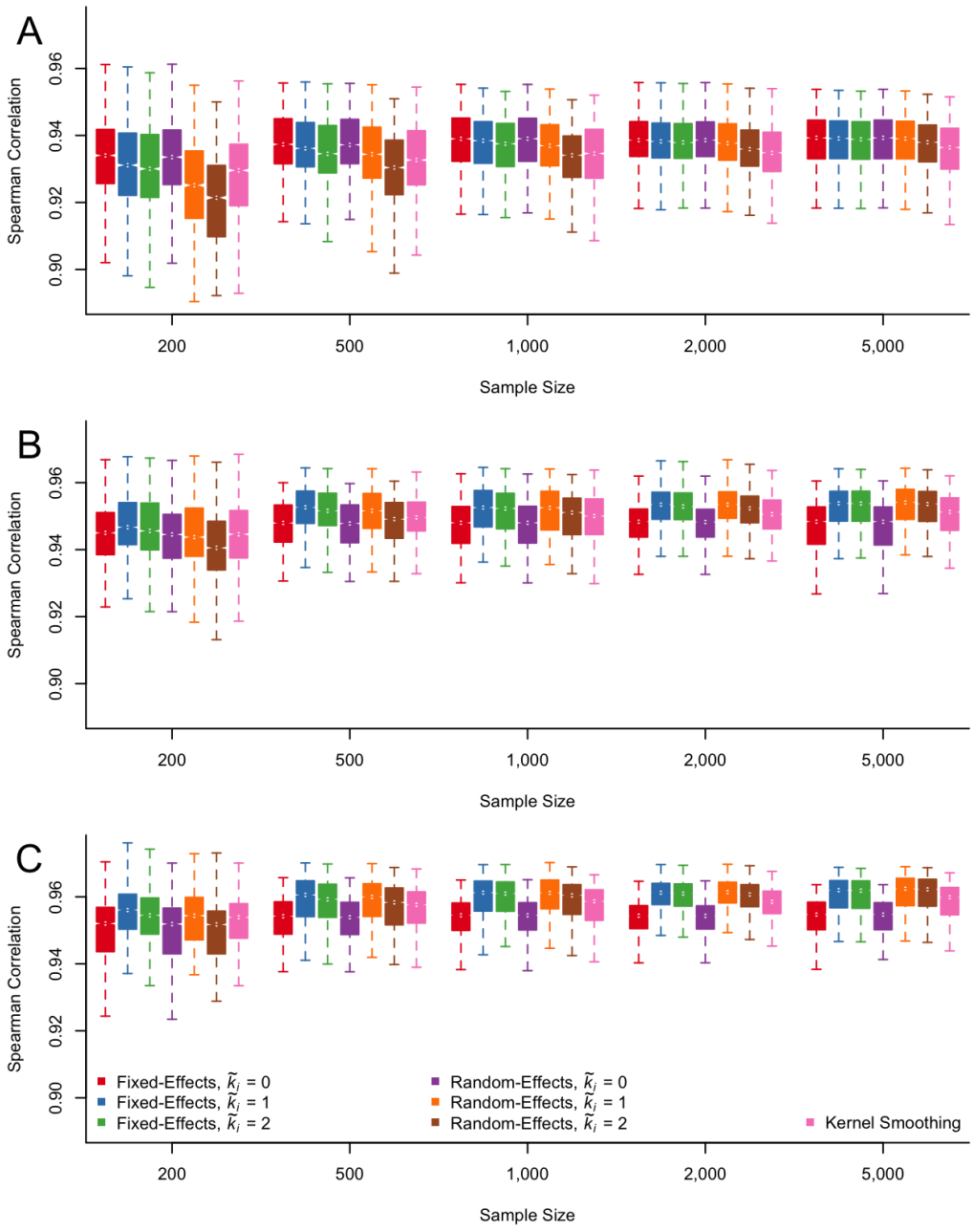

Figure 3.6. Distribution of Spearman correlations between θ and $\hat{\theta}$ for 20-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

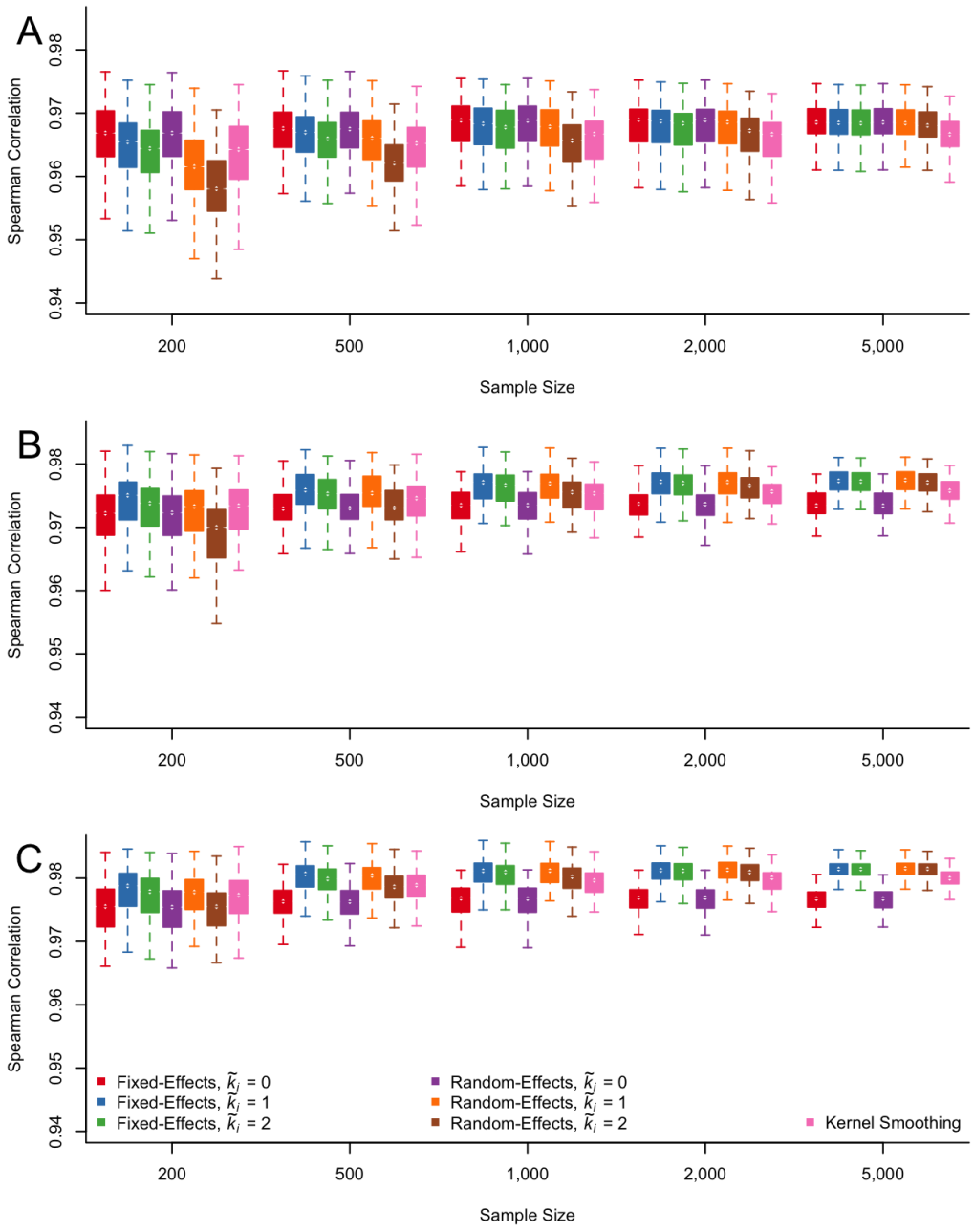

Figure 3.7. Distribution of Spearman correlations between θ and $\hat{\theta}$ for 40-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

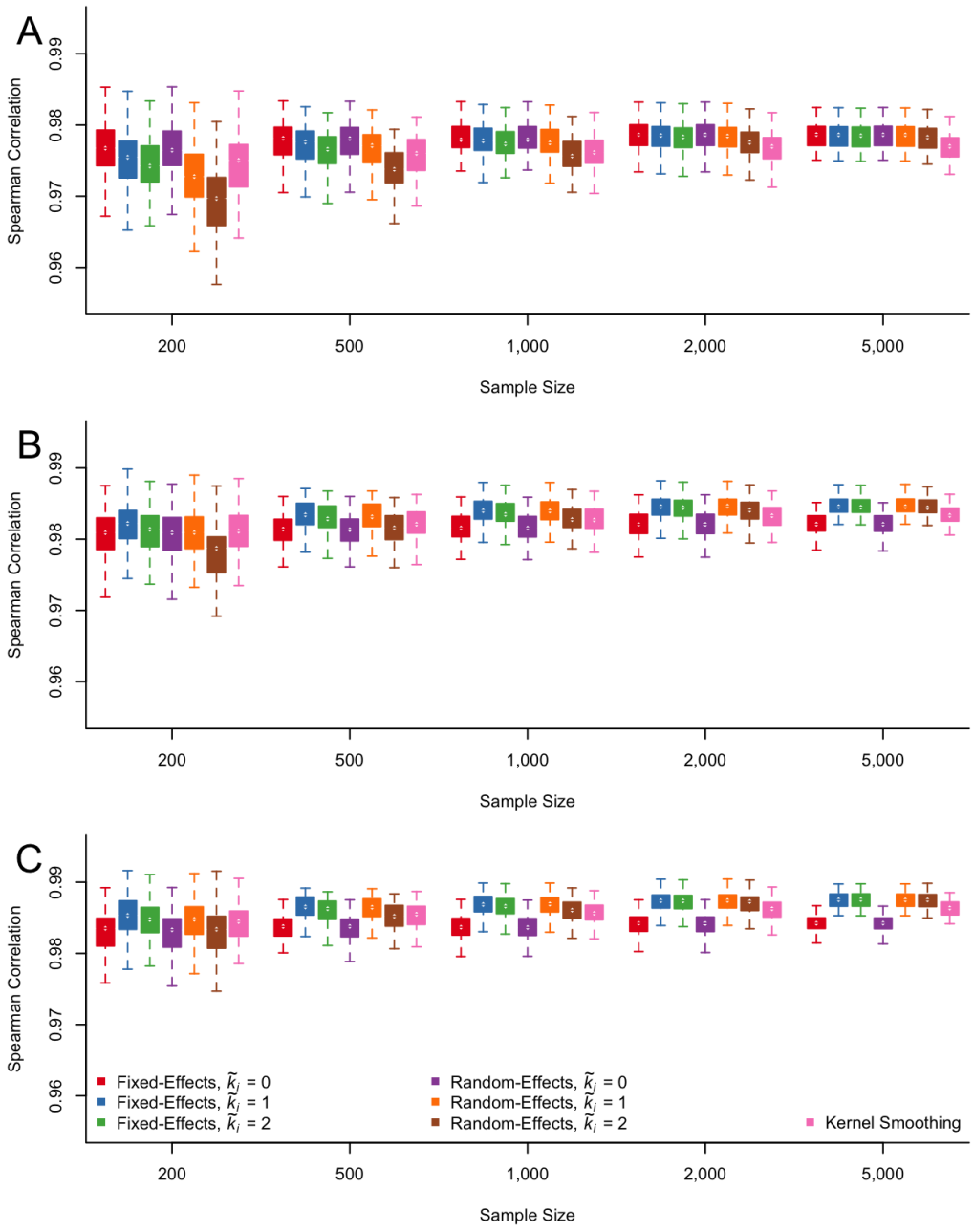

Figure 3.8. Distribution of Spearman correlations between θ and $\hat{\theta}$ for 60-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

The distributions of Kendall's rank-order correlation coefficient are shown in Figures 3.9, 3.10, and 3.11. In general, these correlations are smaller than Spearman or Pearson correlations. Moreover, whereas kernel smoothing tended to have lower Pearson correlations than the other methods, Kendall's τ correlations tend to be higher for kernel smoothing than for the other methods. As with the other correlation coefficients, the distribution of Kendall's correlations stabilize in $N = 1,000$. Moreover, correlations tend to be higher and less variable in $k_i = 2$ conditions than in $k_i = 0$ conditions. As before, $\tilde{k}_i = 0$ is noticeably worse than the other methods when $k_i = 1$ for both fixed-effects and random-effects estimation. Additionally, random-effects estimation with $\tilde{k}_i = 2$ when $k_i = 0$ leads to somewhat lower correlations than the other methods. Thus, so long as a suitable $\tilde{k}_i$ value is chosen and data sets are large enough, Kendall's rank-order correlations indicate good latent trait recovery.

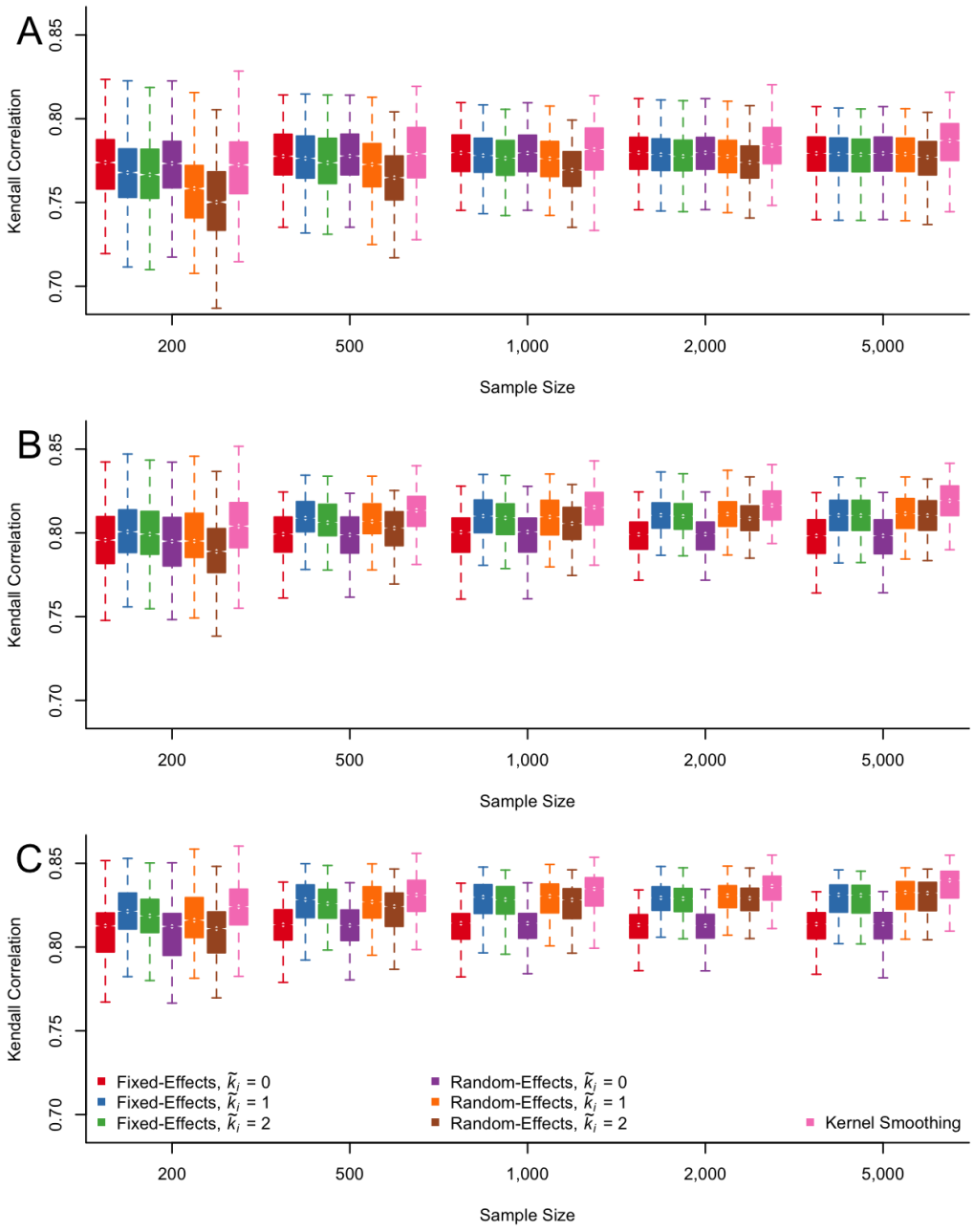

Figure 3.9. Distribution of Kendall correlations between θ and $\hat{\theta}$ for 20-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

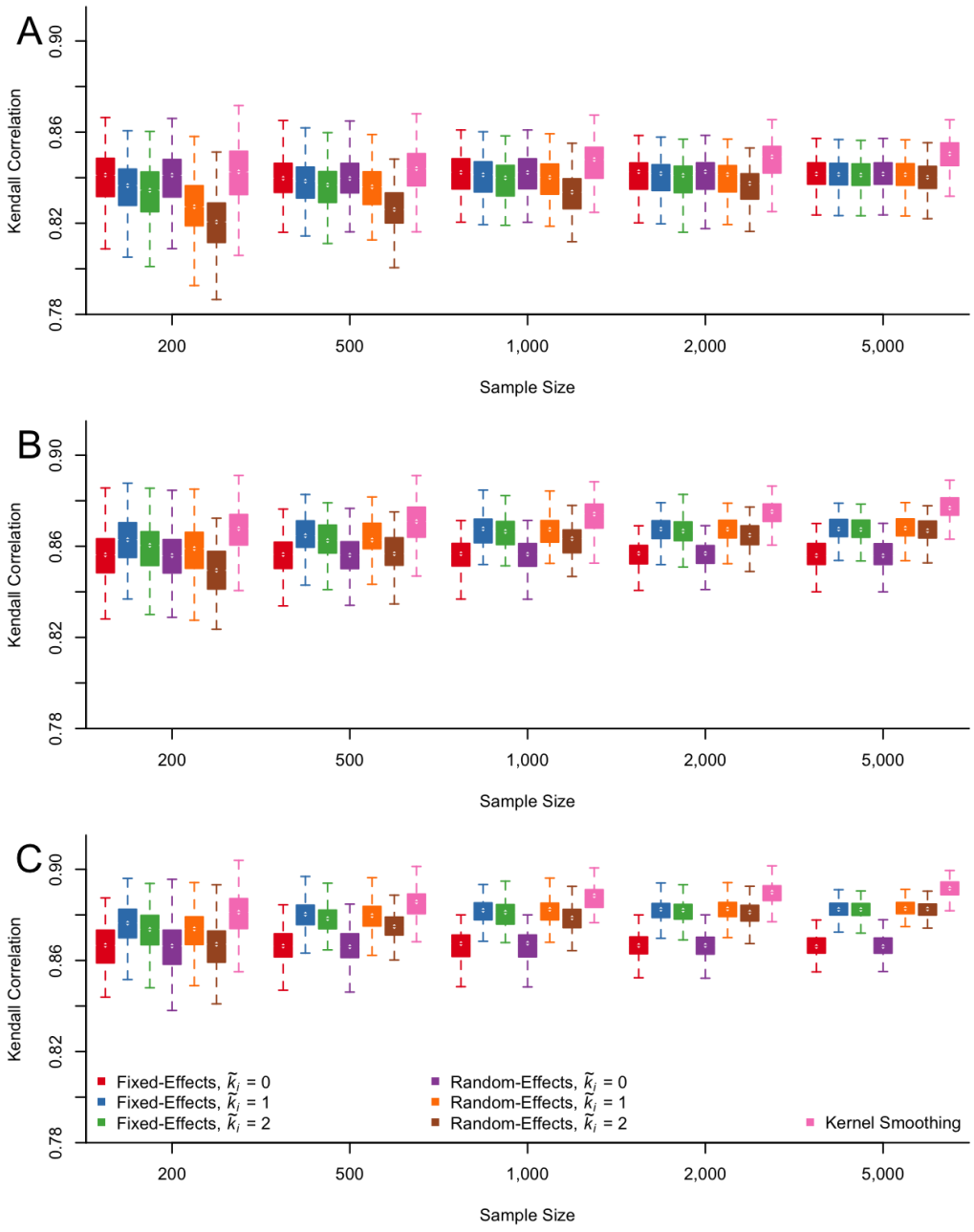

Figure 3.10. Distribution of Kendall correlations between θ and $\hat{\theta}$ for 40-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

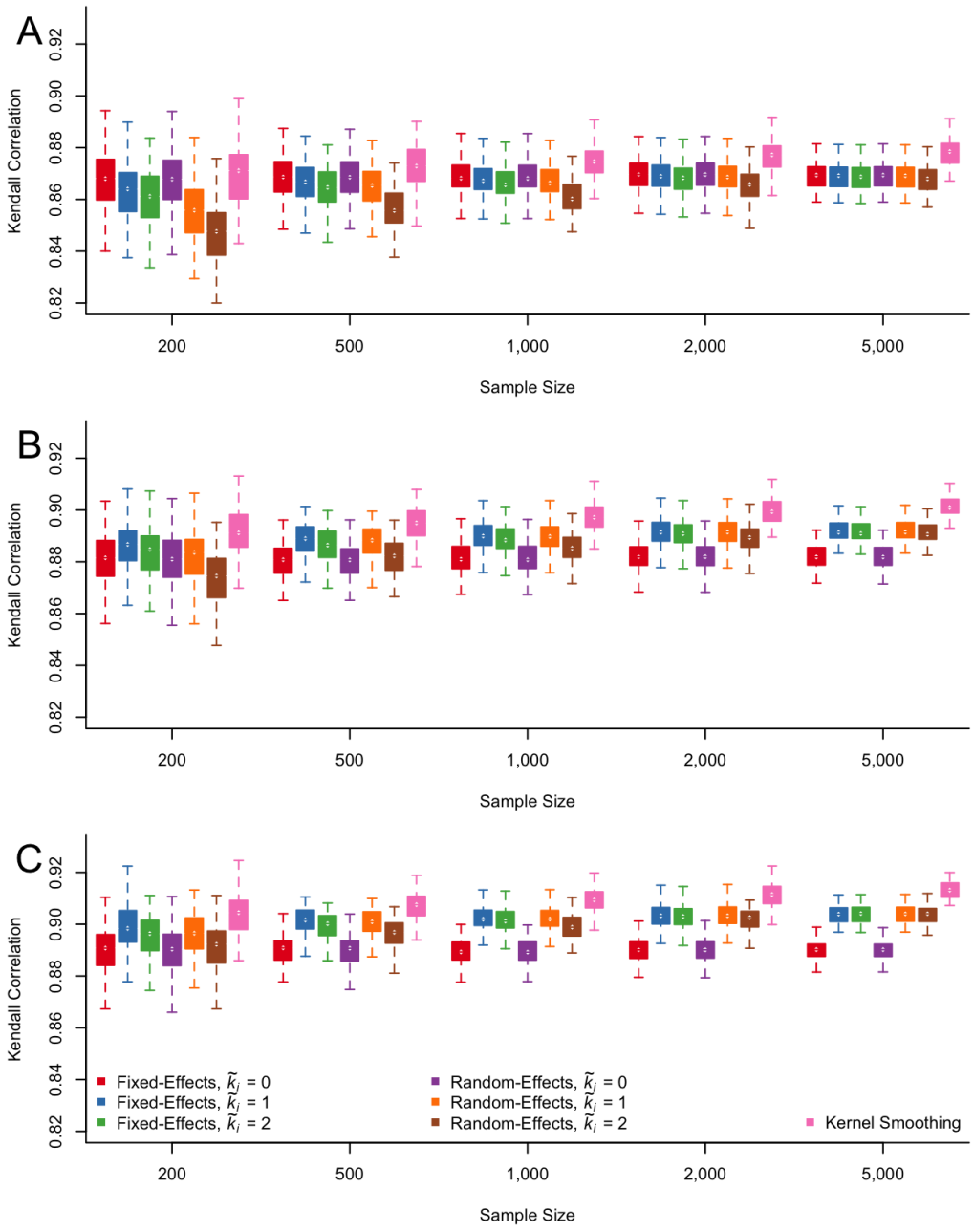

Figure 3.11. Distribution of Kendall correlations between θ and $\hat{\theta}$ for 60-item tests using the $\tilde{k}_i$ values chosen by the AIC and BIC criteria for fixed-effects and random-effects estimation, and using kernel smoothing estimation. Panel A displays results for $k_i = 0$, Panel B displays results for $k_i = 1$, and Panel C displays results for $k_i = 2$.

In general, the three types of correlation coefficients usually lead to the same conclusions about how accurately latent trait scores are recovered. Although there are some conditions in which some estimation methods are systematically lower than others, these differences are small and are not expected to affect substantive conclusions.

3.2.4 FMP model selection

Results shown in Tables 3.1, 3.2, and 3.3 demonstrated that larger $\tilde{k}_i$ values, although they allow greater flexibility in the estimated IRFs, do not necessarily provide better IRF recovery than smaller $\tilde{k}_i$ values. Moreover in practice, one must choose a $\tilde{k}_i$ value based on some model selection criterion such as the AIC or the BIC. To select a $\tilde{k}_i$ value, each data set can be fit multiple times using a sequence of $\tilde{k}_i$ values. The $\tilde{k}_i$ value associated with the smallest AIC or BIC value is said to be "selected" by the AIC or BIC. Below, I evaluate FMP model recovery in terms of the $\tilde{k}_i$ values selected by both the AIC and the BIC. Note that in this simulation design, $\tilde{k}_i$ is selected separately for each item under fixed-effects estimation, but $\tilde{k}_i$ is selected for the entire test under random-effects estimation.

The proportion of times each $\tilde{k}_i \in \{0, 1, 2\}$ value is chosen by the AIC or the BIC is shown conditional on sample size, test length, and k_i in Figures 3.12, 3.13, 3.14, and 3.15. Specifically, Figure 3.12 shows results for fixed-effects estimation and the AIC, Figure 3.13 shows results for random-effects estimation and the AIC, Figure 3.14 shows results for fixed-effects estimation and the BIC, and Figure 3.15 shows results for random-effects estimation and the BIC. In these figures, red

indicates that $\tilde{k}_i = 0$ was selected, blue indicates that $\tilde{k}_i = 1$ was selected, and green indicates that $\tilde{k}_i = 2$ was selected. Further, solid shading indicates that the AIC- or BIC-selected $\tilde{k}_i$ value equals the data-generating k_i value, whereas dashed shading represents cases in which $\tilde{k}_i \neq k_i$. Within each figure, bar plots are arranged into five rows representing the five sample sizes, and three columns representing three test lengths. Within each bar plot, the top row represents $k_i = 2$, the middle row represents $k_i = 1$, and the bottom row represents $k_i = 0$.

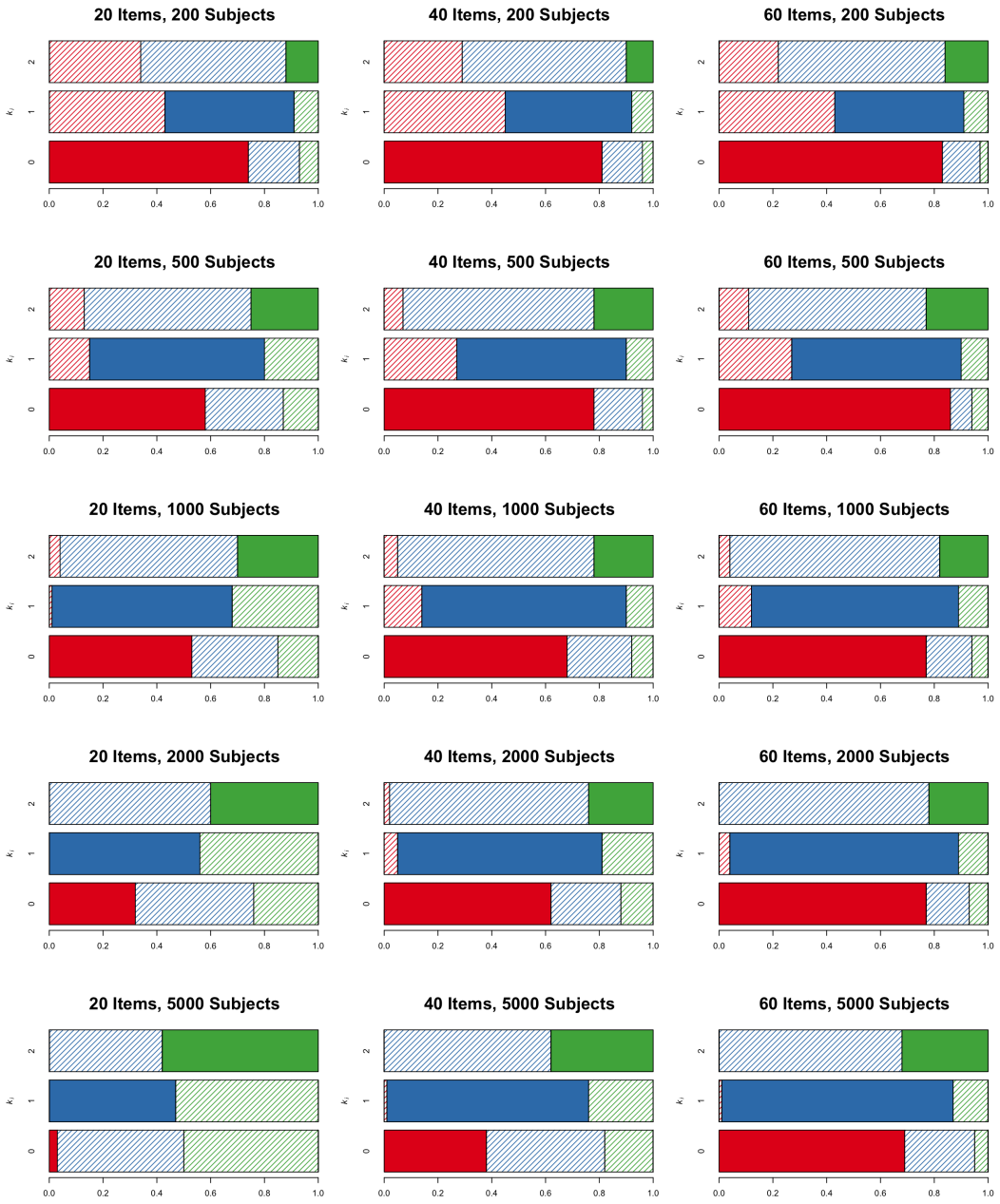

Figure 3.12. Distribution of $\tilde{k}_i$ values selected by the AIC criterion using fixed-effects estimation. Red indicates $\tilde{k}_i = 0$, blue indicates $\tilde{k}_i = 1$, and green indicates $\tilde{k}_i = 2$. Solid shading indicates that the chosen $\tilde{k}_i$ value matches the data-generating $\tilde{k}_i$ value.

Figure 3.13. Distribution of $\tilde{k}_i$ values selected by the AIC criterion using random-effects estimation. Red indicates $\tilde{k}_i = 0$, blue indicates $\tilde{k}_i = 1$, and green indicates $\tilde{k}_i = 2$. Solid shading indicates that the chosen $\tilde{k}_i$ value matches the data-generating $\tilde{k}_i$ value.

Overall, the AIC selected $\tilde{k}_i = k_i$ in 53% of the fixed-effects replications, and in 51% of the random-effects replications. The BIC selected $\tilde{k}_i = k_i$ value in 51% of the fixed-effects replications, and in 58% of the random-effects replications. Although random-effects estimation paired with the BIC has the highest proportion of $\tilde{k}_i = k_i$ replications, it is also informative to look at the types of mismatches associated with each method. Specifically, if the selected $\tilde{k}_i$ value is larger than the data-generating k_i value, the chosen model is too complex and is not the most parsimonious representation of the data. For this reason, it may be preferable to select a $\tilde{k}_i$ value that is too small rather than to select a $\tilde{k}_i$ value that is too large. Bear in mind that the proportion of times each $\tilde{k}_i$ value is chosen is only one indicator of recovery, and later I compare these model selection criteria in terms of their effects on IRF recovery.

Returning to Figures 3.12–3.15, we see that fixed-effects estimation paired with AIC model selection (Figure 3.12) is the only condition in which some nonzero proportion of $\tilde{k}_i$ values equal k_i in each combination of sample size, test length, and k_i conditions. For fixed-effects estimation and $k_i = 2$, AIC selects $\tilde{k}_i = 2$ with higher frequency as sample size increases, especially for short tests. In contrast, the proportion of times $\tilde{k}_i = 0$ is selected when $k_i = 0$ decreases as sample sizes increases. This unexpected result suggests that when $k = 0$ and $\tilde{k} \geq 0$, large samples lead to overfitted models and that the AIC does not impose a strong enough penalty in these conditions. A comparable result was reported by Liang (2007), who found that for fixed-effects estimation and $k = 0$, AIC selected $= \tilde{k} = 0$ in 81% of replications when $N = 300$ but only in 57% of replications when $N = 2,000$. A similar trend occurs for random-effects estimation and the AIC (see Figure 3.13). Namely,

for both fixed-effects and random-effects estimation, the AIC criterion tends to select simpler models (i.e., smaller $\tilde{k}_i$ values) in smaller samples and more complex models (i.e., larger $\tilde{k}_i$ values) in larger samples. However, for random-effects estimation, AIC selects $\tilde{k}_i = 2$ when the data-generating k_i value equals 0 or 1 in the vast majority of replications, particularly for tests with 20 or 40 items and at least 1,000 subjects. In 60-item tests estimated with random-effects, $\tilde{k}_i = 1$ and $\tilde{k}_i = 0$ values are selected more often than they are in shorter tests, but there are still a large number of cases in which AIC selects $\tilde{k}_i$ values that are larger than the data-generating k_i values. These results indicate that overly complex models are often selected when using AIC model selection with random-effects estimation. Thus, AIC is not recommended in conjunction with random-effects estimation. With fixed-effects estimation, AIC model selection does not select overly complex models as often as with random-effects estimation. However, there is still a sizable proportion of trials for which the AIC-selected $\tilde{k}_i > k_i$. A better solution may be to use a model selection criterion that imposes a greater penalty on model complexity, such as the BIC.

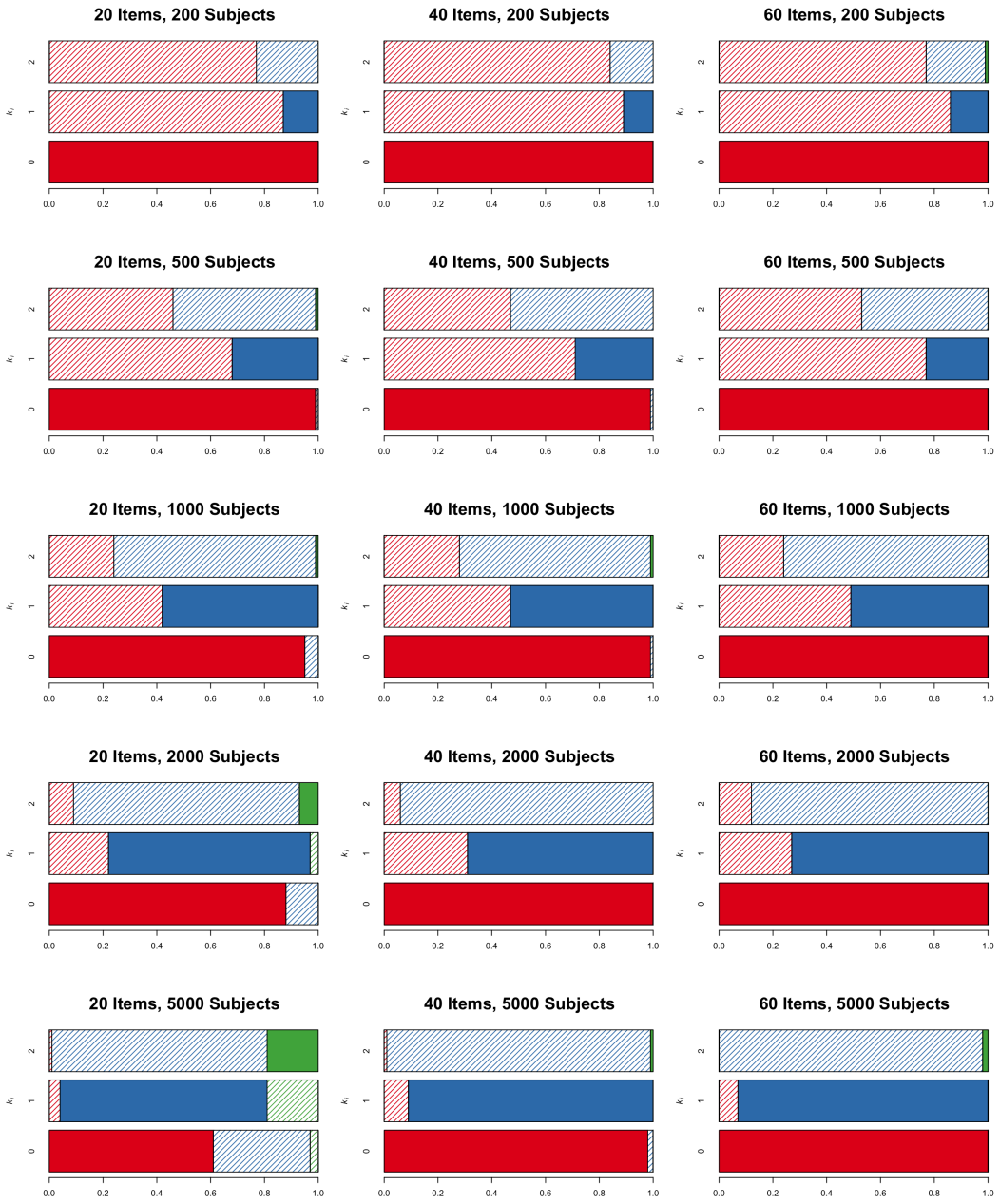

Figure 3.14. Distribution of $\tilde{k}_i$ values selected by the BIC criterion using fixed-effects estimation. Red indicates $\tilde{k}_i = 0$, blue indicates $\tilde{k}_i = 1$, and green indicates $\tilde{k}_i = 2$. Solid shading indicates that the chosen $\tilde{k}_i$ value matches the data-generating $\tilde{k}_i$ value.

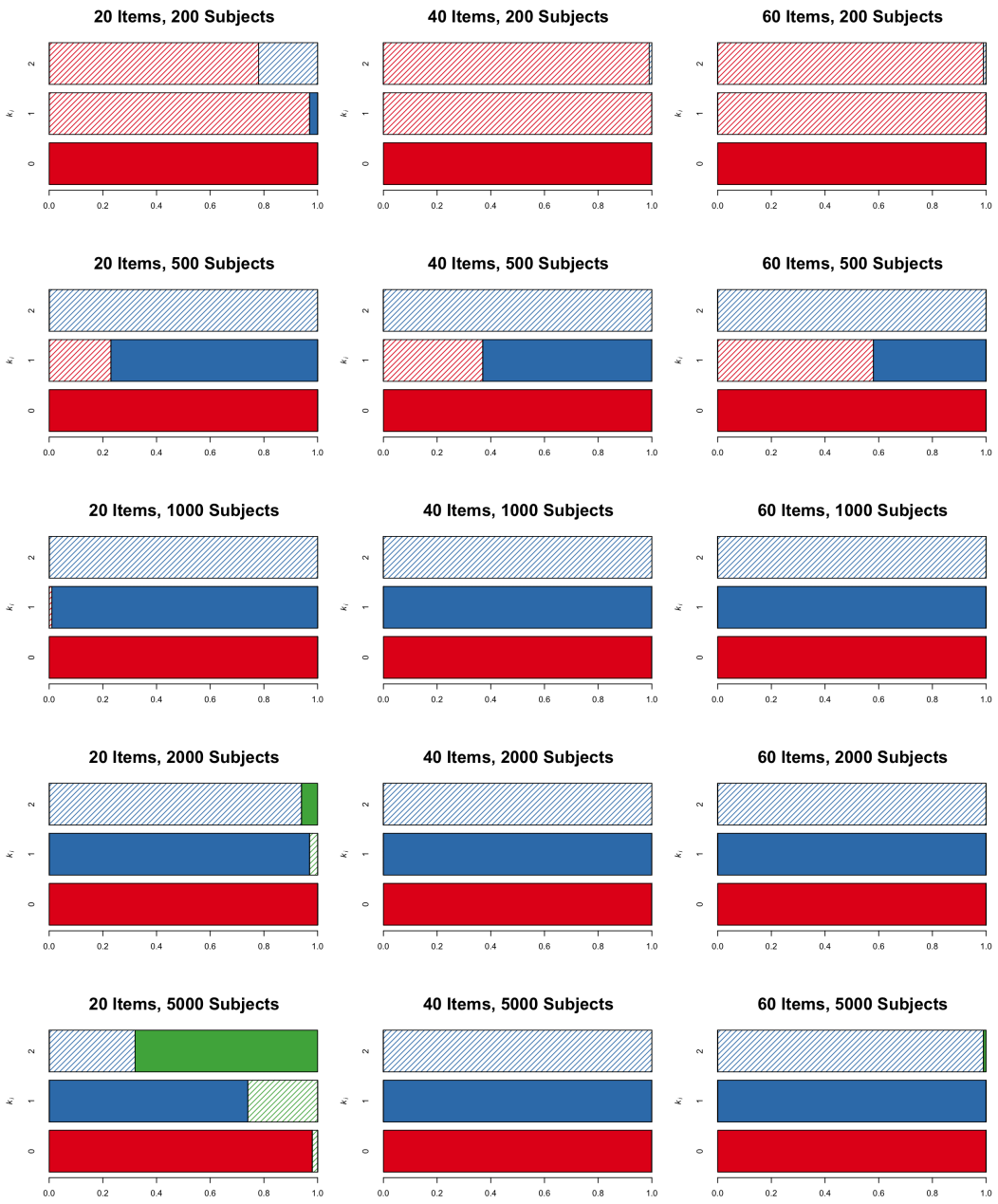

Figure 3.15. Distribution of $\tilde{k}_i$ values selected by the BIC criterion using random-effects estimation. Red indicates $\tilde{k}_i = 0$, blue indicates $\tilde{k}_i = 1$, and green indicates $\tilde{k}_i = 2$. Solid shading indicates that the chosen $\tilde{k}_i$ value matches the data-generating $\tilde{k}_i$ value.

As Figure 3.14 shows, fixed-effects estimation paired with the BIC almost always selects $\tilde{k}_i = 0$ when $k_i = 0$. When $k_i = 1$ or $k_i = 2$, $\tilde{k}_i = 1$ is selected more frequently as sample size increases, with $\tilde{k} = 0$ being selected a majority of times in the smallest sample sizes. In contrast, $\tilde{k}_i = 2$ is selected only rarely, and most often in large samples and short tests. In general, these results suggest that BIC model selection may be preferable to AIC model selection for fixed-effects estimation because, although both criteria choose the correct model with approximately the same frequency, the BIC rarely selects an overly complex model. Similar results are found for the BIC criterion paired with random-effects estimation (see Figure 3.15). Recall that BIC paired with random-effects estimation has a higher proportion of cases in which $\tilde{k}_i = k_i$ than do the other methods. Figure 3.15 reveals that this result occurs because the BIC almost always selects the correct value in the $k_i = 0$ and $k_i = 1$ conditions, but only correctly selects $\tilde{k}_i = 2$ in 3% of replications. In samples of size 1,000 and larger, $\tilde{k}_i = 0$ continues to be selected when $k_i = 0$ in nearly all cases, and $\tilde{k}_i = 1$ is often selected when $k_i = 1$. In contrast, when $k_i = 2$, $\tilde{k}_i = 0$ tends to be selected in the smallest sample, and $\tilde{k}_i = 1$ tends to be selected in the larger samples. In the shortest test length and largest sample size condition, however, $\tilde{k}_i = 2$ is selected a sizable number of times when $k_i = 2$. In this case, $\tilde{k}_i = 2$ is also incorrectly selected when $k_i = 1$ a small proportion of times. In aggregate, random-effects estimation with BIC model selection is tentatively recommended because this method has the highest number of cases in which $\tilde{k}_i = k_i$ and because overly complex models are rarely selected. However, as stated earlier, it is not necessarily desirable to have highly accurate model selection criteria in terms of selecting the data-generating model.

Specifically, it could be that other recovery measures indicate better recovery for a non-data-generating $\tilde{k}_i$ values. Next, I evaluate the performance of the AIC and BIC model selection criteria in terms of IRF recovery as indexed by RIMSE_i.

Each of the studied model selection methods fails to choose the data-generating k_i values in some proportion of cases. Even if $\tilde{k}_i$ values that are selected by the AIC or BIC criterion differ from the data-generating k_i values, it is not necessarily the case that important quantities are misestimated. Instead of comparing the data-generated k_i value to the AIC- or BIC-selected $\tilde{k}_i$ values, it may be more informative to compare the obtained RIMSE_i values associated with the AIC- and BIC-selected $\tilde{k}_i$ values. Figures 3.16, 3.17, and 3.18 display box plots of the RIMSE_i values for the AIC- and BIC-selected $\tilde{k}_i$ values for fixed-effects and random-effects methods. These figures also include box plots of the distribution of RIMSE_i values for the kernel smoothing estimation method.

Figure 3.16 shows results for $k_i = 0$, Figure 3.17 shows results for $k_i = 1$, and 3.18 shows results for $k_i = 2$. In many cases, there are few differences among estimation methods and model selection criteria. The most prominent trend is that RIMSE_i values decrease in mean and variability as sample size increases. Further, kernel smoothing leads to higher average RIMSE_i values than fixed-effects or random-effects estimation when $k_i = 0$, especially in large samples. As noted in the earlier discussion of Tables 3.1, 3.2, and 3.3, this finding may be due to the fact that kernel smoothing estimation approximates IRF conditional on crude surrogate θ values. Specifically, in large samples, conditioning on normalized ranked sum scores may not represent the latent trait metric as well as the methods used in fixed-effects or random-effects FMP estimation. In contrast, in small

samples, kernel smoothing leads to slightly smaller RIMSE_i values on average for the $k_i = 2$ conditions. Thus, kernel smoothing may be a preferable estimation method for estimating complex IRFs in small samples, but FMP methods tend to perform as well or better than kernel smoothing in the vast majority of conditions.

In all conditions, RIMSE_i values decrease gradually as sample size increases. For most purposes, samples of $N = 200$ appear to be too small to lead to reliable IRF recovery, as many RIMSE_i values are very large ($\text{RIMSE}_i \geq .10$). For some applications, $N = 500$ may be a sufficiently large sample. In general, however, samples of at least $N = 1,000$ are needed so that RIMSE_i values are, on average, reasonably small (e.g., $\text{RIMSE}_i < .05$).

Overall, if using the AIC or BIC model selection criterion, random-effects FMP is recommended for short tests. However, the differences among methods are so small that in most cases, the choice between fixed-effects and random-effects, and between AIC and BIC, is not likely to affect conclusions. Combining all conditions, average RIMSE_i values equal .0342 for fixed-effects estimation and the AIC, .0375 for random-effects estimation and the AIC, .0349 for fixed-effects estimation and the BIC, and .0339 for random-effects estimation and the BIC. Random-effects estimation paired with the BIC model selection criterion leads to somewhat smaller RIMSE_i values when $k_i = 0$, and performs comparably to other methods when $k_i = 1$ or $k_i = 2$. Moreover, for short tests and large samples, random-effects estimation paired with the AIC criterion leads to lower RIMSE_i values than the other methods. In terms of IRF recovery, there do not appear to be major differences in the AIC or BIC criterion for fixed-effects estimation. Thus, these simulations suggest that random-effects estimation and the BIC lead

to the best IRF recovery on average, but differences among these four methods are small.

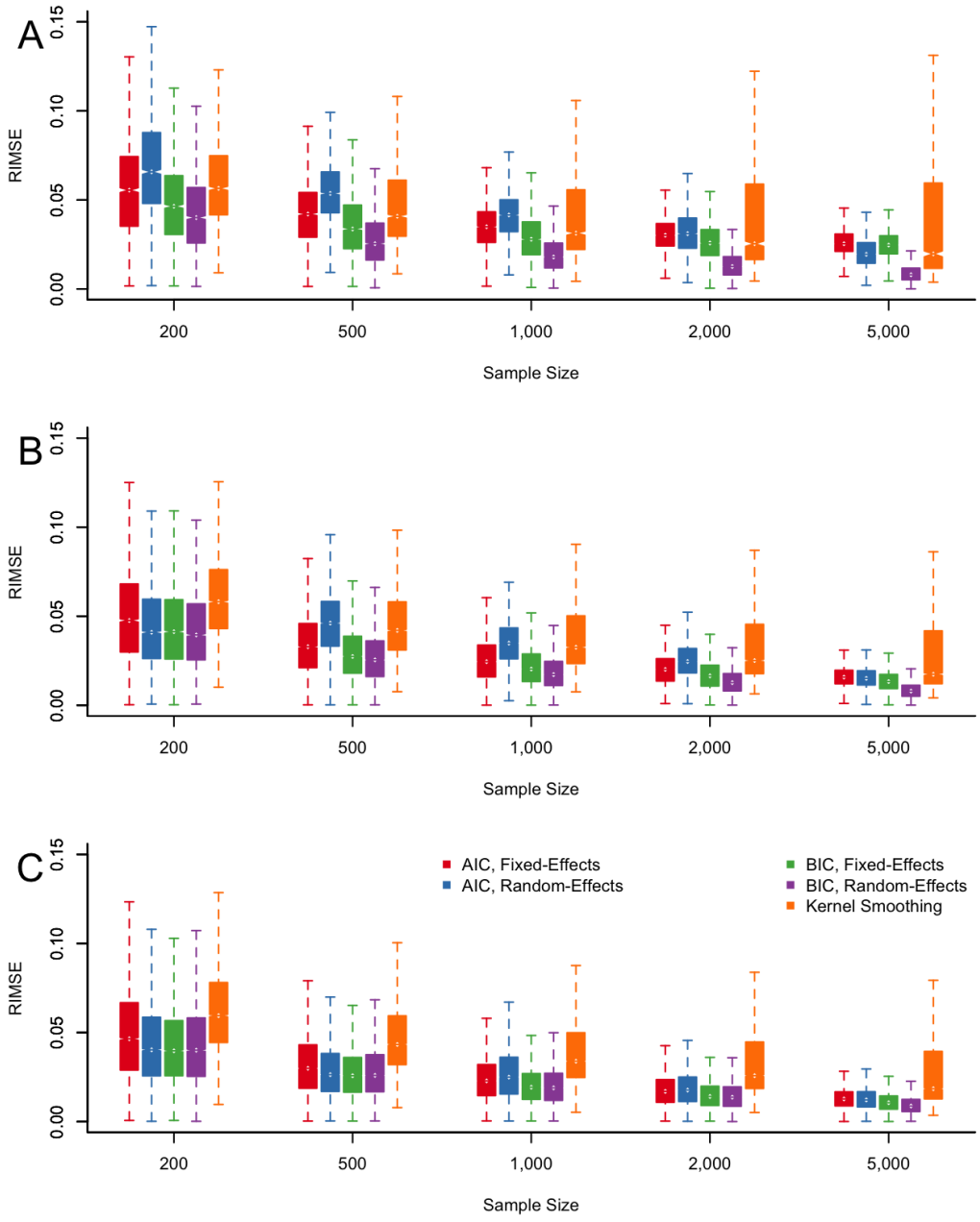

Figure 3.16. Distribution of RIMSE_i values for items generated with $k_i = 0$. Panel A displays results for 20-item tests, Panel B displays results for 40-item tests, and Panel C displays results for 60-item tests.

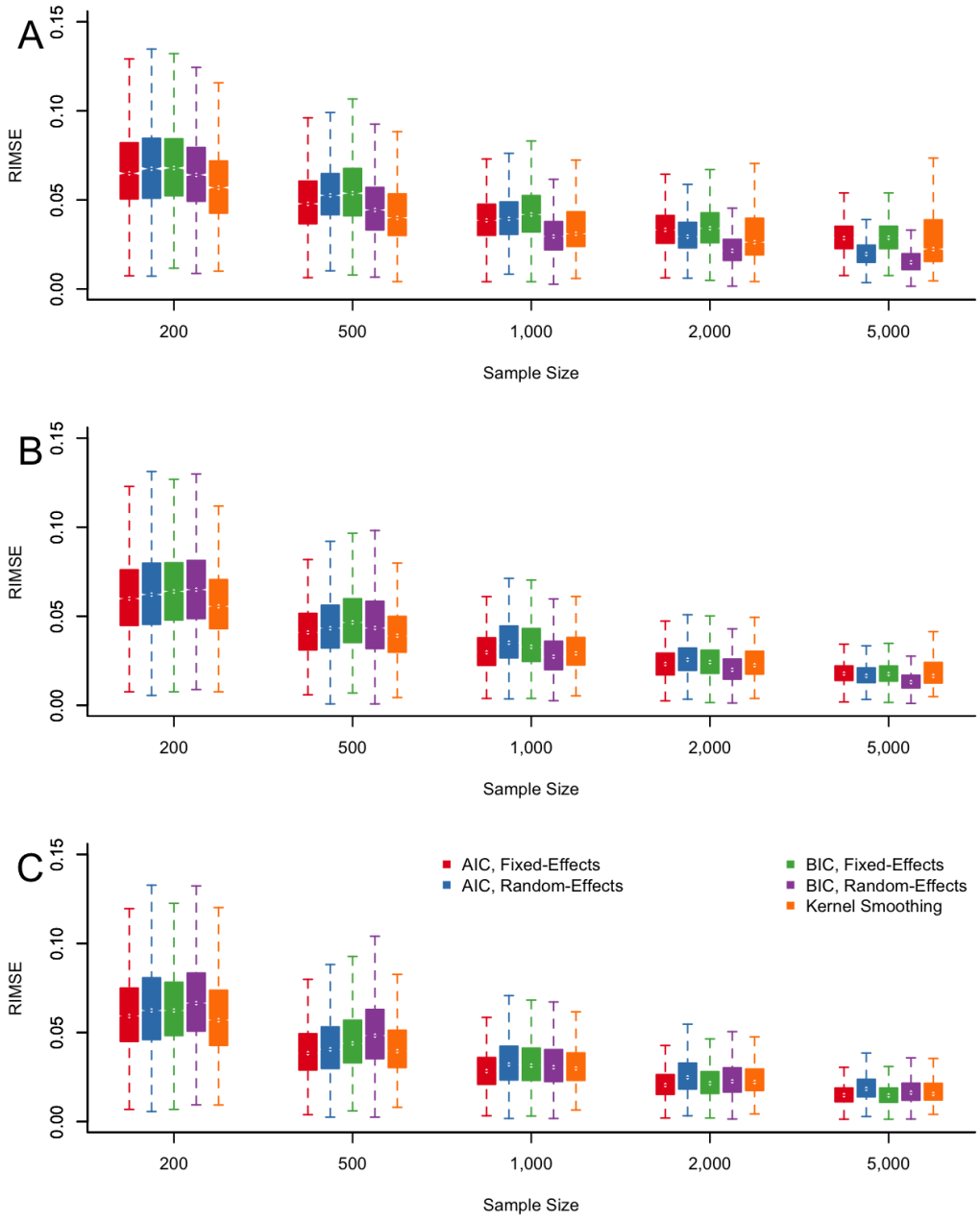

Figure 3.17. Distribution of RIMSE_i values for items generated with $k_i = 1$. Panel A displays results for 20-item tests, Panel B displays results for 40-item tests, and Panel C displays results for 60-item tests.

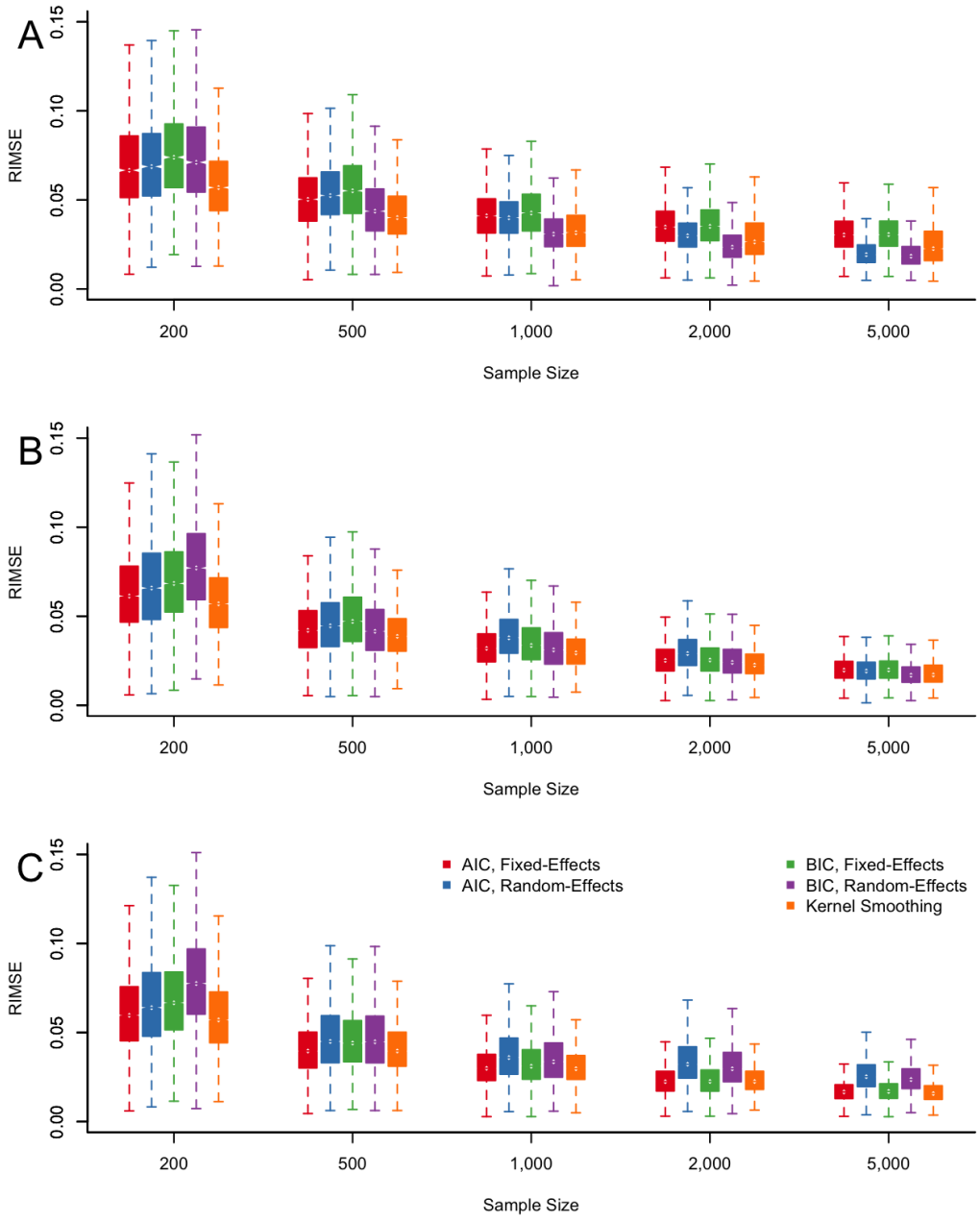

Figure 3.18. Distribution of RIMSE_i values for items generated with $k_i = 2$. Panel A displays results for 20-item tests, Panel B displays results for 40-item tests, and Panel C displays results for 60-item tests.

Chapter 4

Item Parameter Linking

4.1 Item Linking and Model Identification

In a previous section, various methods for identifying the latent trait metric were discussed. For NIRT models, the latent trait is identified by specifying the latent trait distribution (e.g., to be standard normal). For PIRT models, the IRF functional form identifies the shape of the latent trait distribution, but the location and unit size are arbitrary—often, the sample latent trait distribution is considered to be standardized. In both cases, the latent trait metric is sample-dependent to some extent. That is, two samples of examinees that have different latent trait distributions will yield estimated IRFs with different underlying metrics. The technique of item parameter linking aims to reconcile these different metrics by transforming item parameters that have been estimated on one metric to their values on the other metric. In other words, item linking aims to resolve the metric differences that result from differences in model identification restrictions (van

der Linden & Barrett, in press).

The above characterization of item linking differs from how the topic is sometimes discussed. For instance, item linking is often taught in the context of number-correct score equating. IRT-based score equating proceeds in three steps: (a) estimate two sets of item parameters, (b) use IRT-based linking to place the two sets of item parameters on the same θ metric, and (c) transform number-correct scores from one test to be on the same true score metric as scores on the second test (Kolen & Brennan, 2004, p. 172). Although number-correct score equating is an important application, item linking is a more general procedure that transforms one set of item parameter estimates to be on the same metric as another set of item parameter estimates. In other words, IRT-based item linking allows the researcher to alternate between sets of IRT parameters that make equivalent predictions but have been estimated using different identification restrictions.

The vast majority of IRT item linking procedures consider only PIRT models and linear transformations of the θ metric. A linear metric transformation is appropriate when the fitted IRT model determines the latent trait metric up to a linear transformation—as is usually the case for PIRT models—and when the model fits the data. Linear metric transformations are also appropriate under the FMP model. As will be detailed below, transforming the metric linearly does not change FMP item complexities (i.e., k_i values). However, if the FMP item complexities are allowed to vary across the two sets of estimated item parameters, the metric transformation may not be linear. Thus, within the FMP IRT model, both linear and nonlinear transformations of the θ metric can be modeled explicitly.

The linear linking problem is illustrated in Figure 4.1, and the nonlinear linking problem is illustrated in Figure 4.2. For both figures, θ and $\theta^\star$ denote two latent trait metrics that are related by either a linear or nonlinear monotonically increasing function. First consider Figure 4.1. This figure corresponds to the familiar linking problem that is well-described in the literature (e.g., Kolen & Brennan, 2014, pp. 171–245). In this figure, Panel A displays three population IRFs on the θ metric, and Panel B displays the same three IRFs on the $\theta^\star$ metric. In Panels A and B, corresponding IRFs are displayed by the same color (e.g., the red IRF in Panel A corresponds to the red IRF in Panel B). Notice that although all three IRFs appear to be steeper on the $\theta^\star$ metric, the relationships among the IRFs are unchanged. For example, on both metrics the red IRF is less steep than than the green and blue IRFs. In Panel C, the the linear relationship between θ and $\theta^\star$ is displayed by the black line. Finally, Panel D displays a standard normal density on the θ metric and a *transformed* standard normal density on the $\theta^\star$ metric, where the transformation from $\theta^\star$ to θ is defined in Panel C. Notice that because the relationship between θ and $\theta^\star$ is linear, the transformed density in Panel D still follows a normal distribution, albeit one with a non-zero mean and a non-one variance.

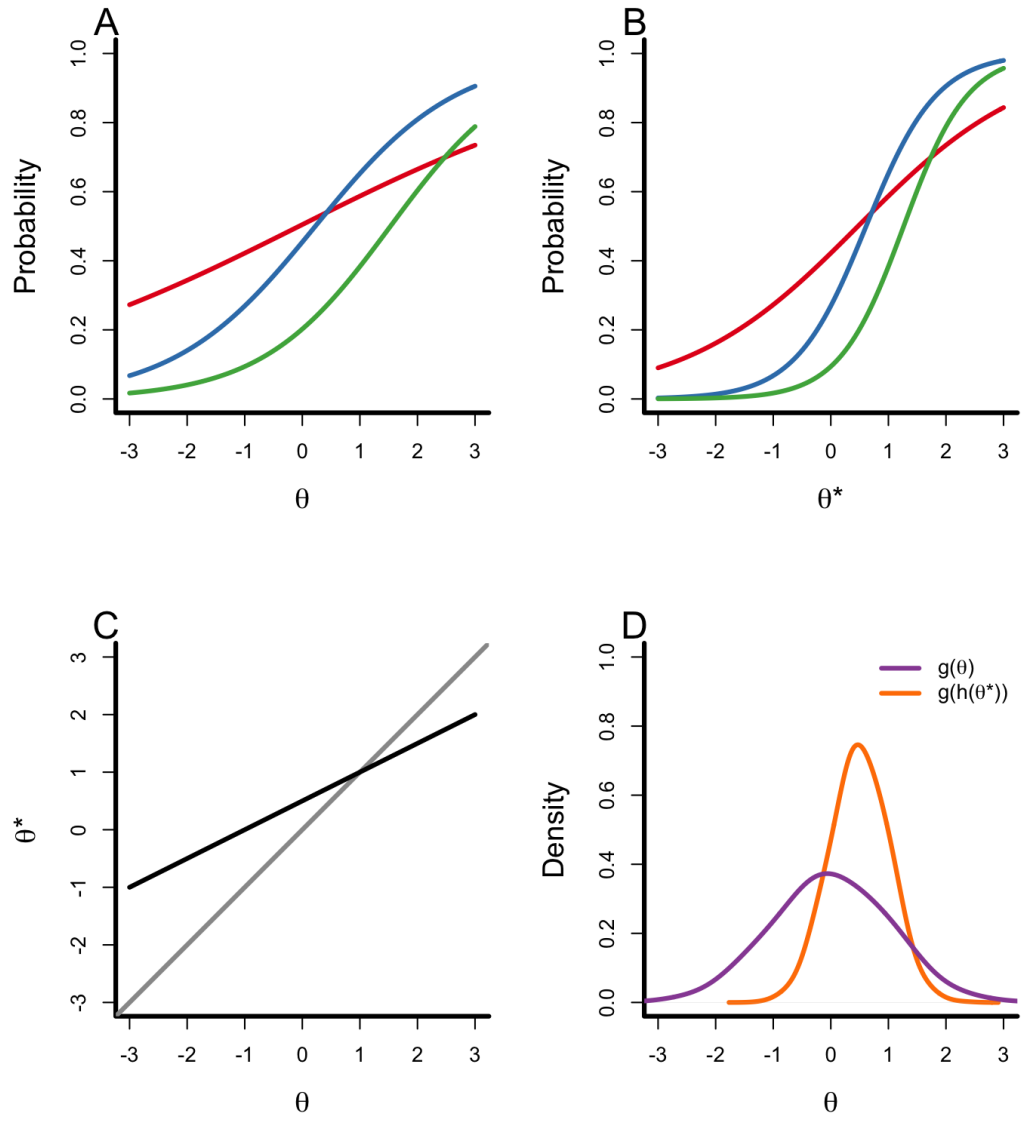

Figure 4.1. Linear item linking illustration. Panel A displays three IRFs on the θ metric and Panel B displays the same three IRFs on the $\theta^{\star}$ metric. In Panels A and B, IRFs with the same color represent the same item. In Panel C, the linear relationship between θ and $\theta^{\star}$ is illustrated by the black line. Finally, Panel D displays a standard normal distribution of θ against a standard normal distribution of $\theta^{\star}$ that has been transformed to the θ metric.

The panels in Figure 4.2 display the same information as the panels in Figure 4.1 with the exception that the relationship between θ and $\theta^\star$ is nonlinear. Although the IRFs in Panel A are instances of the 2PL (i.e., an FMP model with $k_i = 0$), the corresponding IRFs in Panel B take on non-2PL functional forms. The nonlinear relationship between θ and $\theta^\star$ is shown by the black curve in Panel C. In this case, θ is a 3^{rd} degree polynomial function of $\theta^\star$, and the nonlinear relationship between θ and $\theta^\star$ is most prominent at values of θ and $\theta^\star$ that are far away from zero. Finally, Panel D displays a standard normal distribution of θ against a standard normal density of $\theta^\star$ that has been transformed to the θ metric. Notice that the transformed density of $\theta^\star$ is platykurtic and bimodal. As will be detailed later, these characteristics are common when θ is a 3^{rd} degree polynomial function of a normally distributed $\theta^\star$.

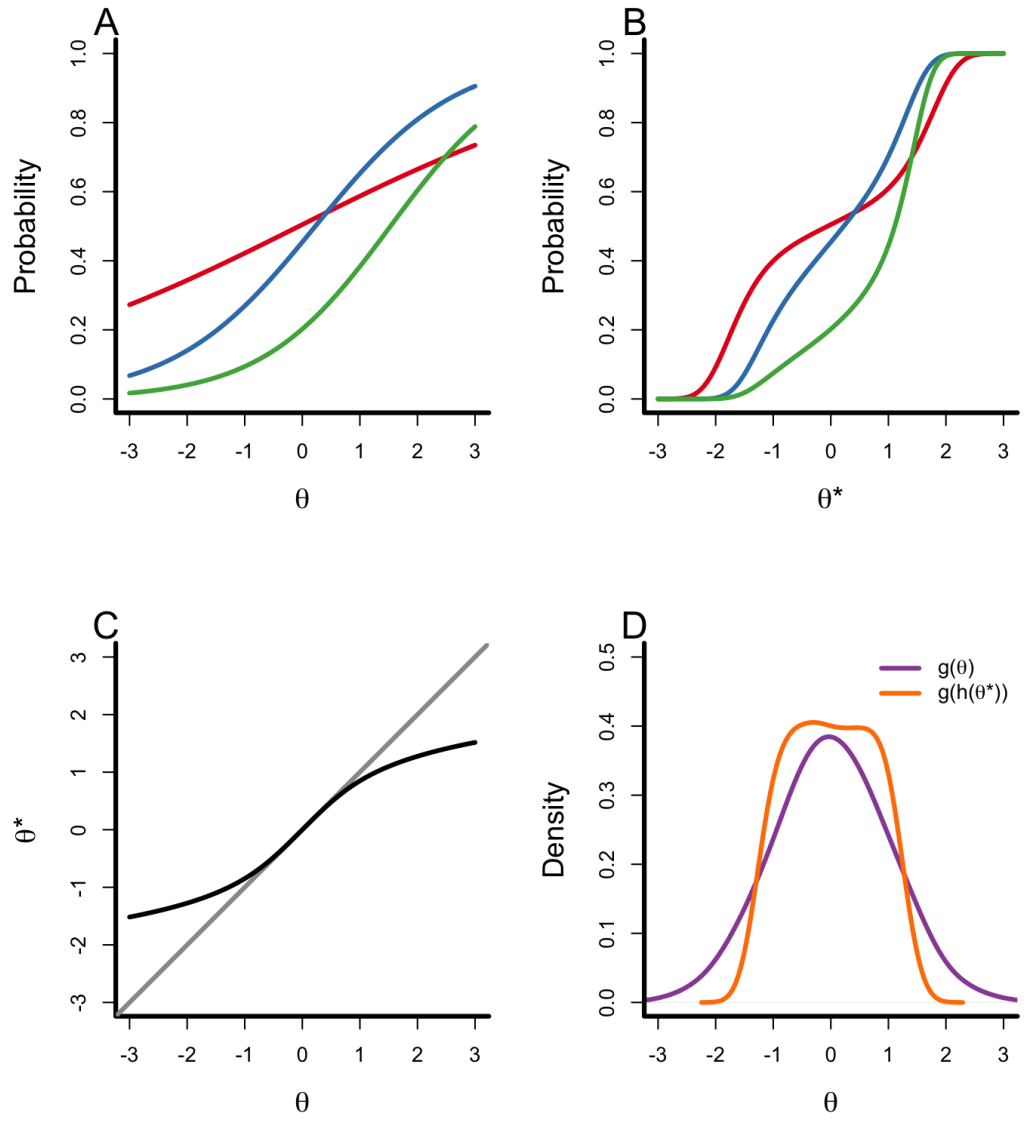

Figure 4.2. Nonlinear item linking illustration. Panel A displays three IRFs on the θ metric and Panel B displays the same three IRFs on the $\theta^{\star}$ metric. In Panels A and B, IRFs with the same color represent the same item. In Panel C, the nonlinear relationship between θ and $\theta^{\star}$ is illustrated by the black curve. Finally, Panel D displays a standard normal distribution of θ against a standard normal distribution of $\theta^{\star}$ that has been transformed to the θ metric. In this example, $\theta = \theta^{\star} + .1667\theta^{\star 3} + .1125\theta^{\star 5}$.

These figures illustrate the linking problem *in the population*. In other words, these figures assume that FMP IRFs are known without error. In fact, the mathematics behind the linking problem is defined in terms of parameters, and it is assumed that estimation errors do not systematically affect the estimated linking transformation. In the following derivations, I explore the mathematical relationships among FMP models that make equivalent predictions. Only after these relationships are known in the population can researchers explore how they operate in sampled data and for estimated item parameters.

4.2 Linear Item Linking with FMP

Before nonlinear linking is described, the equations for linear linking will be derived for the FMP model. Let k_i and $k_i^\star$ denote (population-level) item complexities for the same item on the θ and $\theta^\star$ metrics. If there exists a linear relationship between θ and $\theta^\star$, then $k_i = k_i^\star$. Specifically, let θ denote the latent variable on the original scale, and let $\theta^\star$ denote the latent variable on the transformed scale. If θ is linearly related to $\theta^\star$, the metric transformation can be characterized with two linking parameters, denoted t_0 and t_1. Specifically, let

$$\theta = t_0 + t_1\theta^\star, \tag{4.1}$$

which implies that

$$\theta^\star = t_1^{-1}\theta - t_1^{-1}t_0 \tag{4.2}$$

is the inverse transformation. The aim is to solve for the transformed item parameters $\boldsymbol{b}_i^\star = (b_{0i}^\star, b_{1i}^\star, b_{2i}^\star, b_{2k_i^\star+1,i}^\star)'$ for a typical item i in terms of $\boldsymbol{b}_i$, t_0, and t_1. Finding the transformed item parameters requires solving

$$P(y_{in} = 1|\theta_n, \boldsymbol{b}_i) = P(y_{in} = 1|\theta_n^\star, \boldsymbol{b}_i^\star) \tag{4.3}$$

such that $\boldsymbol{b}_i^\star$ (dimension $(2k_i^\star+2)\times 1$) is a function of t_0, t_1, and $\boldsymbol{b}_i$ (dimension $(2k_i+2)\times 1$) only. That is, the transformed item parameters paired with the transformed metric must yield the same predictions as the original item parameters paired with the original metric. Substituting the FMP IRF given by Equation 2.4 into Equation 4.3,

$$\left[1 + \exp\left(-\sum_{r=0}^{2k_i+1} b_{ri}\theta^r\right)\right]^{-1} = \left[1 + \exp\left(-\sum_{s=0}^{2k_i+1} b_{si}^\star\theta^{\star s}\right)\right]^{-1}, \tag{4.4}$$

which implies that

$$\sum_{r=0}^{2k_i+1} b_{ri}\theta^r = \sum_{s=0}^{2k_i+1} b_{si}^\star\theta^{\star s}. \tag{4.5}$$

It is helpful to express both sides of Equation 4.5 in terms of the same latent variable. By substituting Equation 4.1 into θ^r, we find that

$$\begin{aligned} \theta^r &= (t_0 + t_1\theta^\star)^r \\ &= \sum_{u=0}^{r} \binom{r}{u} (t_1\theta^\star)^{r-u}\, {t_0}^u \end{aligned} \tag{4.6}$$

by an application of the binomial theorem. Substituting Equation 4.6 into Equation 4.5,

$$\sum_{r=0}^{2k_i+1} b_{ri} \sum_{u=0}^{r} \binom{r}{u} (t_1\theta^\star)^{r-u}\, t_0{}^u = \sum_{s=0}^{2k_i+1} b^\star_{si}\theta^{\star s} \tag{4.7}$$

now defines the linking equality.

Notice that both sides of Equation 4.7 are polynomial functions of $\theta^\star$ and that $\boldsymbol{b}^\star_i = (b^\star_{0i}, b^\star_{1i}, \ldots, b^\star_{2k_i+1,i})'$ are the polynomial coefficients for item i. Note also that we can rewrite the left-hand side (LHS) of Equation 4.7 explicitly as a polynomial function of $\theta^\star$. Because both sides of Equation 4.7 can be expressed as polynomial functions of $\theta^\star$, the LHS term that is multiplied by $\theta^{\star s}$ will equal $b^\star_{si}$, where $s = r - u$. For example, let us solve for the highest-order coefficient $b^\star_{2k_i+1,i}$. On the right-hand side of Equation 4.7, $b^\star_{2k_i+1,i}$ is multiplied by $\theta^{\star 2k_i+1}$. On the LHS, $\theta^{\star 2k_i+1}$ only occurs when $r - u = 2k_i + 1$, that is, when $u = 0$ and $r = 2k_i + 1$ (because $r \leq 2k_i + 1$). Substituting in these values of r and u, we find that

$$b^\star_{2k_i+1,i} = t_1^{2k_i+1} b_{2k_i+1,i}. \tag{4.8}$$

In general, note in the LHS side of Equation 4.7, that the quantity $\theta^{\star s}$ occurs only for those sums for which $s = r - u$. Thus, we can express a typical coefficient on the transformed scale, $b^\star_{si}$, as a sum over the index r. Further notice that $u \geq 0$; thus, $r \geq s$. Retaining only those terms that contain $\theta^{\star s}$, a typical coefficient $b^\star_{si}$

on the transformed scale equals

$$b^{\star}_{si} = \sum_{r=s}^{2k_i+1} \binom{r}{r-s} t_1^s t_0^{r-s} b_{ri}. \tag{4.9}$$

It is important to establish that the item parameter transformations given by Equation 4.9 are identified. The t_0 and t_1 parameters are identified with one common item across two calibrations. It is simplest to express this identification in terms of the coefficients $b_{2k_i,i}$, $b_{2k_i+1,i}$, $b^{\star}_{2k_i,i}$, and $b^{\star}_{2k_i,+1i}$. Specifically, by application of Equation 4.9,

$$b^{\star}_{2k_i+1,i} = t_1^{2k_i+1} b_{2k_i+1,i} \tag{4.10}$$

and

$$b^{\star}_{2k_i,i} \;=\; t_1^{2k_i} b_{2k_i,i} + (2k_i+1) t_1^{2k_i} t_0 b_{2k_i+1,i}. \tag{4.11}$$

Thus,

$$t_1 = \left(\frac{b^{\star}_{2k_i+1,i}}{b_{2k_i+1,i}} \right)^{\frac{1}{2k_i+1}} \tag{4.12}$$

and

$$t_0 = \frac{b^{\star}_{2k_i,i} - t_1^{2k_i} b_{2k_i,i}}{(2k_i+1) t_1^{2k_i} b_{2k_i+1,i}} \tag{4.13}$$

is sufficient to establish that the linking transformation is identified.

4.3 Nonlinear Item Linking with FMP

The previous section detailed the linking transformations when IRFs on the θ metric are of the same complexity as IRFs on the $\theta^\star$ metric. In that scenario, $k_i = k_i^\star$ for all items, and we considered linear relationships between θ and $\theta^\star$. If instead $k_i \neq k_i^\star$, we can consider nonlinear relationships between θ and $\theta^\star$. Suppose that $\theta^\star$ is a metric on which IRFs are more complex than the corresponding IRFs on the θ metric. In other words, $k_i^\star > k_i$ for all $i = 1, \ldots, I$. Note that in the population, if an IRF on the $\theta^\star$ metric is more complex than the corresponding IRF on the θ metric, then all IRFs on the $\theta^\star$ metric are more complex than all IRFs on the θ metric. Note also that this relationship is a mathematical property of the FMP model, and need not hold when FMP curves are estimated.

In the population, if $k_i^\star > k_i$ for all items, then, θ is a polynomial function of $\theta^\star$. Stated formally,

$$\theta = h(\theta^\star) = \sum_{l=0}^{2k_\theta+1} t_l \theta^{\star l}, \tag{4.14}$$

where k_θ is the complexity parameter for the θ transformation, and $t_0, t_1, \ldots, t_{2k_\theta+1}$ are the associated linking coefficients.

Consider a polynomial function on the θ scale, for which

$$\begin{aligned} m_i(\theta) &= m_i[h(\theta^\star)] \\ &= \sum_{s=0}^{2k_1+1} b_{si}\theta^s. \end{aligned} \tag{4.15}$$

For an item i, let $2k_i + 1$ denote the degree of $m_i(\theta)$ on the θ scale, and let $2k_i^\star + 1$ denote the degree of $m_i^\star(\theta)$ on the corresponding $\theta^\star$ scale. That is, on the original scale,

$$m_i(\theta) = \sum_{s=0}^{2k_i+1} b_{si}\theta^s \tag{4.16}$$

and on the transformed $\theta^\star$ scale,

$$m_i^\star(\theta^\star) = \sum_{s=0}^{2k_i^\star+1} b_{si}^\star\theta^{\star s}. \tag{4.17}$$

To link the $\theta^\star$ scale to the θ scale, it is necessary that $m_i(\theta) = m_i^\star(\theta^\star)$ for all items $i = 1, \ldots, I$. Substituting Equation 4.14 into Equation 4.16, we find that

$$m_i(\theta) = \sum_{s=0}^{2k_i+1} b_{si}\left(\sum_{l=0}^{2k_\theta+1} t_l\theta^{\star l}\right), \tag{4.18}$$

which, expanding sums, implies that $m_i(\theta)$ can be expressed as a polynomial function of $\theta^\star$. Because $m_i(\theta) = m_i^\star(\theta^\star)$, we set

$$\sum_{s=0}^{2k_i+1} b_{si}\left(\sum_{l=0}^{2k_\theta+1} t_l\theta^{\star l}\right) = \sum_{s=0}^{2k_i^\star+1} b_{si}^\star\theta^{\star s}. \tag{4.19}$$

As was the case in the linear linking problem, the solution to the nonlinear linking problem involves expressing each side of Equation 4.19 as a polynomial function of $\theta^\star$. Then, it is possible to set the left-hand side polynomial coefficients equal to the right-hand side polynomial coefficients, and solve for $\boldsymbol{b}_i^\star$ in terms of $\boldsymbol{b}_i$. This implies that the order of the polynomial on the left-hand side of Equation 4.19 must equal the order of the polynomial on the right-hand side of Equation 4.19.

Now, the highest power associated with $\theta^\star$ on the left-hand side of Equation 4.19 equals $(2k_\theta + 1)(2k_i + 1)$. Thus,

$$2k_i^\star + 1 = (2k_\theta + 1)(2k_i + 1) \tag{4.20}$$

implies that

$$\begin{aligned} k_i^\star &= \frac{(2k_\theta + 1)(2k_i + 1) - 1}{2} \\ &= 2k_i k_\theta + k_i + k_\theta. \end{aligned} \tag{4.21}$$

This result implies that the polynomial degree needed on the transformed scale depends on k_θ, the complexity of the latent trait transformation. This means that, for example, if $k_\theta = 1$ and $k_i = 1$, then $k_i^\star = 4$, which corresponds to a 9^{th} degree polynomial. Note however, that it is usually not necessary (or even advisable) to fit FMP curves using polynomials of degree 9 or higher. As will be detailed later, neither linear nor nonlinear linking requires the theoretical relationship shown in Equation 4.21 to hold precisely for estimated curves.

When (population values of) $k_i^\star > k_i \geq 1$, equating two sets of polynomial coefficients involves solving Equation 4.19, which requires expanding sums of multinomial series. A general solution will be offered later, but first, let us consider a useful special case solution that occurs when $k_i = 0$ for all items. In this case, $k_i^\star = k_\theta$ for all items. Further, when $k_\theta > 0$, the scale transformation is nonlinear,

and may be characterized as follows:

$$\theta = \sum_{l=0}^{2k_\theta+1} t_l \theta^{\star l}. \tag{4.22}$$

Thus, the linking equality in Equation 4.19 becomes

$$\begin{aligned} \sum_{s=0}^{2k_i^\star+1} b_{si}^\star \theta^{\star l} &= b_{0i} + b_{1i}\theta \\ &= b_{0i} + b_{1i}\left(\sum_{l=0}^{2k_i^\star+1} t_l \theta^{\star l}\right). \end{aligned} \tag{4.23}$$

Matching coefficients, we find that

$$b_{0i}^\star = b_{0i} + b_{1i} t_0 \tag{4.24}$$

and

$$b_s^\star = b_{1i} t_s, \;\; s = 1, \ldots 2k_i^\star + 1. \tag{4.25}$$

Thus, although the 2PL may be re-expressed as an FMP model of degree $k_i^\star$, the polynomial coefficients (excluding $b_{0_i}^\star$) are multiples of b_{1i}. Because the $t_0, t_1, \ldots, t_{2k_\theta+1}$ linking coefficients do not vary across items, the $\boldsymbol{b}_2^\star, \boldsymbol{b}_3^\star, \ldots, \boldsymbol{b}_{2k^\star+1}^\star$, item parameters are, across items, perfectly correlated with the $\boldsymbol{b}_1^\star$ coefficients. To clarify this result, an example 10-item test is shown in Table 4.1. Here, the b_{0i}, b_{1i}, $b_{0i}^\star$, $b_{1i}^\star$, $b_{2i}^\star$, and $b_{3i}^\star$ coefficients are given for a test with $\theta = .5 + .5\theta^\star + .1\theta^{\star 2} + .1\theta^{\star 3}$. That is, the transformation is characterized by $t_0 = t_1 = .5$, $t_2 = .2$, and $t_3 = .1$. In this

example, $k_i = 0$ for $i = 1, \ldots, 10$. Although setting $k_i = 0$ is a necessary condition to take advantage of the special solution, it is not necessary that k_i equals the same value for all items. That is, the $\boldsymbol{t}$ vector found using the special solution for one item can be used to transform other sets of item parameters with $k_i \geq 1$.

Table 4.1

Example equivalent FMP parameters

Item	b_{0i}	b_{1i}	$b^{\star}_{0i}$	$b^{\star}_{1i}$	$b^{\star}_{2i}$	$b^{\star}_{3i}$
1	-2	1	-1.5	0.5	0.2	0.1
2	-2	2	-1.0	1.0	0.4	0.2
3	-1	1	-0.5	0.5	0.2	0.1
4	-1	2	0.0	1.0	0.4	0.2
5	0	1	0.5	0.5	0.2	0.1
6	0	2	1.0	1.0	0.4	0.2
7	1	1	1.5	0.5	0.2	0.1
8	1	2	2.0	1.0	0.4	0.2
9	2	1	2.5	0.5	0.2	0.1
10	2	2	3.0	1.0	0.4	0.2

Note. $\boldsymbol{t} = (.5, .5, .2, .1)'$, $k_i = 1$, $i = 1, \ldots, 10$

Following the logic for the linear solution, the linking transformation is the result of matching polynomial coefficients for the equality in Equation 4.19. The needed transformations can be expressed in matrix notation by

$$\boldsymbol{b}^{\star}_i = \boldsymbol{W}\boldsymbol{b}_i \tag{4.26}$$

where $\boldsymbol{W}$ is a weight matrix of dimension $(2k_i^\star + 2) \times (2k_i + 2)$. Let

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_{2k_i+2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \vdots & \vdots & \vdots \\ & \mathbf{0} & \vdots & \vdots \\ & & \mathbf{0} & \vdots \end{bmatrix} \tag{4.27}$$

where $\boldsymbol{w}_s$ is of length $(s-1)(2k_\theta + 1) + 1$, $s = 1, \ldots, 2k_i + 2$, and the remaining elements in the s^{th} column of $\boldsymbol{W}$ are set equal to zero. The column vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{2k_i+2}$ may be found recursively. Define a new class of matrices $\boldsymbol{V}^{(s)}$, with $s(2k_\theta + 1) + 1$ rows and $2k_\theta + 2$ columns. This class of matrices has the following form:

$$\boldsymbol{V}^{(s)} = \begin{bmatrix} \boldsymbol{w}_s & 0 & 0 & \cdots & 0 \\ \vdots & \boldsymbol{w}_s & 0 & \cdots & 0 \\ \mathbf{0} & \vdots & \boldsymbol{w}_s & \cdots & 0 \\ & \mathbf{0} & \vdots & & 0 \\ & & \mathbf{0} & & \boldsymbol{w}_s \end{bmatrix}, \tag{4.28}$$

where for column l, $\boldsymbol{w}_s$ begins at element s of the l^{th} row, and all other elements are set equal to zero. Now, set $\boldsymbol{w}_1 = 1$ and $\boldsymbol{w}_2 = (t_0, t_1, \ldots, t_{2k_\theta+1})'$. Then,

$$\boldsymbol{w}_s = \boldsymbol{V}^{(s-1)}\boldsymbol{t} \tag{4.29}$$

where $\boldsymbol{t} = (t_0, t_1, \ldots, t_{2k_\theta+1})'$.

To illustrate, suppose $k_i = 1$ and $k_\theta = 1$. Then,

$$\boldsymbol{b}_i^\star = \boldsymbol{W}\boldsymbol{b}_i \tag{4.30}$$

where

$$\boldsymbol{W} = \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 \\ 0 & t_1 & 2t_0t_1 & 3t_0^2t_1 \\ 0 & t_2 & 2t_0t_2 + t_1^2 & 3t_0t_1^2 + 3t_0^2t_2 \\ 0 & t_3 & 2t_0t_3 + 2t_1t_2 & t_1^3 + 3t_0^2t_3 + 6t_0t_1t_2 \\ 0 & 0 & t_2^2 + 2t_1t_3 & 3t_1^2t_2 + 3t_0t_2^2 + 6t_0t_1t_3 \\ 0 & 0 & 2t_2t_3 & 3t_1t_2^2 + 3t_1^2t_3 + 6t_0t_2t_3 \\ 0 & 0 & t_3^2 & t_2^3 + 3t_0t_3^2 + 6t_1t_2t_3 \\ 0 & 0 & 0 & 3t_2^2t_3 + 3t_1t_3^2 \\ 0 & 0 & 0 & 3t_2t_3^2 \\ 0 & 0 & 0 & t_3^2 \end{bmatrix}. \tag{4.31}$$

The columns $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ have already been defined as $\boldsymbol{w}_1 = 1$ and $\boldsymbol{w}_2 = (t_0, t_1, \ldots, t_{2k_\theta+1})'$. The vector $\boldsymbol{w}_3$ equals

$$\boldsymbol{w}_3 = \boldsymbol{V}^{(2)}\boldsymbol{t} \tag{4.32}$$

where

$$\boldsymbol{V}^{(2)} = \begin{bmatrix} t_0 & 0 & 0 & 0 \\ t_1 & t_0 & 0 & 0 \\ t_2 & t_1 & t_0 & 0 \\ t_3 & t_2 & t_1 & t_0 \\ 0 & t_3 & t_2 & t_1 \\ 0 & 0 & t_3 & t_2 \\ 0 & 0 & 0 & t_3 \end{bmatrix}. \tag{4.33}$$

Thus,

$$\boldsymbol{w}_3 = \begin{bmatrix} t_0 & 0 & 0 & 0 \\ t_1 & t_0 & 0 & 0 \\ t_2 & t_1 & t_0 & 0 \\ t_3 & t_2 & t_1 & t_0 \\ 0 & t_3 & t_2 & t_1 \\ 0 & 0 & t_3 & t_2 \\ 0 & 0 & 0 & t_3 \end{bmatrix} \begin{bmatrix} t_0 \\ t_1 \\ t_2 \\ t_3 \end{bmatrix} = \begin{bmatrix} t_0^2 \\ 2t_0t_1 \\ 2t_0t_2 + t_1^2 \\ 2t_0t_3 + 2t_1t_2 \\ 2t_1t_3 + t_2^2 \\ 2t_2t_3 \\ t_3^2 \end{bmatrix}. \tag{4.34}$$

Th column vector $\boldsymbol{w}_3$ can now be used to find $\boldsymbol{w}_4$. Specifically,

$$\boldsymbol{V}^{(3)} = \begin{bmatrix} t_0^2 & 0 & 0 & 0 \\ 2t_0t_1 & t_0^2 & 0 & 0 \\ 2t_0t_2 + t_1^2 & 2t_0t_1 & t_0^2 & 0 \\ 2t_0t_3 + 2t_1t_2 & 2t_0t_2 + t_1^2 & 2t_0t_1 & t_0^2 \\ 2t_1t_3 + t_2^2 & 2t_0t_3 + 2t_1t_2 & 2t_0t_2 + t_1^2 & 2t_0t_1 \\ 2t_2t_3 & 2t_1t_3 + t_2^2 & 2t_0t_3 + 2t_1t_2 & 2t_0t_2 + t_1^2 \\ t_3^2 & 2t_2t_3 & 2t_1t_3 + t_2^2 & 2t_0t_3 + 2t_1t_2 \\ 0 & t_3^2 & 2t_2t_3 & 2t_1t_3 + t_2^2 \\ 0 & 0 & t_3^2 & 2t_2t_3 \\ 0 & 0 & 0 & t_3^2 \end{bmatrix}, \qquad (4.35)$$

and

$$
\begin{aligned}
\boldsymbol{w}_4 &= \boldsymbol{V}^{(3)}\boldsymbol{t} \\
&= \begin{bmatrix}
t_0^2 & 0 & 0 & 0 \\
2t_0t_1 & t_0^2 & 0 & 0 \\
2t_0t_2+t_1^2 & 2t_0t_1 & t_0^2 & 0 \\
2t_0t_3+2t_1t_2 & 2t_0t_2+t_1^2 & 2t_0t_1 & t_0^2 \\
2t_1t_3+t_2^2 & 2t_0t_3+2t_1t_2 & 2t_0t_2+t_1^2 & 2t_0t_1 \\
2t_2t_3 & 2t_1t_3+t_2^2 & 2t_0t_3+2t_1t_2 & 2t_0t_2+t_1^2 \\
t_3^2 & 2t_2t_3 & 2t_1t_3+t_2^2 & 2t_0t_3+2t_1t_2 \\
0 & t_3^2 & 2t_2t_3 & 2t_1t_3+t_2^2 \\
0 & 0 & t_3^2 & 2t_2t_3 \\
0 & 0 & 0 & t_3^2
\end{bmatrix}
\begin{bmatrix} t_0 \\ t_1 \\ t_2 \\ t_3 \end{bmatrix} \qquad (4.36) \\
&= \begin{bmatrix}
t_0^3 \\
3t_0^2t_1 \\
3t_0t_1^2+3t_0^2t_2 \\
t_1^3+3t_0^2t_3+6t_0t_1t_2 \\
3t_1^2t_2+3t_0t_2^2+6t_0t_1t_3 \\
3t_1t_2^2+3t_1^2t_3+6t_0t_2t_3 \\
t_2^3+3t_0t_3^2+6t_1t_2t_3 \\
3t_2^2t_3+3t_1t_3^2 \\
3t_2t_3^2 \\
t_3^2
\end{bmatrix}. \qquad (4.37)
\end{aligned}
$$

Thus,

$$\boldsymbol{W} = \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 \\ 0 & t_1 & 2t_0t_1 & 3t_0^2t_1 \\ 0 & t_2 & 2t_0t_2 + t_1^2 & 3t_0t_1^2 + 3t_0^2t_2 \\ 0 & t_3 & 2t_0t_3 + 2t_1t_2 & t_1^3 + 3t_0^2t_3 + 6t_0t_1t_2 \\ 0 & 0 & t_2^2 + 2t_1t_3 & 3t_1^2t_2 + 3t_0t_2^2 + 6t_0t_1t_3 \\ 0 & 0 & 2t_2t_3 & 3t_1t_2^2 + 3t_1^2t_3 + 6t_0t_2t_3 \\ 0 & 0 & t_3^2 & t_2^3 + 3t_0t_3^2 + 6t_1t_2t_3 \\ 0 & 0 & 0 & 3t_2^2t_3 + 3t_1t_3^2 \\ 0 & 0 & 0 & 3t_2t_3^2 \\ 0 & 0 & 0 & t_3^2 \end{bmatrix} \tag{4.38}$$

and

$$\boldsymbol{b}_i^\star = \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 \\ 0 & t_1 & 2t_0t_1 & 3t_0^2t_1 \\ 0 & t_2 & 2t_0t_2 + t_1^2 & 3t_0t_1^2 + 3t_0^2t_2 \\ 0 & t_3 & 2t_0t_3 + 2t_1t_2 & t_1^3 + 3t_0^2t_3 + 6t_0t_1t_2 \\ 0 & 0 & t_2^2 + 2t_1t_3 & 3t_1^2t_2 + 3t_0t_2^2 + 6t_0t_1t_3 \\ 0 & 0 & 2t_2t_3 & 3t_1t_2^2 + 3t_1^2t_3 + 6t_0t_2t_3 \\ 0 & 0 & t_3^2 & t_2^3 + 3t_0t_3^2 + 6t_1t_2t_3 \\ 0 & 0 & 0 & 3t_2^2t_3 + 3t_1t_3^2 \\ 0 & 0 & 0 & 3t_2t_3^2 \\ 0 & 0 & 0 & t_3^2 \end{bmatrix} \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix}. \tag{4.39}$$

The $\boldsymbol{W}$ matrices for higher values of k_i and k_θ may be found using Equations 4.26, 4.27, 4.28, and 4.29. Moreover, the elements of $\boldsymbol{W}$ for $k_i = \{0, 1, 2\}$ and $k_\theta = \{0, 1, 2\}$—conditions that are expected to be of the greatest practical benefit—are printed in the Appendix. However, the accuracy of this procedure for finding elements of $\boldsymbol{b}_i^\star$ has been verified by the author for up to $k_i = 10$ and $k_\theta = 10$. Finally, note that these equations reduce to the previously described special solution when $k_i = 0$.

The linking coefficients are identified in the following manner. Notice that $b_{0i}^\star$ is a function of t_0 and $\boldsymbol{b}_i$ only (and of no higher linking coefficients). The first row of $\boldsymbol{W}$ will always contain the elements $(1, t_0, t_0^2, t_0^3, \ldots, t_0^{2k_i+1})$. Thus,

$$b_{0i}^\star = \sum_{s=0}^{2k_i+1} b_{si} t_0^s \tag{4.40}$$

will always hold. Because the $\boldsymbol{b}_i$ coefficients define a monotonic polynomial, the above function is invertible, and so there exists a unique value of t_0 that satisfies the above equation. Next, notice in $\boldsymbol{W}$ that the linking coefficient t_l will occur in rows with indices $\geq (l + 1)$. Further, in row $l + 1$, t_l will not be taken to any power higher than 1. Using these facts, it is possible to solve for the $\boldsymbol{t}$ coefficients iteratively. Thus, the linking transformation is identified.

The $\boldsymbol{W}$ matrix may also be used to find the $\boldsymbol{b}_i$ matrix if the vectors $\boldsymbol{b}_i^\star$ and $\boldsymbol{t}$ are known. Specifically,

$$\boldsymbol{b}_i = (\boldsymbol{W}'\boldsymbol{W})^{-1} \boldsymbol{W}'\boldsymbol{b}_i^\star. \tag{4.41}$$

For $k_\theta \in \{0, 1, 2\}$ and $k_i \in \{01, 2\}$ (i.e., the matrices presented in the appendix),

$\boldsymbol{W}$ has full column rank. This was established by symbolically (i.e., for any $\boldsymbol{t}$) finding the reduced row echelon form of $\boldsymbol{W}$ using Mathematica (Wolfram Research Inc., 2015). In the reduced row echelon form of $\boldsymbol{W}$, the number of nonzero rows equals the number of columns, suggesting that $\boldsymbol{W}$ is full column rank for any nontrivial $\boldsymbol{t}$. This appears to be the case generally for any $\boldsymbol{t}$ vector regardless of whether $\boldsymbol{t}$ contains the coefficients of a monotonic polynomial. This result implies that, for conformable $\boldsymbol{W}$ and $\boldsymbol{b}_i^\star$ matrices, the matrix $(\boldsymbol{W}'\boldsymbol{W})$ is invertible and that $\boldsymbol{b}_i$ may be found uniquely using Equation 4.41.

4.4 Implementation

There are a variety of experimental designs for which linking is appropriate (see Kolen & Brennan, 2014, pp. 182–183). In this paper, it is assumed that two sets of estimated item parameters are obtained using two examinee groups that have different latent trait distributions. Further suppose that the two examinee groups respond to the same subset of items (which could be the entire test). In this scenario, we can transform the estimated item parameters computed from one group of examinees to be on the same metric as the item parameters from the other group of examinees.[1] Two linking methods, the Haebara method (Haebara, 1980) and the Stocking-Lord method (Stocking & Lord, 1983) will be considered here. Both of these methods estimate the linking coefficients by comparing the

[1] In this development, I assume the absence of item bias or differential item functioning (DIF; Raju, van der Linden, & Fleer, 1995) between the two groups. When linking item parameters, the presence of DIF may affect the accuracy of the estimated linking transformation (Candell & Drasgow, 1988). If DIF is suspected for some items, the metric transformation parameters can be estimated by applying the following methods to a subset of items that are assumed to not exhibit DIF.

estimated response functions from the two calibrations of the anchor test. In simpler IRT models such as the 2PL and 3PL, the Haebara and Stocking-Lord methods have consistently been found to produce more stable estimates than the computationally simpler mean/sigma (Marco, 1977) and mean/mean (Loyd & Hoover, 1980) methods (e.g., Baker & Al-Karni, 1991; Hanson & Béguin, 2002; S.-H. Kim & Cohen, 1992). Moreover, both the Haebara and Stocking-Lord methods have been found to work well in real data, and there is no clear advantage of one over the other (Kim & Kolen, 2007).

In the Haebara approach, the $\boldsymbol{t}$ coefficients are found by minimizing the following integral over some trait distribution, $g(\theta^{\star})$, where

$$HB_c = \int \sum_{i=1}^{I} \left[P(\boldsymbol{y}_i|\theta^{\star}\hat{\boldsymbol{b}}_i^{\star}) - P(\boldsymbol{y}_i|\theta^{\star}, \boldsymbol{W}\hat{\boldsymbol{b}}_i) \right]^2 g(\theta^{\star}) d\theta^{\star}, \tag{4.42}$$

$\boldsymbol{y}_i$ is the vector of N binary item responses to item i, and $\boldsymbol{t}$ factors into Equation 4.42 via the $\boldsymbol{W}$ matrix (see Equations 4.27–4.29). This approach minimizes the overall sum of squared differences between individual IRFs. The Stocking-Lord approach is a minor modification of the Haebara approach wherein the sum of squared differences between test response functions (i.e., the sum of item response functions) is minimized. In the Stocking-Lord approach, the linking coefficients $\boldsymbol{t}$ are found by minimizing

$$SL_c = \int \left[\sum_{i=1}^{I} P(\boldsymbol{y}_i|\theta^{\star}, \hat{\boldsymbol{b}}_i^{\star}) - \sum_{i=1}^{I} P(\boldsymbol{y}_i|\theta^{\star}, \boldsymbol{W}\hat{\boldsymbol{b}}_i) \right]^2 g(\theta^{\star}) d\theta^{\star}. \tag{4.43}$$

Note that neither Equation 4.42 nor Equation 4.43 requires that the item complexities need to be of the same degree after they are transformed to the same

scale. This is because Equations 4.42 and 4.43 compare differences in the model-predicted probabilities. Thus, although in the population, the k_θ value is determined by k_i and $k_i^\star$, scales can be linearly or nonlinearly linked for any values of k_i and $k_i^\star$. Put another way, the researcher can nonlinearly link scales using any desired k_θ value.

To implement these item linking procedures, it is necessary to ensure that the linking parameters $\boldsymbol{t}$ define a monotonic polynomial. When $k_\theta = 0$ (linear linking), this transformation is monotonically increasing if and only if $t_1 > 0$. When $k_\theta > 0$ (nonlinear linking), monotonicity of the θ transformation can be ensured by using the same parameter transformations described in a previous section to ensure monotonicity of the item parameters $\boldsymbol{b}_i$. First, recall that

$$h(\theta^\star) = \sum_{l=0}^{2k_\theta+1} t_l \theta^{\star l} \tag{4.44}$$

defines the linking transformation. Following the same derivation used earlier, $h(\theta^\star)$ is monotone if and only if its first derivative is strictly nonnegative. Let

$$\frac{\partial h(\theta^\star)}{\partial \theta} = a_{0\theta} + a_{1\theta}\theta^\star + \cdots + a_{2k_\theta+1,\theta}\theta^{\star 2k_\theta+1}, \tag{4.45}$$

$$\xi_\theta = t_0, \tag{4.46}$$

and

$$t_l = \frac{a_{l-1,\theta}}{l} \;\; l = 1, 2, \ldots, 2k_\theta + 1, \tag{4.47}$$

be analogous to Equations 2.8, 2.9, and 2.10. Expressing $\boldsymbol{a}_\theta = (a_{0\theta}, a_{1\theta}, \ldots, a_{2k_\theta,\theta})'$ in matrix notation,

$$\boldsymbol{a}_\theta = \boldsymbol{T}_\theta^{(k_\theta)} \boldsymbol{T}_\theta^{(k_\theta - 1)} \cdots \boldsymbol{T}_\theta^{(2)} \boldsymbol{T}_\theta^{(1)} \exp(\omega_\theta). \tag{4.48}$$

Thus, instead of estimating

$$\boldsymbol{t} = (t_0, t_1, t_2, \ldots t_{2k_\theta+1})' \tag{4.49}$$

directly, we can solve for the elements of $\boldsymbol{\gamma}_\theta$, where

$$\boldsymbol{\gamma}_\theta = (\xi_\theta, \omega_\theta, \alpha_{1\theta}, \tau_{1\theta}, \ldots, \alpha_{2k_\theta+1,\theta}, \tau_{2k_\theta+1,\theta})', \tag{4.50}$$

such that Equation 4.42 or Equation 4.43 is minimized.

Chapter 5

A Composite FMP Model

5.1 Model Specification

The equalities used for nonlinear item linking suggest that, before k_i is specified, there exists a nonlinear indeterminacy in the θ metric associated with the FMP model. Like other model indeterminacies, this indeterminacy is resolved by the model identification method. As stated earlier, IRT models are identified by specifying either the parametric form of the IRF or the latent trait distribution. However, because in IRT the location and interval spacing of the latent trait metric depends on how the model is identified, the method used to resolve this indeterminacy has important consequences for drawing inferences about trait scores. In this section I propose a composite FMP model that explicitly models two tenable, and equally admissible, latent trait metrics that are related to each other by a nonlinear monotonic function.

A composite FMP model can be defined based on the nonlinear linking transformations derived in the previous section, such that

$$\begin{aligned} P(y_{in}|\theta_n^\star, \boldsymbol{b}_i, \boldsymbol{t}) &= H[m_i(\theta)] \\ &= H\{m_i[h(\theta_n^\star)]\}, \end{aligned} \tag{5.1}$$

where

$$\theta = h(\theta^\star) = \sum_{l=0}^{2k_\theta+1} t_l \theta^{\star l} \tag{5.2}$$

defines the linking transformation, and all other quantities are as previously defined. Notice that on the $\theta^\star$ scale, items are of different complexities than the corresponding items on the θ scale. However, because the increased flexibility is introduced via a transformation of the latent trait metric, the added complexity is somewhat artificial. For example, an IRF transformed from $k_i = 1$ to $k_i^\star = 4$ implies transforming a polynomial of degree 3 to a polynomial of degree 9, but is defined by only four linking parameters $\boldsymbol{t} = (t_0, t_1, t_2, t_3)'$. Further, in the same way that $H[m_i(\theta)]$ can approximate any monotonic IRF to arbitrary precision, $h(\theta^\star)$ can approximate any monotonic (i.e., order-preserving) transformation of the latent trait for large enough k_θ values.

The composite FMP model implies two latent trait metrics, θ and $\theta^\star$, and two sets of item parameters, $\boldsymbol{b}$ and $\boldsymbol{b}^\star$. In this model, θ is a related to $\theta^\star$ by a monotone polynomial transformation. Because the polynomial transformation is monotone,

the transformation is invertible. Thus, there is a one-to-one mapping of scores on θ to scores on $\theta^\star$. Suppose that individual n has a score of θ_n on the θ metric and $\theta_n^\star$ on the $\theta^\star$ metric. The probability of a keyed response for individual n is invariant to the choice of metric. By specifying that θ is a polynomial function of $\theta^\star$ (as defined in Equation 5.1 and Equation 5.2), the item complexity of $\boldsymbol{b}_i$ will always be less than the item complexity of $\boldsymbol{b}_i^\star$. In practice, this will always be the case even when item parameters are estimated because $\boldsymbol{b}_i$ and $\boldsymbol{b}_i^\star$ will be estimated simultaneously. To understand this point, note that the composite FMP model includes three test-level vectors of parameters: $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_I)'$, $\boldsymbol{b}^\star = (\boldsymbol{b}_1^\star, \ldots, \boldsymbol{b}_I^\star)'$, and $\boldsymbol{t}$. Fixing the elements of any two of these vectors determines the values of the third vector. When linking item parameter estimates, as described in the previous section, we considered the case in which $\boldsymbol{b}$ and $\boldsymbol{b}^\star$ are estimated from separate groups and $\boldsymbol{t}$ is unknown. In contrast, for the composite FMP model, we simultaneously estimate $\boldsymbol{t}$ and $\boldsymbol{b}$, the values of which determine $\boldsymbol{b}^\star$.

An illustration will be helpful at this junction. Suppose θ is a polynomial function of $\theta^\star$ with $k_\theta = 1$ and coefficient vector $\boldsymbol{t} = (0, .1, .1, .3)'$. The relationship between θ and $\theta^\star$ is shown in Figure 5.1. In this figure, Panel A illustrates the mapping from θ to $\theta^\star$ by the black curve. Panel B illustrates a standard normal distribution of θ and a standard normal distribution of $\theta^\star$ that has been transformed to be on the θ metric. Notice that when transformed to the θ metric, the distribution of $\theta^\star$ is bimodal.

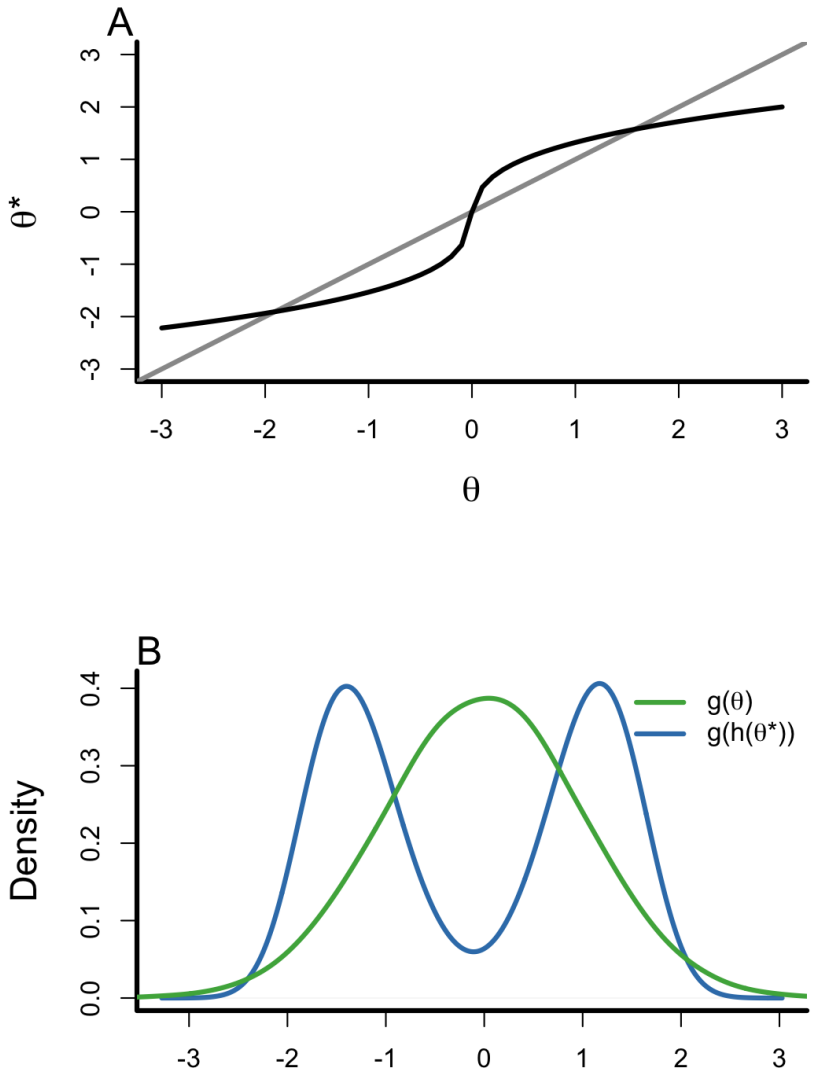

Figure 5.1. Illustration of a polynomial metric transformation where θ is a polynomial function of $\theta^\star$ and $\boldsymbol{t} = (0, .1, .1, .3)'$. In Panel A, the 3^{rd} degree polynomial relationship between θ and $\theta^\star$ is shown by the black curve, and the gray line indicates the identity mapping. In Panel B, a standard normal distribution of θ is displayed against a standard normal distribution of $\theta^\star$ that has been transformed to the θ metric.

A three-item example of the composite FMP model is presented in Figure 5.2 and Table 5.1. In this table and figure, θ and $\theta^\star$ are related by the metric transformation shown in Figure 5.1 (i.e., θ is a polynomial function of $\theta^\star$ with polynomial coefficients $\boldsymbol{t} = (0, .1, .1, .3)'$). Moreover, the corresponding item parameter vectors on the two metrics are related deterministically by Equation 4.26. In this example, all $k_i = 1$ on the θ metric, and all $k_i^\star = 4$ on the $\theta^\star$ metric. The FMP item parameters for this example are reported in Table 5.1. Notice that for these curves, a nonlinear metric transformation alters the shapes and steepnesses of the IRFs. However, because model-predicted probabilities do not change under metric transformations, the relative relationships among the curves do not change. For instance in Figure 5.2, on both the θ and $\theta^\star$ metrics, the red IRFs have equal or higher model-predicted response probabilities than the green IRFs at all trait values.

Table 5.1

Item parameters for composite FMP example

s	b_{s1}	b_{s2}	b_{s3}	$b^{\star}_{s1}$	$b^{\star}_{s2}$	$b^{\star}_{s3}$
0	.019	-.184	-1.371	.019	-.184	-1.371
1	.359	.087	.152	.036	.009	.015
2	.434	.032	.247	.040	.009	.018
3	.183	.039	.163	.117	.027	.051
4				.031	.002	.018
5				.028	.002	.017
6				.043	.004	.025
7				.007	.001	.006
8				.005	.001	.004
9				.005	.001	.004

Note. In Figure 5.2, item 1 is shown in red, item 2 is shown in blue, and item 3 is shown in green.

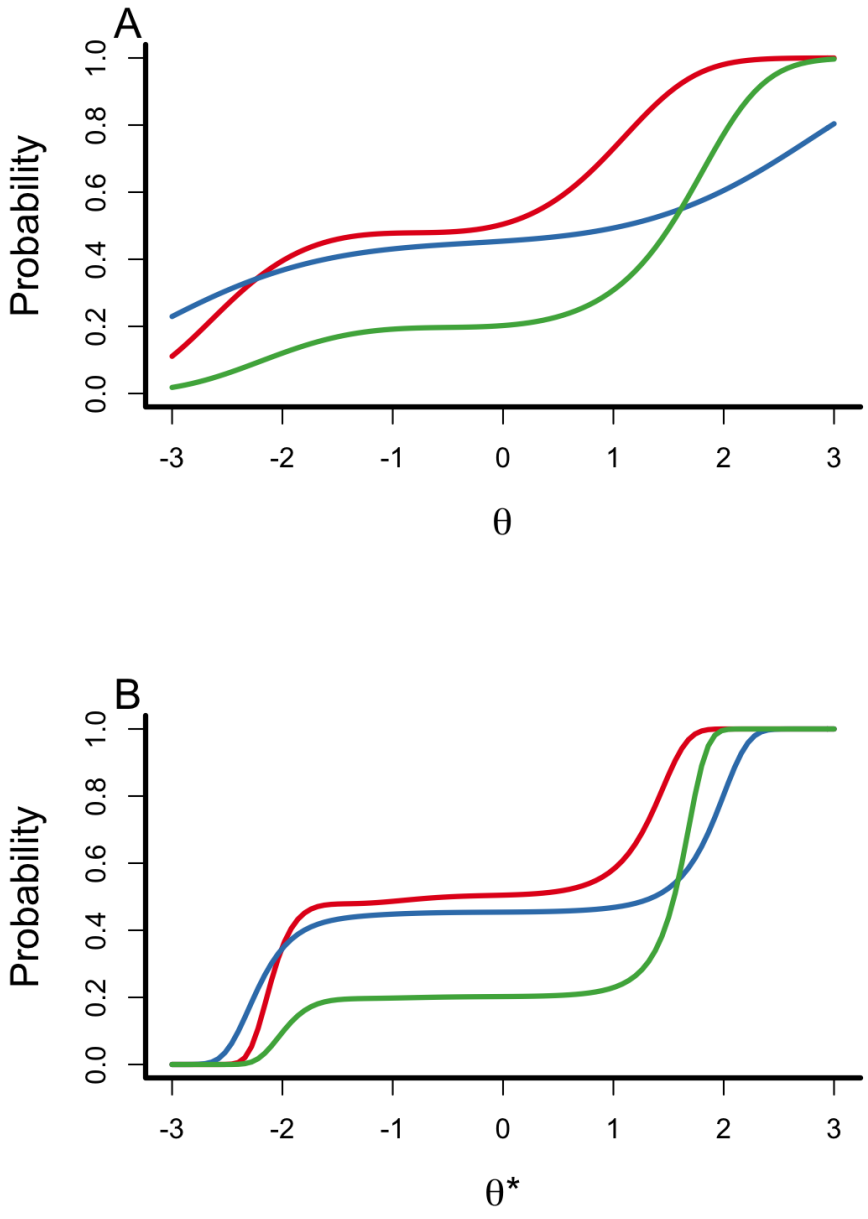

Figure 5.2. An example of three FMP IRFs on the θ metric (Panel A) and on the $\theta^\star$ metric (Panel B). Items 1, 2, and 3 are displayed in red, blue, and green, respectively. The θ metric is a polynomial transformation of $\theta^\star$ with coefficient vector $\boldsymbol{t} = (0, .1, .1, .3)'$. The item parameters for these curves are reported in Table 5.1.

There are two situations in which the composite FMP model will be useful. One scenario—which will be described in detail in a later section—involves transforming the latent trait metric to a metric with more desirable properties. Another scenario—which is described next—is item parameter estimation. Recall that it is commonplace to assume that examinees follow a normal distribution when estimating item parameters. Further recall that, for any univariate continuous latent trait, there always exists a monotonic transformation of the latent trait such that scores are normally distributed (Duncan & MacEachern, 2008). Let $\theta^\star$ be the transformation of the latent trait such that scores are normally distributed. It is possible that there exists a monotonic transformation of the latent trait, denoted θ, such that items may be modeled with smaller item complexities. In this case, fitting an FMP model conditional on a normally distributed $\theta^\star$ could require large k_i values for many items. As an alternative, we could estimate $\boldsymbol{t}$—the polynomial coefficients that determine the transformation from $\theta^\star$ to θ—and a smaller number of item parameters on the θ metric. Because $\boldsymbol{t}$ affects all test items, this use of the composite FMP model may involve estimating a smaller number of parameters than estimating FMP item parameters directly on the $\theta^\star$ metric. That is, a more parsimonious model may be fit by estimating an appropriate metric transformation than by flexibly estimating the IRF shapes. Model parsimony may be particularly desirable in small samples. Simpler models may be preferred in small samples because increasing the number of estimated parameters does not necessarily lead to greater predictive accuracy when sample sizes are small (Browne, 2000; Cudeck & Henly, 1991).

5.2 Fixed-Effects Estimation

When fitting the composite FMP model, we seek to simultaneously estimate the metric transformation coefficients $\boldsymbol{t}$ and the item parameters $\boldsymbol{b}_i$, $i = 1, \ldots, I$. Estimates of $\boldsymbol{b}_i$ and $\boldsymbol{t}$ are sufficient to determine the estimated values of $\boldsymbol{b}_i^\star$ by an application of Equations 4.26–4.29. Recall that in the composite FMP model, $k_i < k_i^\star$ for all items i. As a consequence, any given $\boldsymbol{b}_i$ vector contains fewer estimated coefficients than the corresponding $\boldsymbol{b}_i^\star$ vector. Further, let $\theta^\star$ denote the normally distributed surrogate trait scores that are obtained in the same way as the surrogate trait scores used in the ordinary fixed-effects FMP model (see Liang & Browne, 2015). Because θ is a nonlinear transformation of $\theta^\star$, the composite FMP model is most useful when IRFs with small item complexities are appropriate for a non-normally distributed latent trait.

Recall that the metric transformation is defined by the polynomial coefficients $\boldsymbol{t} = (t_0, t_1, \ldots, t_{2k_\theta+1})'$. It is necessary to ensure that these estimated coefficients define a monotonic transformation of the latent trait. Conveniently, monotonicity can be ensured by estimating $\boldsymbol{\gamma}_\theta$ (see Equation 4.50) instead of estimating $\boldsymbol{t}$ directly. However, when estimating $\boldsymbol{\gamma}_\theta$, there exists a linear indeterminacy in the metric transformation. One way to resolve this indeterminacy is to fix the item parameter values for one item (e.g., the first item). However, it is unclear what parameter values to choose for the first item, particularly if $k_1 > 0$. Another way to resolve this linear indeterminacy is to is to fix $t_0 = 0$ and $t_1 = 1$. That is,

$$t_0 = \xi_\theta = 0 \tag{5.3}$$

and

$$t_1 = \exp(\omega_\theta) = 1. \tag{5.4}$$

Fixing $t_0 = 0$ ensures that scale scores on the θ metric are centered around the same value as the corresponding set of scores on the $\theta^\star$ metric. Moreover, setting both $t_0 = 0$ and $t_1 = 1$ ensures that the estimated metric transformation reduces to an identity mapping as higher-order coefficients approach zero. With these fixed values, estimating $\alpha_{s\theta}$ and $\tau_{s\theta}$, $s = 1, \ldots, k_\theta$, is sufficient to identify the metric transformation.

Recall that we seek to estimate the metric transformation using $\alpha_{s\theta}$ and $\tau_{s\theta}$, $s = 1, \ldots, k_\theta$, and the item parameters on the θ metric. The item parameters may be estimated in the same manner as in the ordinary FMP model. Specifically, the $\boldsymbol{\gamma}_i$ vector (see Equation 2.21) can be estimated instead of $\boldsymbol{b}_i$, $i = 1, \ldots, I$ to ensure that the estimated IRFs are monotonically increasing. Thus, the estimated parameter vector for the composite FMP model equals

$$\boldsymbol{\zeta}^\star = (\alpha_{1\theta}, \tau_{1\theta}, \ldots \alpha_{2k_\theta,\theta}, \tau_{2k_\theta,\theta}, \gamma_1, \gamma_2, \ldots, \gamma_I)'. \tag{5.5}$$

As before, the log likelihood for person n on item i equals

$$F_n(\boldsymbol{\gamma}_i) = y_{in} \ln\{P(\theta_n|\boldsymbol{\gamma}_i)\} + (1 - y_{in}) \ln\{1 - P(\theta_n|\boldsymbol{\gamma}_i)\}. \tag{5.6}$$

To compute $P(\theta_n|\boldsymbol{\gamma}_i)$ in Equation 5.6, it is necessary to transform each surrogate trait score, denoted $\theta_n^\star$ to the corresponding surrogate trait score on the θ metric,

denoted θ_n. Specifically,

$$\theta_n = \sum_{l=0}^{2k_\theta+1} t_l \theta_n^{\star l}. \tag{5.7}$$

Using the transformed surrogate trait scores in Equation 5.6, the complete-data negative log likelihood of the composite FMP model equals

$$F(\boldsymbol{\zeta}^\star) = -\sum_{n=1}^{N}\sum_{i=1}^{I}\left[F_n(\gamma_i)\right]. \tag{5.8}$$

The maximum likelihood estimate of $\boldsymbol{\zeta}^\star$ is obtained by minimizing Equation 5.8. Elements of the estimated parameter vector $\hat{\boldsymbol{\zeta}}^\star$ may then be transformed using Equations 2.9–2.15, 4.26, 4.47, and 4.48 to find the $\hat{\boldsymbol{t}}$, $\hat{\boldsymbol{b}}$, and $\hat{\boldsymbol{b}}^\star$ parameter vectors.

5.3 Random-Effects Estimation

Random-effects estimation of the composite FMP model can be implemented by finding estimates of the same parameter vector as in fixed-effects estimation. Specifically, an estimate of $\boldsymbol{\zeta}^\star$, as defined in Equation 5.5, may be found using an application of the EM algorithm. In the random-effects composite FMP model, it will be convenient to redefine all quantities on the $\theta^\star$ metric.

Suppose that $\theta^\star$ is assumed to follow a standard normal distribution (i.e., the standard normal latent trait prior for the EM algorithm is specified on the $\theta^\star$ metric). Further, let $\boldsymbol{X}^\star = (X_1^\star, X_2^\star, \ldots X_Q^\star)$ denote the quadrature points on the $\theta^\star$ metric, and let $A(X_1^\star), A(X_2^\star), \ldots A(X_Q^\star)$ denote the Q quadrature weights. Recall that the goal of the E step is to obtain estimates of $\bar{n}_q$ and $\bar{r}_{iq}$ based on a

provisional estimate of $\boldsymbol{\zeta}^\star$. Thus, the elements of the provisional $\hat{\boldsymbol{\zeta}}^\star$ vector may be transformed to find provisional estimates of $\boldsymbol{t}$, $\boldsymbol{b}_i$, and $\boldsymbol{b}_i^\star$. This is done as follows. First, as in the fixed-effect model, set $\hat{t}_0 = 0$ and $\hat{t}_1 = \exp(\hat{\omega}_\theta) = 1$ (i.e., $\hat{\omega}_\theta = 0$) to resolve the linear indeterminacy in the estimated metric transformation. Second, transform

$$\hat{\boldsymbol{\gamma}}_\theta = (0, 0, \hat{\alpha}_{1\theta}, \hat{\tau}_{1\theta}, \ldots \hat{\alpha}_{2k_\theta,\theta}, \hat{\tau}_{2k_\theta,\theta})' \tag{5.9}$$

to $\hat{\boldsymbol{t}}$ by an application of Equations 4.46–4.50. This estimated transformation can then be used to find $\hat{\boldsymbol{b}}_i^\star$ by an application of Equations 2.9–2.15 (to find $\hat{\boldsymbol{b}}_i$ from $\hat{\boldsymbol{\gamma}}_i$) and Equation 4.26 (to find $\hat{\boldsymbol{b}}_i^\star$ from $\hat{\boldsymbol{b}}_i$). Next, Equations 2.56 and 2.57 can be redefined such that $\bar{n}_q$ and $\bar{r}_{iq}$ are computed based on the $\theta^\star$ metric. First, define

$$L(\boldsymbol{\zeta}^\star|\boldsymbol{y}_n, X_q^\star) = \prod_{i=1}^{I} L(\boldsymbol{b}_i^\star|y_{in}, X_q^\star), \tag{5.10}$$

where

$$L(\boldsymbol{b}_i^\star|\boldsymbol{y}_{in}, X_q^\star) = P(y_{in}|\boldsymbol{b}_i^\star, X_q)^{y_{in}}[1 - P(y_{in}|\boldsymbol{b}_i^\star, X_q)]^{1-y_{in}}. \tag{5.11}$$

The expected number of persons at quadrature point q, denoted $\bar{n}_q$, equals

$$\bar{n}_q = \sum_{n=1}^{N} \left[\frac{L(\boldsymbol{\zeta}^\star|\boldsymbol{y}_n, X_q^\star)A(X_q^\star)}{\sum_{q=1}^{Q} L(\boldsymbol{\zeta}^\star|\boldsymbol{y}_n, X_q^\star)A(X_q^\star)} \right], \tag{5.12}$$

and the number of correct responses to item i at quadrature point q, denoted $\bar{r}_{iq}$,

equals

$$\bar{r}_{iq} = \sum_{n=1}^{N} \left[\frac{y_{in} L(\boldsymbol{\zeta}^\star | \boldsymbol{y}_n, X_q^\star) A(X_q^\star)}{\sum_{q=1}^{Q} L(\boldsymbol{\zeta}^\star | \boldsymbol{y}_n, X_q^\star) A(X_q^\star)} \right]. \tag{5.13}$$

Thus, performing the E step for the random-effects estimation of the composite FMP model involves finding $\bar{n}_q$ and $\bar{r}_{iq}$, $i = 1, \ldots, I$, $q = 1, \ldots, Q$, as defined in Equations 5.12 and 5.13.

In the M step, the $\hat{\boldsymbol{\zeta}}^\star$ vector is obtained by maximizing

$$\ln L(\boldsymbol{r}, \boldsymbol{n} | \boldsymbol{X}^\star, \boldsymbol{\zeta}^\star) = -\sum_{i=1}^{I} \sum_{q=1}^{Q} \{ \bar{r}_{iq} \ln P(X_q^\star | \boldsymbol{b}_i^\star) + (\bar{n}_q - \bar{r}_{iq}) \ln[1 - P(X_q^\star | \boldsymbol{b}_i^\star)] \}, \tag{5.14}$$

where $\boldsymbol{r} = (\bar{r}_{11}, \bar{r}_{12}, \ldots, \bar{r}_{IQ})'$ and $\boldsymbol{n} = (\bar{n}_1, \ldots, \bar{n}_Q)'$ are computed in the previous E step, and $P(X_q^\star | \boldsymbol{b}_i^\star)$ values are computed using the same parameter transformations that are used to find likelihoods in the E step. As in the ordinary FMP model, the E and M steps are computed iteratively until $\hat{\boldsymbol{\zeta}}^\star$ stabilizes, for example, until no element of $\hat{\boldsymbol{\zeta}}^\star$ changes by more than .0001 across successive iterations. Additionally, as in the fixed-effects method, the final estimate of $\boldsymbol{\zeta}^\star$ can be transformed to give the estimated metric transformation $\hat{\boldsymbol{t}}$ and estimated item parameters on both the θ and $\theta^\star$ metrics.

5.4 Properties of Transformed Scales

5.4.1 Latent trait distribution

In the composite FMP model, θ is specified as a monotonically increasing polynomial function of $\theta^\star$. If the polynomial is of a sufficiently large degree, any monotonic transformation with continuous first derivatives can be approximated with arbitrary precision (Elphinstone, 1983, 1985). In practice, however, it is likely that a polynomial approximation to a monotonic transformation will be of a small degree (e.g., $k_\theta = 1$ or $k_\theta = 2$), largely because it can be difficult to reliably estimate the coefficients of a high-degree polynomial (Murray, Müller, & Turlach, 2013). It is therefore worthwhile to explore the properties of transformed variables when using monotonic polynomial transformations of relatively small degrees.

Assume that θ is a polynomial function of $\theta^\star$ such that

$$\theta = h(\theta^\star) = \sum_{l=0}^{2k_\theta+1} t_l \theta^{\star l}. \tag{5.15}$$

The effects of polynomial transformations on the the distributions of variables have been described in the literature. Specifically, Fleishman (1978) described using 3$^{\text{rd}}$ degree polynomials to transform a normally distributed variable $\theta^\star$ such that the transformed variable θ has user-specified skewness and kurtosis values. Fleishman's procedure, though widely used, is limited in several ways. First, as emphasized by Tadikamalla (1980), a third degree polynomial transformation of a normally distributed variable cannot generate the full range of possible skewness

and kurtosis combinations. Specifically, define

$$\text{skew} = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{5.16}$$

and

$$\text{kurtosis} = \frac{\kappa_4}{\kappa_2^2}, \tag{5.17}$$

where for a variable X,

$$\kappa_2 = E(X^2) - [E(X)]^2, \tag{5.18}$$

$$\kappa_3 = E(X^3) - 3E(X^2)E(X) + 2[E(X)]^3, \tag{5.19}$$

and

$$\kappa_4 = E(X^4) - 4E(X^3)E(X) - 3\left[E(X^2)\right]^2 + 12E(X^2)[E(X)]^2 + 6[E(X)]^4. \tag{5.20}$$

The combinations of skewness and kurtosis values that can be produced by Fleishman's procedure are bounded by the parabola

$$\text{kurtosis} \geq 1.588 \times \text{skew}^2 - 1.139 \tag{5.21}$$

(Fleishman, 1978), meaning that for a given skewness value, there is a lower limit to kurtosis. The lower limit possible using Fleishman's transformation is higher

than the theoretical lower limit on kurtosis for a given skewness value (i.e., for any distribution), which is defined by the parabola

$$\text{kurtosis} \geq \text{skew}^2 - 2 \tag{5.22}$$

(Headrick, 2002). One way to extend the lower limit on kurtosis is to increase the order of the polynomial transformation. Headrick (2002) demonstrated that using polynomial transformations of degree 5 moves the lower bound on kurtosis closer to its theoretical limit and allows for some control over the 5^{th} and 6^{th} moments of the transformed distribution.

One limitation relevant to the current application is that neither the Fleishman nor the Headrick procedure guarantee monotonic transformations. This is because these procedures were developed for the purpose of simulating non-normal data, a situation for which monotonicity is not a major concern. However, when transforming the $\theta^{\star}$ metric under the FMP model, only monotonic polynomial transformations are permissible. Limiting polynomial transformations to monotonic transformations further limits the skewness and kurtosis values that transformed scales can have. The combinations of skewness and kurtosis that are available with 3^{rd} degree and 5^{th} degree polynomial transformations are illustrated in Figure 5.3. In this figure, blue dots indicate monotonic 3^{rd} degree polynomial transformations, green dots indicate non-monotonic 3^{rd} degree polynomial transformations, and purple dots indicate skewness and kurtosis combinations available with 5^{th} degree transformations but not 3^{rd} degree transformations. Finally, the red curve

indicates the theoretical lower bound of kurtosis values. Note that because Headrick's procedure allows some control over 5^{th} and 6^{th} distributional moments, there may be multiple $\boldsymbol{t}$ vectors that give the same skewness and kurtosis values. This fact makes it difficult to determine the skewness and kurtosis values available with 5^{th} degree *monotonic* polynomial transformations, and thus this information is not shown in the figure. For the 3^{rd} degree polynomial transformations, the coefficients found using Fleishman's procedure were checked for monotonicity by determining whether the following conditions were satisfied:

$$t_2^2 - 3t_1t_3 \leq 0 \tag{5.23}$$

and

$$t_3 > 0. \tag{5.24}$$

If the conditions in Equations 5.23 and 5.24 are satisfied, then $\boldsymbol{t}$ defines a monotonically increasing polynomial transformation (Chen & Tung, 2003). In this figure, all of the blue dots occur at kurtosis values greater than zero, which suggests that transforming a normally distributed variable using a monotonic polynomial of degree 3 guarantees a leptokurtic distribution. Further, although 5^{th} degree polynomial transformations extend the bounds implied by Fleishman's procedure, there are still many skewness and kurtosis combinations that are not possible under Headrick's procedure. Thus, even with 5^{th} degree polynomial transformations of the latent trait metric, there are limitations to the distributional properties of the transformed metric.

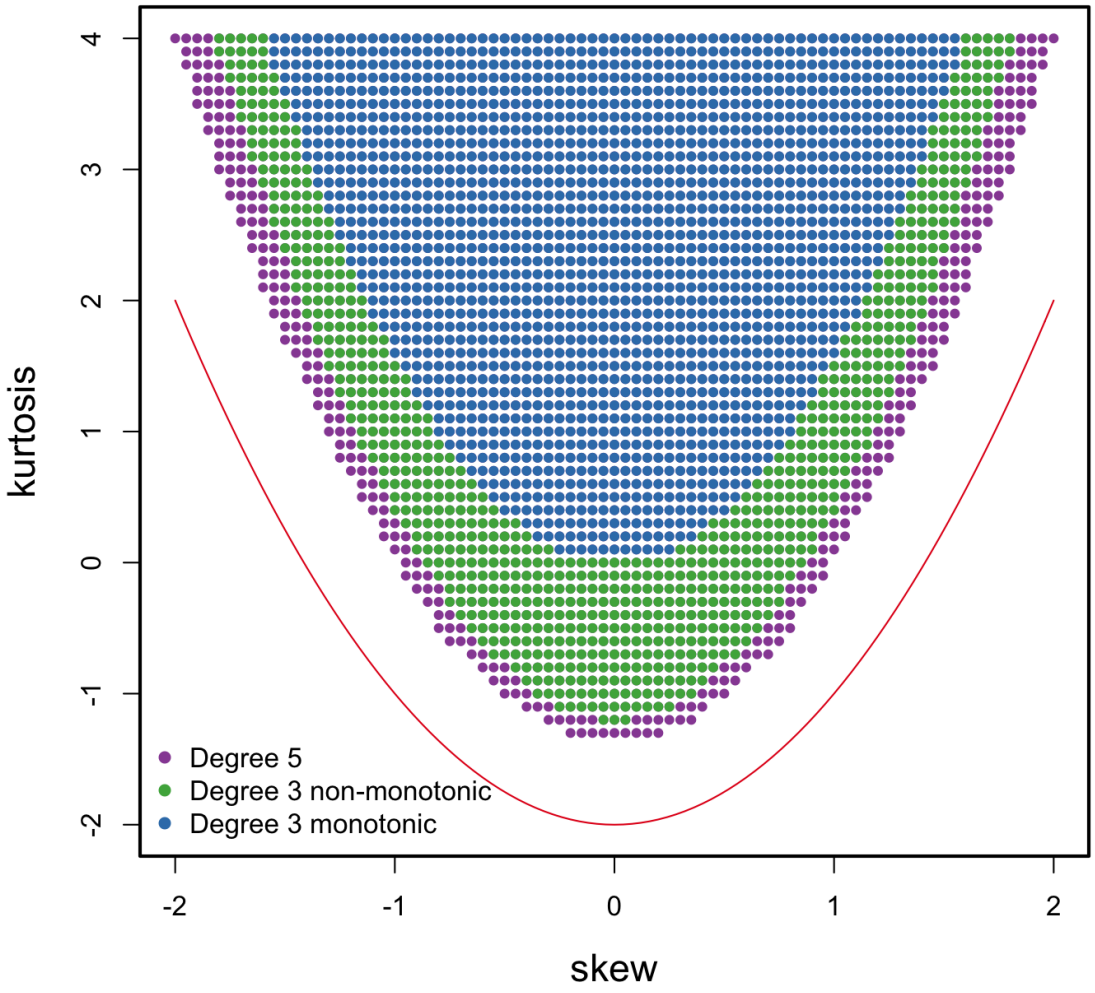

Figure 5.3. Scatter plot of skewness and kurtosis values for 3rd and 5th degree polynomial transformations using Fleishman's and Headrick's methods. Blue dots indicate monotonic (i.e., invertible) 3rd degree polynomial transformations, green dots indicate non-monotonic 3rd degree polynomial transformations. Purple dots occur at skewness and kurtosis pairs that are possible with 5th degree polynomial transformations (i.e., Headrick's method) that cannot occur with 3rd degree polynomial transformations (i.e., Fleishman's method). The red curve defines the theoretical lower bound of skewness and kurtosis values.

Another limitation to the methods proposed by Fleishman and Headrick is that both assume that $\theta^\star$ in Equation 5.15 is the normally distributed variable. We may instead be interested in the distribution of $\theta^\star$ when θ is normally distributed, that is in the properties of *inverse* polynomial transformations away from normality.

The range of skewness and kurtosis values produced by inverse polynomial transformations was explored empirically. Specifically, a sample of 100,000 θ values was randomly generated from a standard normal distribution. Next, a range of monotonic polynomials of degree 3 and degree 5 was generated from a grid of $\boldsymbol{\gamma}_\theta$ parameters (see Equation 4.50) generated uniformly from the ranges $\xi_\theta \in [-1, 1]$, $\omega_\theta \in [-1, 1]$, $\alpha_{1\theta}, \alpha_{2\theta} \in [-.5, .5]$, and $\tau_{1\theta}, \tau_{2\theta} \in [-10, -2]$. Exploration of these parameter ranges suggested that all possible combinations of these parameters were sufficient to explore the range of monotonic polynomial transformations (up to linear transformations, which do not affect skewness and kurtosis). For each polynomial of degree 3 and 5, the 100,000 θ values were transformed to $\theta^\star$ values via an inverse polynomial transformation computed using numerical root-finding methods. The empirical skewness and kurtosis values of the transformed variables were computed using the `fungible` package (Waller & Jones, 2015) in `R` (R Core Team, 2015). Results are displayed in Figure 5.4. In Panel A, $k_\theta = 1$; in Panel B, $k_\theta = 2$. In contrast to the polynomial power transformations proposed by Fleishman (1978) and Headrick (2002), the inverse polynomial power transformation appears to have an *upper* boundary to the amount of kurtosis possible for a given skewness. As with the lower boundaries, this upper boundary appears to be defined by a parabola such that $c_2\text{kurt} - \text{skew}^2 < c_1$ for some constants c_1 and c_2. No attempt is made here to assign exact numbers to c_1 and c_2, and these

values are subject to some error due to the empirical nature of the illustration. However, these graphs suggest that degree 3 inverse polynomial transformations have limited skewness and kurtosis ranges, and that these limits are broadened by the use of degree 5 inverse polynomial transformations.

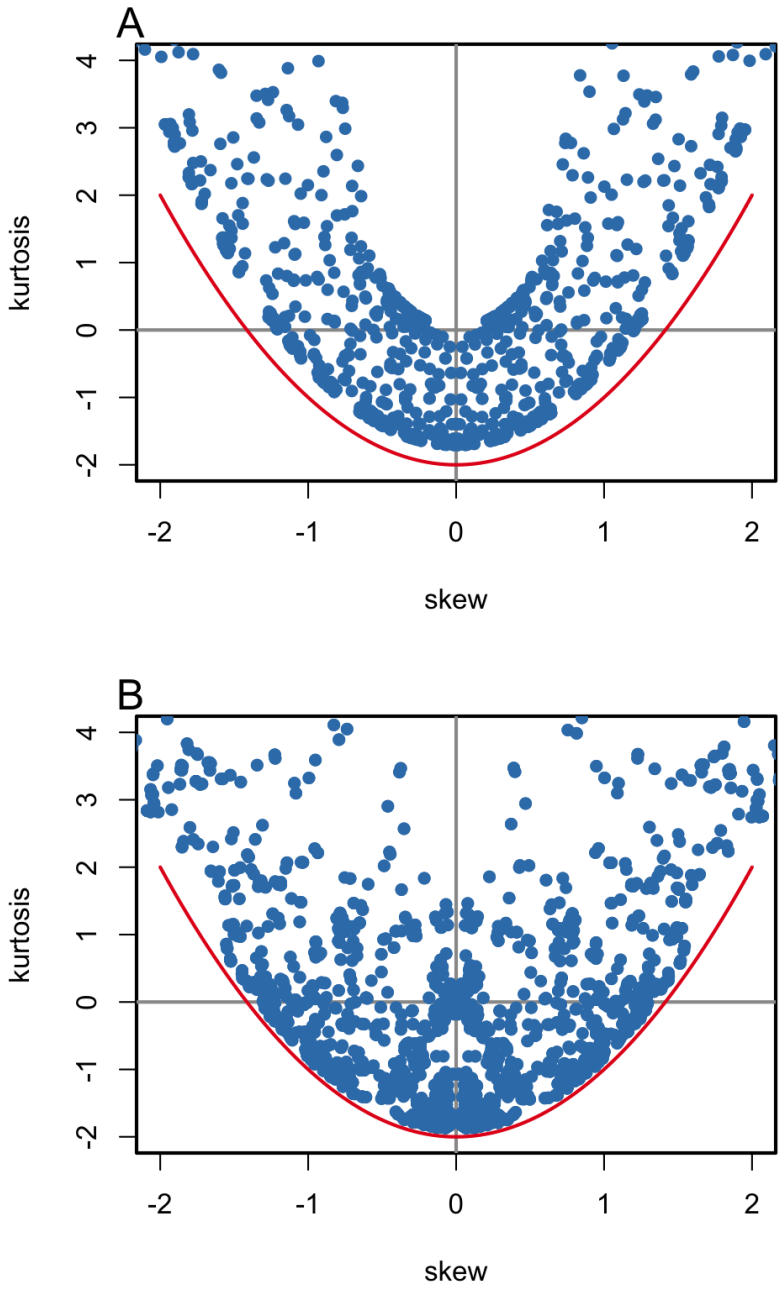

Figure 5.4. Scatter plots of skewness and kurtosis values when transforming normally distributed scored by several inverse polynomial transformations. In Panel A, all $k_\theta = 1$ and in Panel B, all $k_\theta = 2$. The red curves define the theoretical lower bound of skewness and kurtosis values.

5.4.2 Item information

When θ and $\theta^\star$ are related by a nonlinear transformation, the information function on the θ metric need not have the same shape or height as the information function on the $\theta^\star$ metric. In fact, the information function on the θ metric can be distorted to any other continuous univariate information function by some monotonic transformation (Lord, 1980, p. 86), although there will be some limitations when modeling the θ transformation with polynomials of small degrees.

The relationship between the two IIFs was derived by Lord (1974, p. 353) as

$$\mathcal{I}(\theta) = \frac{\mathcal{I}(\theta^\star)}{\left(\frac{\partial\theta}{\partial\theta^\star}\right)^2}. \tag{5.25}$$

Generally speaking, the IIF can take on any desired shape or height for a suitably chosen transformation of the latent trait. Not only this, but the trait level that maximizes $\mathcal{I}(\theta)$ need not be the corresponding trait level that maximizes $\mathcal{I}(\theta^\star)$.

Figure 5.5 displays the IRF and IIF for a single item on both the θ and $\theta^\star$ metrics. In this example, $\boldsymbol{b}_i = (0, 1)'$ and $\boldsymbol{t} = (0, 0.1, 0.1, 0.3)'$. These quantities imply that $k_i^\star = 1$ and $\boldsymbol{b}_i^\star = (0, 0.1, 0.1, 0.3)'$. Panel A displays the IRF on the θ metric, and Panel B displays the IRF on the $\theta^\star$ metric. The vertical lines illustrate the locations of five examinees on each metric. Notice that the relative to the θ metric, the $\theta^\star$ metric appears to be stretched in mid-ranges of the latent trait and squeezed at the extremes. Whereas the examinee scores are evenly spaced on the θ metric, extreme scores on the $\theta^\star$ are closer to each other.

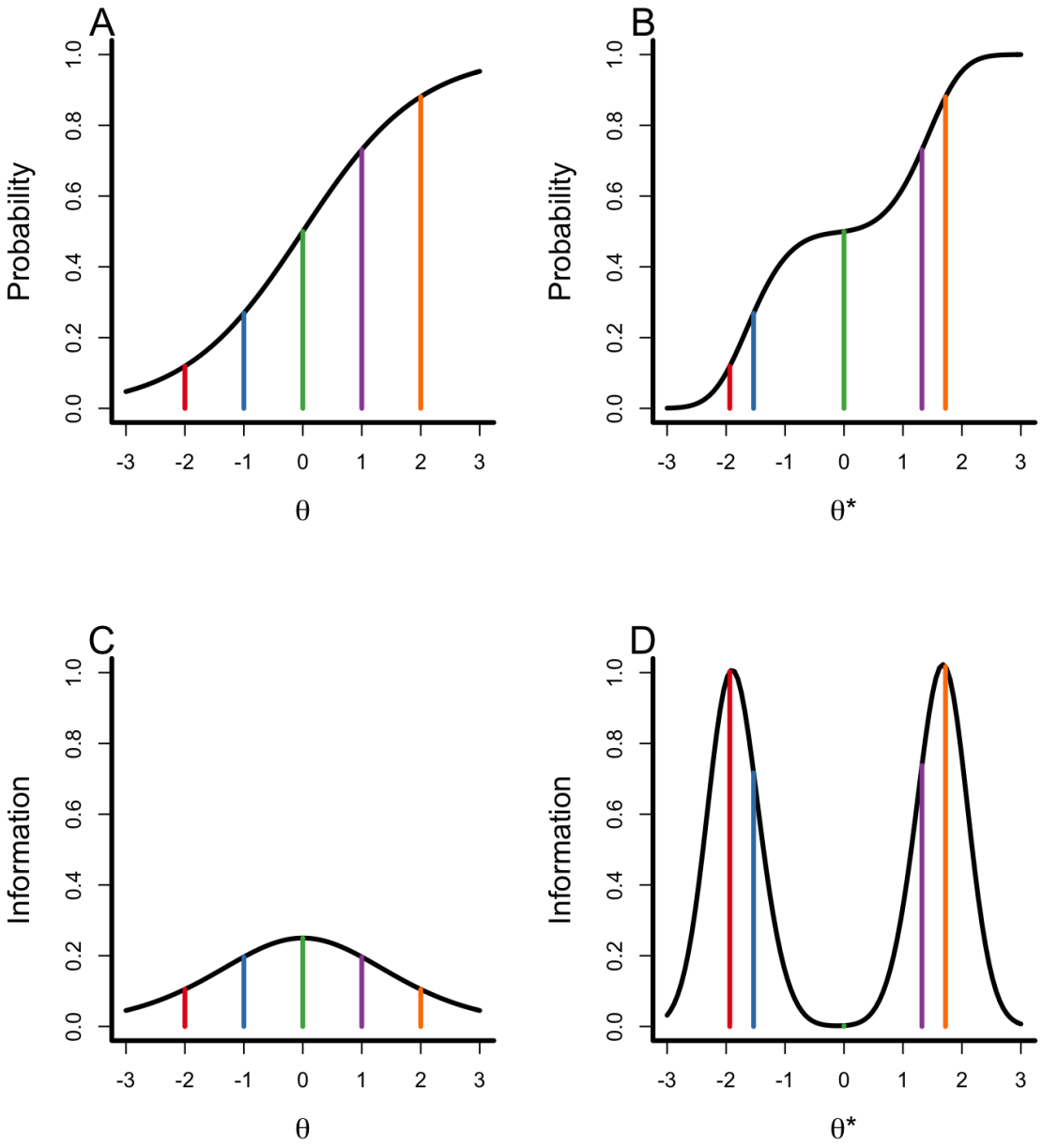

Figure 5.5. Example FMP item IRF and IIF on the θ and $\theta^\star$ metrics. Panels A and B show IRFs, and Panels C and D show IIFs. Panels A and C display functions on the θ metric, and Panels B and D display functions on the $\theta^\star$ metric. In this example, $\boldsymbol{b}_i = (0, 1)'$, $\boldsymbol{b}_i^\star = (0, 0.1, 0.1, 0.3)'$, and $\boldsymbol{t} = (0, 0.1, 0.1, 0.3)'$.

Although information functions can be greatly altered by metric transformations, it is the case that the relative efficiency (RE) of a test remains unchanged under monotonic scale transformations. Let $\theta_n^\star$ denote the latent trait value on the $\theta^\star$ metric that corresponds to the value θ_n on the θ metric (i.e., θ_n is a polynomial transformation of $\theta_n^\star$ using the polynomial coefficients $\boldsymbol{t}$). The relative efficiency, that is the ratio of item information functions conditional on a trait level, is unaffected by the metric transformation:

$$RE = \frac{\mathcal{I}_1^\star(\theta_n^\star)}{\mathcal{I}_2(\theta_n^\star)} = \frac{\mathcal{I}_1(\theta_n)}{\mathcal{I}_2(\theta_n)} \tag{5.26}$$

(Lord, 1974, 1980, p. 89). Put another way, this result implies that the relative information provided by each item is invariant across monotonic transformations of the latent trait. This result is illustrated in Figure 5.6, which displays the IRFs and IIFs for two sets of item parameters on both the θ and $\theta^\star$ metrics. In this figure, Panels A and B display IRFs, and Panels C and D display IIFs. Moreover, Panels A and C display curves on the θ metric, and Panels B and D display curves on the $\theta^\star$ metric. In this example, $\boldsymbol{b}_1 = (0, 1)'$, $\boldsymbol{b}_2 = (0.5, 1.5)'$, $\boldsymbol{b}_1^\star = (0, 0.1, 0.1, 0.3)'$, $\boldsymbol{b}_2^\star = (0.5, 0.15, 0.15, 0.45)'$, and $\boldsymbol{t} = (0, 0.1, 0.1, 0.3)'$. For all panels, the green curves represent item 1 and the blue curves represent item 2. The purple vertical dotted lines are shown at an example location on the latent trait. This value equals -0.5 on the θ metric and -1.2 on the $\theta^\star$ metric. At this example latent trait value, the model-predicted response probabilities do not change: $P_1(-0.5) = P_1^\star(-1.2) = .44$ and $P_2(-0.5) = P_2^\star(-1.2) = .38$. Notice however, that item information functions on the θ metric look very different than the

item information functions on the $\theta^\star$ metric. Specifically, the IIFs are unimodal on the θ metric and bimodal on the $\theta^\star$ metric. The IIFs on the $\theta^\star$ metric also have noticeably higher maximum information values than the IIFs on the θ metric. The amount of information at the example trait level (represented by the vertical purple line) also varies across metrics. Namely, $\mathcal{I}_1(-0.5) = .55$, $\mathcal{I}_1^\star(-1.2) = .75$, $\mathcal{I}_2(-0.5) = .24$, and $\mathcal{I}_2^\star(-1.2) = .32$. Note, however, that the *relative* heights of these curves do not change: $\mathcal{I}_1(-0.5)/\mathcal{I}_2(-0.5) = \mathcal{I}_1^\star(-1.2) = \mathcal{I}_2^\star(-1.2) = 2.4$. These figures illustrate that conditional on $\theta = h(\theta^\star)$, neither the model-predicted probabilities nor the ratio of information functions are affected by the choice of metric transformation. In other words, although the shape of the IIF can be dramatically manipulated by metric transformations, the relative amount of information given by certain items does not change. This result has several implications for IRT-based measurement, particularly in an adaptive context. First, recall that the test information function is the simple sum of item information functions. This implies that the shape of the TIF depends on the choice of metric transformation (Lord, 1980, pp. 85–88). Because metric transformations can alter the shape of the TIF, the standard error of measurement (i.e., the inverse square root of test information) depends on the choice of metric. Thus, terminating a computerized adaptive test (CAT) based on a standard error of measurement (SEM) criterion cutoff value (Weiss, 1982) only guarantees a constant SEM if scale scores are measured directly on the desired metric.

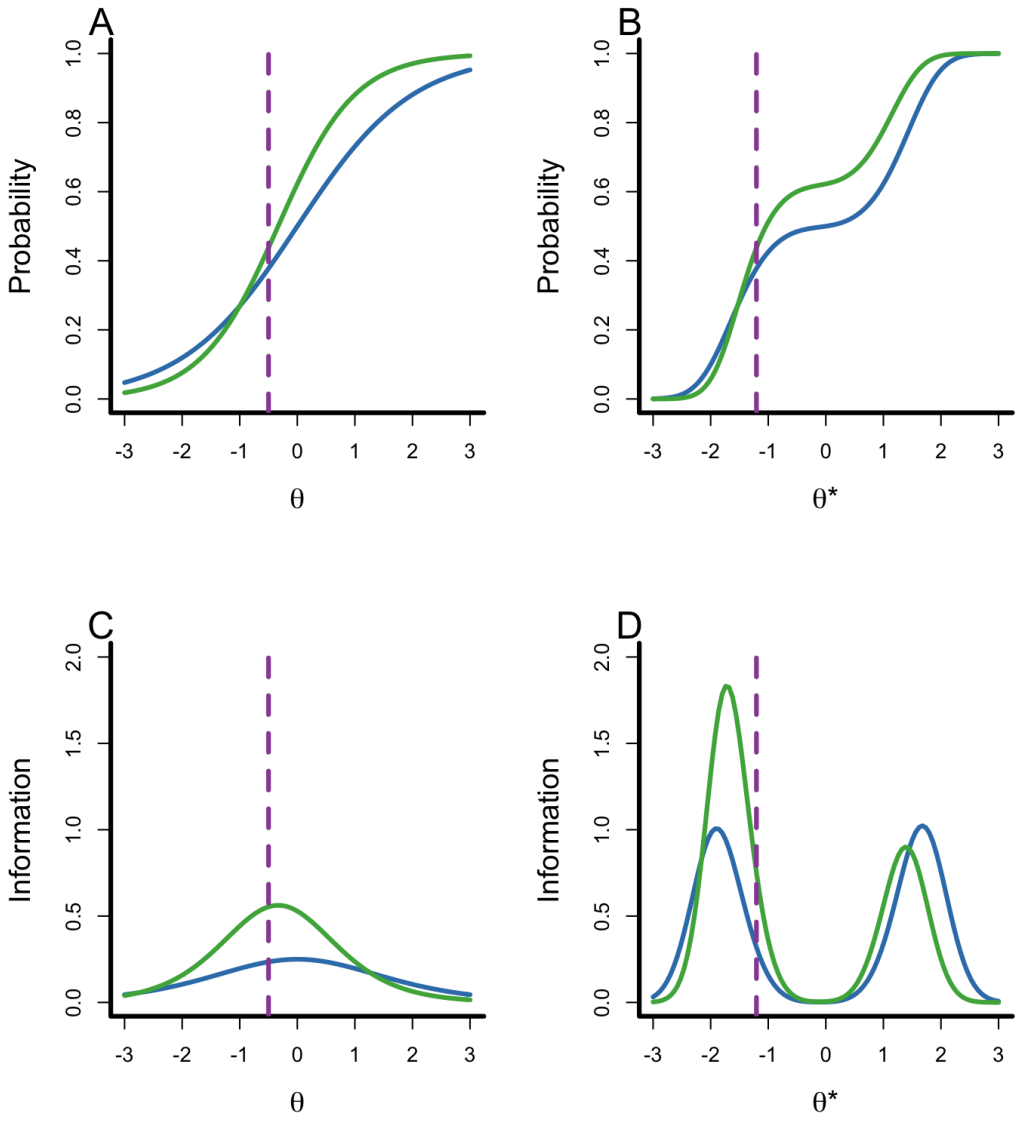

Figure 5.6. Example FMP item IRFs and IIFs on the θ and $\theta^{\star}$ metrics. Panels A and B show IRFs, and Panels C and D show IIFs. Panels A and C display functions on the θ metric, and Panels B and D display functions on the $\theta^{\star}$ metric. Item 1 is represented by the green curves, and item 2 is represented by the blue curves. The relative efficiency of item 1 to item 2 a fixed trait value (represented by the purple line) is equal for the θ and θ^{star} metrics.

5.5 Model Selection

One application of the composite FMP model is to generalize the sequential model-fitting algorithm described earlier for the ordinary FMP model. Specifically, increasing k_θ in the composite FMP model allows the user to explore non-normal distributions of the latent trait. Earlier, I described a method for FMP model selection that compares the fit of a model with $k_i = 0$ to a model with $k_i = 1$ for some item i. If a model with higher item complexities provides better fit as indicated by the AIC criterion, then models with higher complexities can be fit. This procedure is looped over items. Under the composite FMP model, the researcher can compare the fit of models that not only increase k_i for one or more items, but also explore increasing the k_θ value. The AIC-based model selection procedure described by (e.g.) Falk and Cai (2016) can be modified slightly to explore non-normal latent trait distributions. Specifically,

1. Fit the data to a model with $k_\theta = 0$ and $k_i = 0$ for all I.

2. Fit $I + 1$ models by setting exactly one of the following parameters to 1: $k_\theta, k_1, k_2, \ldots, k_I$.

3. Among the original model and the $I + 1$ fitted models, the model that leads to the lowest AIC value becomes the candidate model.

4. Fit another $I + 1$ models by adding 1 to exactly one of the $k_\theta, k_1, k_2, \ldots, k_I$ values that are specified in the candidate model.

5. Among the current candidate model and the $I + 1$ fitted models, the model with the lowest AIC value becomes the candidate model.

6. Repeat steps 3, 4, and 5 until increasing none of the complexity parameters

improves the AIC value over the current candidate model.

Chapter 6

Applications

In the previous section, I described a composite FMP model based on nonlinear transformations of the latent trait metric. In this section, it is demonstrated that the FMP nonlinear linking identities can be useful in constructing item response models on alternate latent trait metrics. As I have argued previously, the θ scale may not be the most useful or interpretable latent trait metric for empirical applications. This is evidenced by the fact that test scores are often reported on transformed metrics such as true-score metrics (Stocking, 1996), grade-equivalent metrics (Schulz & Nicewander, 1997), and a metric that is supposed to stabilize error variance (Kolen, 1988). Although scores are often reported on these alternate metrics, item response models are typically constructed on the θ scale. There are several reasons why it may be desirable to construct the IRT model directly on a transformed metric. First, it cannot be assumed that properties of trait estimates on one metric hold for trait estimates on a nonlinearly transformed metric. For example, as demonstrated by Yi, Wang, and Ban (2001), nonlinear

monotonic scale transformations can change the magnitude and sometimes even the direction of bias in estimated trait scores. The performance of computerized adaptive tests (CATs) is also heavily dependent on the transformation of the latent trait metric. Because standard errors on one metric can change with metric transformations, a standard-error CAT termination rule (Weiss, 1982) is not appropriate if the metric is transformed nonlinearly after trait estimation. A final reason why it is useful to obtain an item response model directly on the desired metric is in test construction. Specifically, the test information function is used to determine the trait levels at which the item pool provides high information. However, because the shape of the test information function is highly sensitive to the choice of metric, a test that appears to have high information for a wide range of trait levels could provide very low information at some trait levels after a metric transformation. For these reasons, there is a need to construct item response models directly on the desired metrics. For instance, Stocking (1996) discusses the challenges and merits of scoring adaptive tests directly on the more widely understood proportion-correct metric. Moreover, Yi et al. (2001, p. 289), commented that "[i]t is desirable to directly control measurement precision on the reported-score scale when making CAT decisions. However it is difficult to implement such an idea in practice, mainly because no existing index can be used to compute the conditional standard error on the reported-score scale to terminate a CAT." The composite FMP model provides a solution to this problem because it allows the quantities used in a CAT (e.g., the conditional standard error) to be computed directly on the desired metric.

One feature of the FMP model that makes it ideal for exploring alternate

latent trait metrics is that it includes the familiar 2PL as a special case. In many situations, however, we may need a model that includes the 3PL as a special case. A guessing-added FMP model has already been described by Falk and Cai (2015). This guessing-added model allows for a non-zero probability of a keyed response for all trait levels (typically attributed to random guessing behavior). Similarly, a guessing-added composite FMP model can be defined as

$$P(y_{in}|\theta, \boldsymbol{b}_i, \boldsymbol{t}, c_i) = c_i + (1 - c_i)H\{m_i[h(\theta^\star)]\}, \tag{6.1}$$

where c_i indicates an item-level lower asymptote parameter. Note that the c_i parameter is unaffected by the choice of metric transformation. For this reason, the linear and nonlinear FMP linking equations are unaffected by the presence of guessing. Thus, the 3PL is a special case of the guessing-added composite FMP model in Equation 6.1, and the linear and nonlinear FMP linking identities can be used to explore metric transformations for the 3PL.

6.1 Uncorrelated Parameters

One attempt to construct an alternate latent trait metric is described by Lord (1975), who argued that the $\theta^\star$ scale produced when fitting an item response model may not be the most useful representation of the latent trait. Specifically, he found that many empirical data sets, when fit to the 3PL, have positively correlated discrimination and difficulty parameters. Lord argued that these correlations are undesirable, and suggested transforming $\theta^\star$ such that across items, difficulties and

slopes are uncorrelated. Below, Lord's solution is demonstrated to be an instance of the guessing-added composite FMP model.

Consider the following parameterization of the 3PL,

$$P_i^\star(\theta^\star) = c_i^\star + (1 - c_i^\star)\left\{1 + \exp[-a_i^\star(\theta^\star - b_i^\star)]\right\}^{-1}, \tag{6.2}$$

where $a_i^\star$, $b_i^\star$, and $c_i^\star$ denote the item discrimination, difficulty, and guessing parameters, respectively. Note that this model is parameterized in terms of $\theta^\star$ for compatibility with the composite FMP model, and that it assumes the difficulty/discrimination, rather than the slope/intercept parameterization of the 3PL. In this parameterization, $b_i^\star$ is the IRF inflection point and equals the trait level at which the item response probability equals $.5(1 + c_i^\star)$, and $.25a_i^\star(1 - c_i^\star)$ is the slope of the IRF at $\theta^\star = b_i^\star$. Lord (1975) suggested the following method to find a transformation of the latent trait metric. First, for a given set of 3PL item parameters, fit a quadratic regression of item discriminations on item difficulties such that

$$a_i^\star = \hat{\beta}_0 + \hat{\beta}_1 b_i^\star + \hat{\beta}_2 b_i^{\star 2} + \epsilon_i. \tag{6.3}$$

Note that under this parameterization of the 3PL, the item difficulties $b_i^\star$ are on the same metric as the latent trait $\theta^\star$. To find an alternate metric, θ, on which difficulties and slopes are uncorrelated, Lord (1975) suggested that for individual n,

$$\theta_n = h(\theta_n^\star) = \int^{\theta_n^\star} \left(\hat{\beta}_0 + \hat{\beta}_1\theta^\star + \hat{\beta}_2\theta^{\star 2}\right) d\theta^\star, \tag{6.4}$$

$$= t_0 + t_1\theta_n^\star + t_2\theta_n^{\star 2} + t_3\theta_n^{\star 3}, \tag{6.5}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the coefficients found in Equation 6.3, $t_3 = \hat{\beta}_2/3$, $t_2 = \hat{\beta}_1/2$, $t_1 = \hat{\beta}_0$, and t_0 is an arbitrarily chosen constant of integration. The estimated coefficients $\boldsymbol{t} = (t_0, t_1, t_2, t_3)'$ can then be used to express the metric transformation generally:

$$\theta = t_0 + t_1\theta^\star + t_2\theta^{\star 2} + t_3\theta^{\star 3}. \tag{6.6}$$

To find an expression for the IRF on the transformed metric, first recall that $c_i^\star$ values are unaffected by metric transformations. Further, let $b_i = h(b_i^\star)$, and let a_i denote the IRF slope at $\theta = b_i$. With these definitions, the a_i values are independent of both b_i and $b_i^\star$. Note however, that b_i need not be an IRF inflection point on the θ metric, and that the a_i values need not be proportional to the maximum IRF slopes.

If the $\boldsymbol{t} = (t_0, t_1, t_2, t_3)'$ parameters define a monotonic polynomial, then these polynomial coefficients can define the metric transformation in the composite FMP model. Moreover, the original 3PL item parameters (transformed to the slope/threshold parameterization) can be used to find an expression for the IRF on the transformed metric.

When transforming the latent trait metric based on a criterion such as the relationship between item difficulties and discriminations, it is important to ensure that the resulting transformation is monotonic. Following the procedure outlined above, a necessary and sufficient condition to ensure monotonicity of $h(\theta^{\star})$ is that the fitted quadratic regression line given by Equation 6.3 is positive at all trait levels. Unfortunately, a positive correlation between item difficulty and discrimination (the problem that motivated Lord's suggested transformation) does not guarantee a positive quadratic regression. This problem was encountered in several sets of published 3PL item parameters investigated by the author. For example, Lord (1968) published 3PL item parameter estimates for an 80-item test. Although the estimated difficulty parameters correlate .43 with the estimated discrimination parameters, the ordinary least squares quadratic regression of discriminations on difficulties is not positive at all trait levels: $\hat{\beta}_0 = .97$, $\hat{\beta}_1 = .22$, and $\hat{\beta}_2 = -.03$. This problem appears to be a limitation of Lord's method that may limit the practicality of this method for lessening the correlation between difficulties and discriminations.

6.2 Approximating a Known Functional Transformation

In some cases, the functional form for a desired metric transformation is already known. In such cases, the metric transformation can be approximated to arbitrary

precision using an inverse polynomial function. Below, two functional transformations of θ are explored. First, the test response function (TRF) can be used to produce scores on the true score (number-correct) metric. Second, an arcsin transformation was proposed by Kolen (1988) as way to produce scales with nearly equal error variance across levels of the latent trait. This arcsin transformation has been used in practice to transform scores on the ACT assessment (Brennan & Kolen, 1989).

In classical test theory (CTT, Crocker & Algina, 1986), trait scores are estimated based on linear combinations of item responses. The simplest of these is the raw sum score. The observed number-correct score is then treated as an estimate of the expected number-correct score, or true score T, where the expectation is taken over a hypothetical infinite number of parallel tests. Both T and θ are latent variables, and under the MHM assumptions, θ is a monotonic transformation of T. Specifically,

$$T = \sum_{i=1}^{I} P_i(\theta) = \text{TRF}(\theta), \tag{6.7}$$

where $P_i(\theta)$ is any item response function that satisfies the MHM assumptions. Note that under the MHM assumptions, TRF is a monotonically increasing and invertible function of θ. Because T is a monotonic transformation of θ, it is possible to build an item response model for which T is the latent variable. This model would have an IRF that follows the general form

$$P_i(\theta) = P_i[\text{TRF}^{-1}(T)]. \tag{6.8}$$

The inverse TRF generally does not have a simple expression. Instead, if a polynomial approximation to $\text{TRF}^{-1}(T)$ is used, then this model can be expressed as an instance of the composite FMP model. That is, the TRF can be approximated by a monotonic polynomial, although a high-degree polynomial may be needed.

Another useful functional transformation of the latent metric is Kolen's (1988) arcsin transformation. This transformation, which is derived from the binomial error model, produces a metric that is supposed to have more equal standard errors across the latent trait continuum. This transformation is defined as a monotonic transformation of T as follows:

$$S = .5\left\{\arcsin\left[\left(\frac{T}{I+1}\right)^{1/2}\right] + \arcsin\left[\left(\frac{T+1}{I+1}\right)^{1/2}\right]\right\}, \tag{6.9}$$

where T is the true score, and I is the number of test items. Although neither the true-score nor arcsin transformations are explicitly polynomial functions, it has been shown that any monotonic function with continuous derivatives can be approximated to an arbitrary degree of precision by a polynomial function (Elphinstone, 1983, 1985). However, the approximation may require a high-degree polynomial for the desired degree of precision. This should not be a problem because the researcher is able to explicitly evaluate the goodness of approximation. The coefficients needed for the metric transformation can be found as follows. The researcher can generate a large number of θ values from a uniform or normal distribution over the θ range to which the test will be applied. For each θ value, the transformed score may be calculated using Equations 6.7 and 6.9. The polynomial coefficients can then be found by regressing θ on the transformed variable

using monotonic polynomial regression methods such as those implemented in the `MonoPoly` package (Murray et al., 2013) for `R` (R Core Team, 2015). This package contains several methods for fitting monotonic regression models. One of the options is based on the work of Elphinstone (1983, 1985) and is closely related to the method for constraining monotonic IRFs that is described earlier in this work.

Polynomial approximations to the true score and arcsin transformations are illustrated using the 3PL item parameters reported by Lord (1968) for an 80-item test. To obtain the polynomial approximation to an inverse TRF, I used 1,000 θ values spaced uniformly between $\theta = -7$ and $\theta = 7$. A wide range of θ values was used so that the corresponding true scores spanned most of the theoretical range of true scores. Given these item parameters, the true scores theoretically range from 12.45 to 80; for $\theta \in [-7, 7]$, the corresponding true scores range from 12.84 to 79.50. The TRF was then used to map the θ values to T. A series of monotonic polynomials with increasing k_θ values was then used to approximate the TRF. In Figure 6.1, T is plotted against the residuals for each regression model. Note that the residuals are on the true score metric. As expected, increasing the polynomial degree leads to a better approximation of the target function; the largest residuals occur with a linear approximation ($k_\theta = 0$) and the smallest maximum absolute residual occurs under a polynomial of degree 11 ($k_\theta = 5$), that is, under the highest-degree model fit to the simulated data.

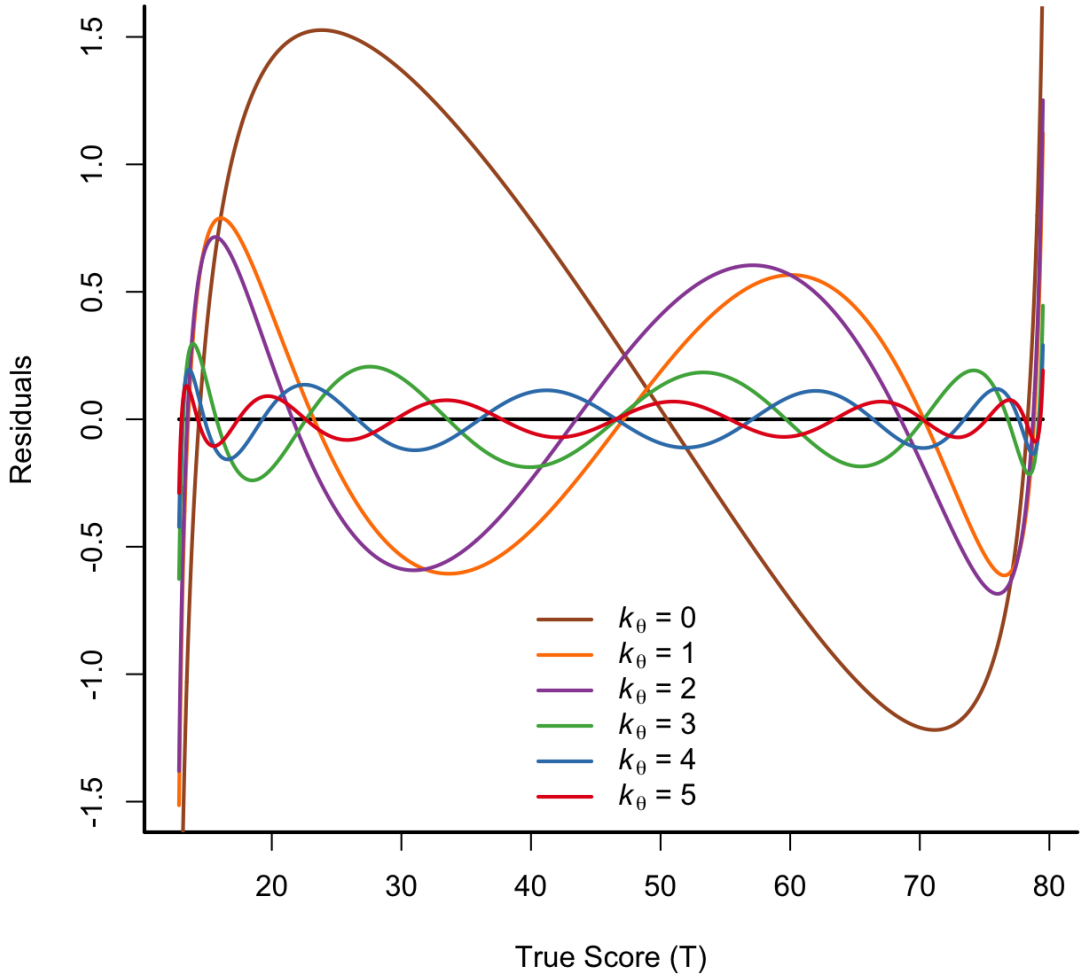

Figure 6.1. Monotonic polynomial approximation to an inverse test response function. A sequence of θ values was regressed on true scores using monotonic polynomials of increasing degrees (i.e., k_θ values). Smaller residuals indicate a better approximation.

Next, the T values computed for the previous figure were transformed using the arcsin transformation in Equation 6.9 to find a set of 1,000 S values. I

then regressed θ on standardized S values using a series of monotonic polynomial regressions with increasing k_θ values. The residuals from these regressions are plotted against S in Figure 6.2. This figure suggests that a very close approximation to the true function can be obtained with $k_\theta = 3$. The maximum absolute residual equalled only .01, a very small value considering that the S scores are standardized. These illustrations demonstrate that polynomial approximations are able to closely approximate a known functional transformation. The closeness of approximation can be controlled by the researcher by increasing the k_θ value. Further, depending on the purpose of the test, the researcher can approximate a functional transformation only for certain regions of the latent trait continuum. In this example, I used 1,000 θ values were used ranging from -7 to 7. However, the researcher can choose to use any number, range, or distribution of θ values that are appropriate for a given test.

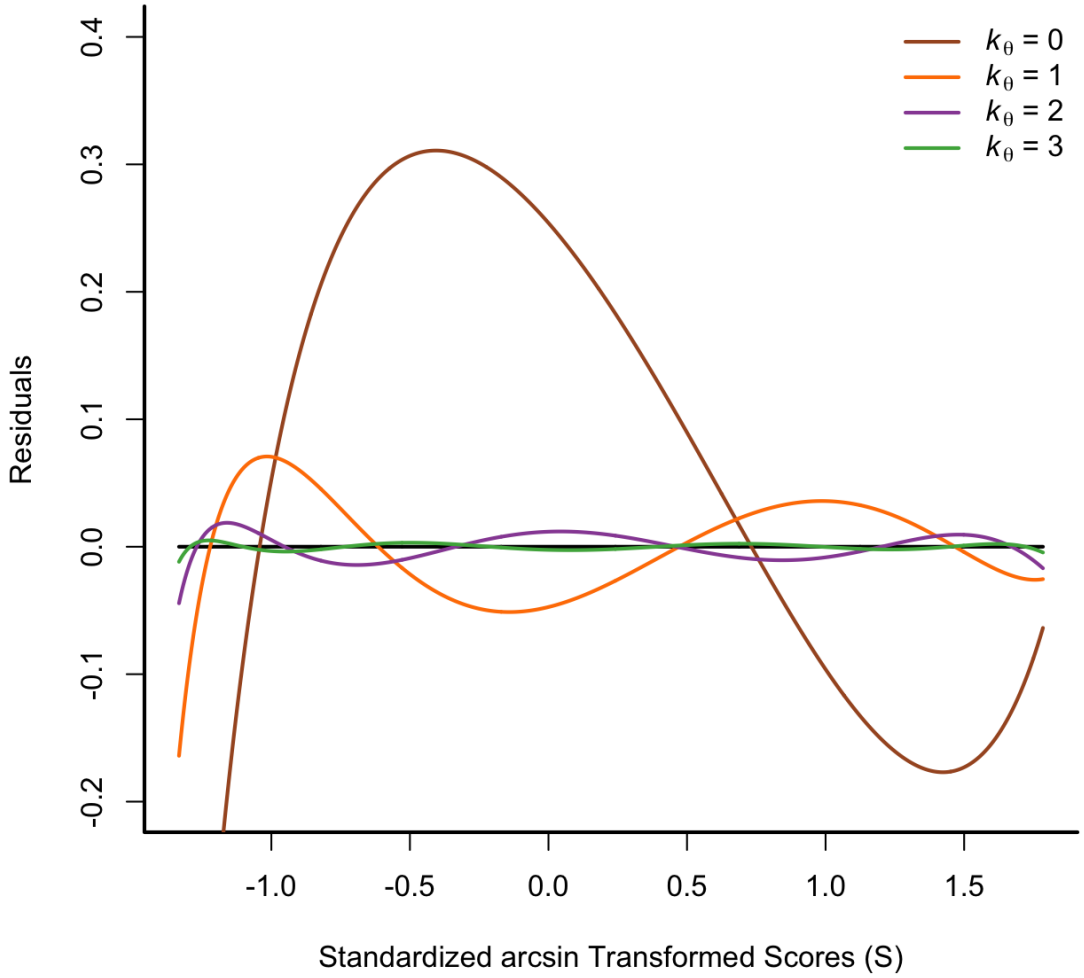

Figure 6.2. Monotonic polynomial approximation to θ as a function of Kolen's (1988) arcsin transformation of true scores. A sequence of θ values was regressed on transformed scores S using monotonic polynomials of increasing degrees (i.e., k_θ values). Smaller residuals indicate a better approximation.

6.3 Grade-Equivalent Scaling

In some cases, a metric may be desired such that the scale units are defined in terms of an external variable. For instance, one may with to construct a grade-equivalent scale wherein, on average, scores for a one-unit increase in grade level correspond to a one-unit increase in the latent trait. This application is similar to that described in the previous subsection except that the metric transformation is unknown in advance and must be approximated from data.

One existing method for constructing grade-equivalent scales is to regress estimated latent trait scores on grade levels using a quadratic regression model (see Schulz & Nicewander, 1997). Specifically, estimated trait scores can be regressed on integer grade levels for a sample of students. One problem with fitting a quadratic model to grade-level data is that the transformation need not be monotonic. In fact, when using quadratic regression (and the leading coefficient is nonzero), the transformation is never monotonic across the entire latent trait continuum. Even if the transformation is monotonic for all trait regions to which the scale will be applied, the lack of monotonicity excludes this metric transformation from the FMP framework. Instead, cubic regression may be used to find the needed transformation. To this end, a monotonic cubic polynomial may be fit to data using methods available, for example, in the `MonoPoly` package for `R` (Murray et al., 2013). If necessary, polynomials of higher degrees can be fit to the data.

To illustrate the application of the composite FMP model to construct grade-equivalent scores, I obtained $\hat{\theta}$ values for 46,667 students in grades 1 to 8 who were

administered an adaptive reading test. The $\hat{\theta}$ values were obtained via a previous administration of a computerized adaptive test. A histogram showing the number of students at each grade level is displayed in Figure 6.3. As this figure shows, all grade levels are well-represented by this data set. The majority of students are in grades 1–5, but even the least represented grade level (grade 8) has over 2,000 students.

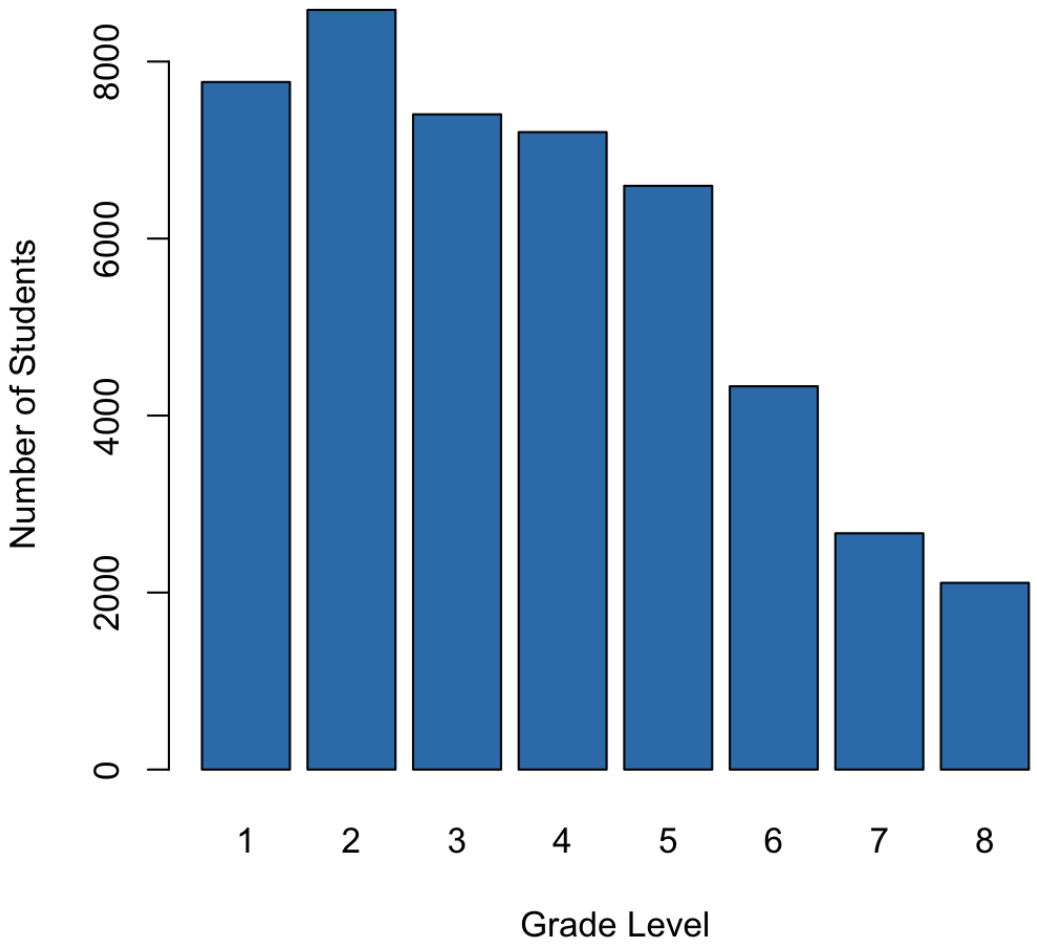

Figure 6.3. Histogram of the number of students belonging to each grade level in reading test data.

The conditional distributions of previously estimated $\hat{\theta}$ values are shown in Figure 6.4. Note that in these data, the mean $\hat{\theta}$ value at some grade levels is smaller than then median $\hat{\theta}$ value. As the figure shows, average $\hat{\theta}$ values increase

as grade levels increase, but the relationship between average $\hat{\theta}$ values and grade level is nonlinear. To elucidate this nonlinear relationship, I next applied a series of monotonic polynomial regression models to these data. Specifically, I regressed $\hat{\theta}$ on grade level using the `MonoPoly` package in `R` and $k_\theta = 0$, $k_\theta = 1$, $k_\theta = 2$, and $k_\theta = 3$. The AIC or BIC model selection criteria can be used to choose which k_θ value provides the best model fit relative to the number of estimated parameters. Comparing these fitted models, AIC selected the $k_\theta = 2$ model, and BIC selected the $k_\theta = 1$ model. The fitted curves are shown in Figure 6.4 for each model (the $k_\theta = 3$ model is not shown because it closely traces the $k_\theta = 2$ model). I decided to retain the $k_\theta = 1$ model as selected by the BIC criterion. Under the $k_\theta = 1$ model, the estimated transformation between grade levels and θ equals

$$\theta = -1.678 + .766 \times \text{grade} - .097 \times \text{grade}^2 + .004 \times \text{grade}^3. \tag{6.10}$$

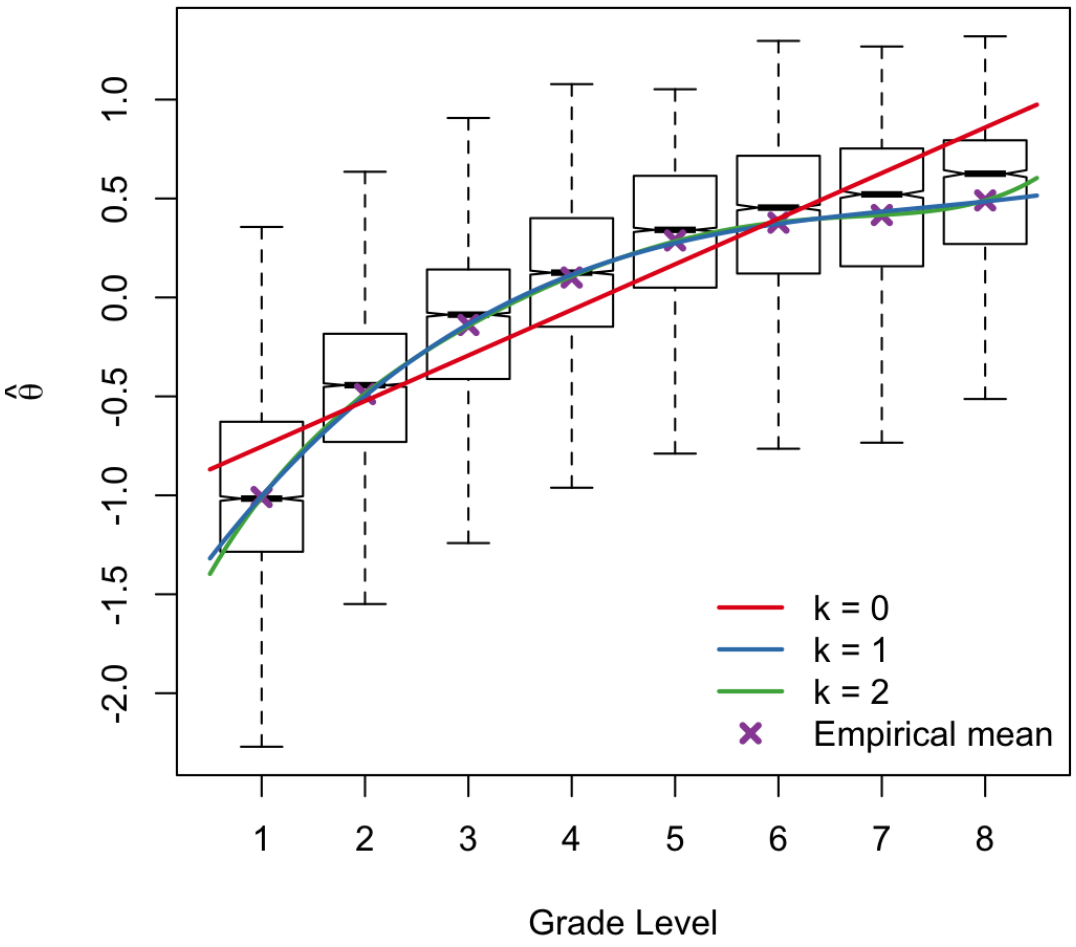

Figure 6.4. Box plots of estimated $\hat{\theta}$ values and the best-fit line computed using monotonic polynomial regression for reading test data.

The nonlinear θ transformation found using monotonic polynomial regression can be used in the composite FMP model to transform item parameters defined on the θ metric to be defined on the grade-equivalent metric. This is illustrated

using the 3PL item parameters that are reported by Lord (1968) for an 80-item test. For these 80 items, the test information function on the θ metric is shown in Panel A of Figure 6.5. The guessing-added composite FMP model was then used in conjunction with the estimated θ transformation to obtain 80 sets of transformed item parameters on the grade-equivalent metric. The transformed test information function on the grade-equivalent metric is shown in Panel B of Figure 6.5. In both panels, the grade-level means are shown by the vertical dotted red lines. In Panel A, we see that on the θ metric, there is substantial information at all mean grade levels. Specifically, all mean grade levels have an expected standard error of measurement between .42 and .58 on the θ metric, and there is more information for higher grade level means. In contrast, on the grade-equivalent metric, there is very little information at high grade levels. On the grade-equivalent metric, expected standard errors of measurement at grade means range from .98 to 7.6. Based on these items, there is very little information to distinguish response behavior for the grade 8 mean level from other grade levels. However, even at the grade level at which there is the most information, grade 1, the expected standard error of measurement is about 1.0 for this 80-item test. These results suggest that, for this transformation between grade-level scores and θ scores, test constructors should be careful to include items that provide high information on the grade-equivalent metric.

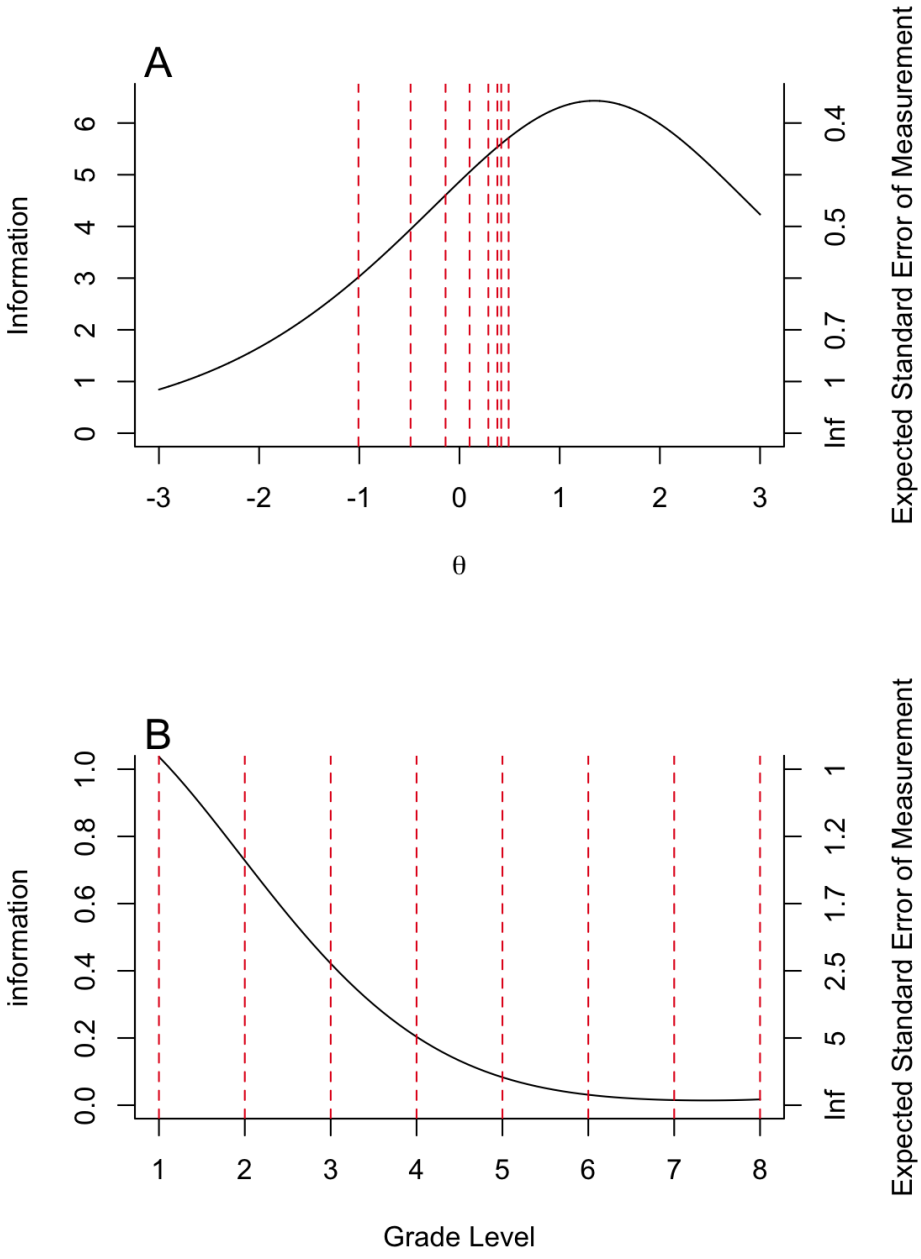

Figure 6.5. Information functions for an 80-item test on the θ metric (Panel A) and on an estimated grade-equivalent metric (Panel B). Red vertical lines indicate grade-level mean scores for grades 1–8.

Chapter 7

Discussion

In this dissertation, I have argued that nonlinear transformations of the IRT latent trait metric θ (a) are permissible under the assumptions underlying many item response models, (b) can result in latent trait metrics that are more useful or interpretable than the θ metric, and (c) can be modeled explicitly via an application of the filtered monotonic polynomial (FMP) item response model proposed by Liang and Browne (2015). Additionally, I have demonstrated that the FMP model is indeterminate such that an infinite number of FMP models can be specified that make identical predictions. Among FMP models that make identical predictions, the underlying latent trait metrics are related by monotonic transformations. Furthermore, I have derived linking equations such that the FMP item parameters for these equally predictive FMP models can be found using matrix algebra. Through an application of these linking equations, IRT models can be specified such that item response functions and item information functions are explicitly defined on a metric other than the θ metric. Consequently, the methods outlined in this

paper give researchers control over the metric underlying IRT models. For instance, researchers can apply nonlinear linking equations to construct tests based on metrics that are more easily communicated or more interpretable, such as true score or grade-equivalent metrics.

In addition to the discussion and application of metric transformations, this dissertation described FMP model estimation using both fixed- and random-effects methods. A simulation study was conducted to evaluate the accuracy of FMP model estimation (item parameter recovery, IRF recovery, and latent trait score recovery) in a number of sample size, test length, and item complexity conditions. The results of the simulation study indicated that individual item parameters were often highly inaccurate, even for large data sets, and especially for more complex IRFs (i.e., $k_i \geq 1$). For this reason, users should avoid interpreting FMP item parameters. Instead, I argue that it is more informative to evaluate recovery in terms of IRF accuracy, as indexed by RIMSE_i. Evaluated in terms of RIMSE_i, fixed-effects and random-effects estimation led to comparable IRF recovery, although random-effects led to more accurate estimated IRFs in short tests. Next, latent trait recovery was assessed using Pearson, Spearman, and Kendall correlation coefficients between the true and estimated latent trait scores. Each type of correlation indicated highly accurate latent trait recovery for data sets with at least 40 items and 1,000 subjects. In addition, the AIC and BIC model selection criteria were evaluated in terms of whether they selected the data-generating model. Among the studied model selection methods, random-effects estimation paired with the BIC criterion led to the highest number of accurately selected model complexities (i.e., $k_i = \tilde{k}_i$). Moreover, the BIC rarely

selected overly complex models whereas the AIC often selected overly complex models. I argued that it is preferable to select overly simply models rather than select overly complex models, and thus I recommend random-effects estimation with BIC model selection in practice. Finally, AIC and BIC were evaluated in terms of IRF recovery for the $\tilde{k}_i$ values selected by each criterion. Conditional on sample size and test length, few differences existed among the conditions. For all conditions, fairly accurate IRFs are obtained with sample sizes of at least 1,000. Overall, random-effects estimation paired with the BIC is recommended when estimating FMP models, and it is recommended that data sets have at least 1,000 subjects and 40 items.

Returning to the primary focus of this work, the metric transformation methods outlined above hinge on an appropriate understanding of the latent trait metric in IRT. Specifically, these methods make use of the generally unacknowledged fact that in IRT models, the unit and interval spacing of the θ scale is identified by model identification restrictions. These identification restrictions are typically chosen for convenience or tradition rather than for substantive reasons (Lord, 1975). Thus, the identified scaling of the latent trait θ may not be the most practical or substantively interesting scaling possible. Although θ estimates are often transformed *after* item and person parameter estimation, properties of items (e.g., IRFs and IIFs) are not routinely transformed. The failure to transform items can lead to misleading conclusions because not all properties of items and tests on the θ metric apply to the transformed metric. For instance, a test that is highly informative for a wide range of trait levels on the θ metric need not be highly informative, or provide the same amount of information, at corresponding

trait levels on the transformed metric (Lord, 1980, pp. 85–88).

The FMP model is not strictly necessary to explicitly model nonlinear transformations of the latent trait. Instead of modeling the metric transformation as a monotonic polynomial, any monotonic transformation of the latent trait may be used (Lord, 1975). Moreover, polynomials are not as flexible as other functional families (e.g., Tadikamalla, 1980) and may introduce biases when using polynomials of small degrees. Despite the limitations associated with monotonic polynomials, the FMP model boasts several properties that justify its use. First, the FMP model includes the 2PL as a special case, and the guessing-added FMP model includes the 3PL as a special case. Thus, the FMP-based methods proposed in this paper can be applied to existing sets of estimated item parameters that belong to the 2PL or 3PL. Second, the composite FMP model provides simple closed-form expressions of transformed IRFs. This is not generally true for nonlinear metric transformations. For instance, note that the true score metric T is a nonlinear transformation of the θ metric. To transform IRFs from the θ metric to the true score metric, the *inverse* test response function is needed, which does not have a simple expression. Third, the composite FMP model is highly flexible and is able to represent increasingly complex response functions and latent trait transformations as model complexity increases. Even if the desired latent trait transformation is not a polynomial, a polynomial function specified to a sufficiently high degree can approximate the latent trait transformation to an arbitrary degree of precision (Elphinstone, 1983, 1985). Finally, unlike many functional approximations, it is possible to guarantee the monotonicity of polynomial approximations. Namely, FMP model estimation is closely related to monotonic

polynomial regression, which estimates the relationship between two variables as a monotonically increasing polynomial function. Monotonic polynomial regression is an active area of research (Murray et al., 2013; Murray, Müller, & Turlach, in press), and it is expected that new advances in this area will further improve the estimation efficiency and accuracy of the FMP item response model.

In conclusion, the composite FMP model developed in this dissertation provides the first comprehensive framework in which nonlinear transformations of the latent trait metric can be modeled explicitly. This composite model gives researchers the freedom to transform IRT models onto a latent trait metric other than the θ metric, and allows researchers to easily transform other important quantities such as information functions and standard errors. The ideas put forth in this work promote a sophisticated understanding of the latent trait metric while allowing researchers to more broadly apply metrics that may have more desirable properties than the θ metric.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox & F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Andersen, E. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Andersen, E., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, *42*, 357–374.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, *26*, 291–309.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*, 147–162.

Baker, F. B., & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.

Barton, M., & Lord, F. (1981). *An upper asymptote for the three-parameter*

logistic item-response model. (Research Report No. 81–20). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *4*6, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

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

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*, 141–162.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 504–516.

Brennan, R., & Kolen, M. (1989). Scaling the ACT assessment and P-ACT*: Rationale and goals. In R. Brennan (Ed.), *Methodology used in scaling the ACT assessment and P-ACT** (pp. 1–17). Iowa City, IA: Author.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108–132.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithm. *IMA Journal of Applied Mathematics*, *6*, 76–90.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253–260.

Chen, X., & Tung, Y. -K. (2003). Investigation of polynomial normal transform. *Structural Safety*, *25*, 423–445.

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

Coombs, C. H. (1964). *A Theory of Data.* New York: Wiley.

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

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika*, *42*, 523–548.

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

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*, 1-19.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. L.

(1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, *13*, 285–299.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modeling for item response. *Statistical Modelling*, *8*, 41–66.

Duncan, K. A., & MacEachern, S. N. (2013). Nonparametric Bayesian modeling of item response curves with a three-parameter logistic prior mean. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (pp. 108–125). New York, NY: Routledge.

Elphinstone, C. D. (1983). A target distribution model for nonparametric density estimation. *Communication in Statistics–Theory and Methods*, *12*, 161–198.

Elphinstone, C. D. (1985). *A method of distribution and density estimation* (Unpublished dissertation). University of South Africa, Pretoria, South Africa.

Falk C. F., & Cai, L. (2015, April). *Semi-parametric item response functions in the context of guessing* (CRESST Report 884). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika, 81*, 434–460.

Fleishman, A. J. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, *13*, 317–322.

Goldfarb, D. (1970). A family of variable metric methods derived by variational

means. *Mathematics of Computation*, *24*, 23–26.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics*, *13*, 243–271.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis*, *40*, 685–711.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*, 367–386.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Ip, E. H., & Chen, S.-H. (2015). Using projected locally dependent unidimensional models to measure multidimensional response data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 226–251). New York, NY: Routledge.

Ip, E. H., Wang, Y. J., de Boeck, P., & Meulders, M. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of

hostility. *Psychometrika*, *69*, 191–216.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*, 389–423.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*, 371–397.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 51–66.

Kirisci, L., Hsu, T.-C, & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146–162.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*, 97–110.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking (3rd. Ed)*. New York: Springer.

Lathrop, Q. N. (2015). Abstract: IRT and SVD: Implementing psychological methods in new and complex situations. *Multivariate Behavioral Research*, *50*, 138.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, *61*, 273–287.

Lawley, D. N. (1944). The factorial analysis of multiple test items. *Proceedings of the Royal Society of Edinburgh*, *62-A*, 74–82.

Levine, M. V. (1984). *An introduction to multilinear formula score theory.* (Personnel and Training Research Programs, Office of Naval Research, Measurement Series No. 84-4). Arlington, VA: Personnel and Training Research Programs.

Liang, L. (2007). *A semi-parametric approach to estimating item response functions* (Unpublished doctoral dissertation). The Ohio State University: Columbus, OH.

Liang, L., & Browne, M. W. (2015). A quasi-parametric method for fitting flexible item response functions. *Journal of Educational and Behavioral Statistics*, *40*, 5–34.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*, 989–1020.

Lord, F. M. (1974). The relative efficiency of two tests as a function of ability level. *Psychometrika*, *39*, 351–358.

Lord, F. M. (1975). The 'ability' scale in item characteristic curve theory. *Psychometrika*, *40*, 205–217.

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

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new

type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160.

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*, 1–34.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*, 354–368.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Mislevy, R. J. (1987). Recent developments in item response theory with implications for teacher certification. *Review of Research in Education*, *14*, 239–275.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague: Mouton.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement*, *25*, 295–299.

Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression*, Boston: Addison-Wesley.

Murray, K., Müller, S., & Turlach, B. A. (2013). Revisiting fitting monotone polynomials to data. *Computational Statistics*, *28*, 1989–2005.

Murray, K., Müller, S., & Turlach, B. A. (in press). Fast and flexible methods for

monotone polynomial fitting. *Journal of Statistical Computation and Simulation*, Manuscript accepted for publication.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

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

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement, *Applied Psychological Measurement*, *3*, 237–255.

R Core Team. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353–368.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630.

Ramsay, J. O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, *84*, 906–915.

Ramsay, J. O., & Winsberg, S. (1991). Maximum marginal likelihood estimation

for semiparametric item analysis. *Psychometrika*, *56*, 365–379.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press.

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

Reise, S. P. (2010). Thurstone might have been right about attitudes, but Drasgow, Chernyshenko, and Stark fail to make the case for personality. *Industrial and Organizational Psychology*, *3*, 485–488.

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13–40). New York, NY: Routledge.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*, 164–184.

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

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

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, *20*, 231–255.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *27*, 291–317.

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recovery from early mistakes in CAT? *Applied Psychological Measurement*, *33*, 83–101.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, *65*, 319–355.

Schulz, E. M., & Nicewander, W. A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement*, *34*, 315–331.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299–311.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, *24*, 647–656.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 719–746). Amsterdam: Elsevier.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall: London.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25–39.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, *21*, 365–389.

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

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1–16.

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement*, *25*, 300–306.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, *44*, 313–324.

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika*, *45*, 273–279.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement

models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology*, *94*, 1287–1304.

Thissen, D. (2009). On interpreting the parameters for any item response model. *Measurement*, *7*, 106–110.

van den Oord, E. J. C. G. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement*, *29*, 45–64.

van der Linden, W. J., & Barrett, M. D. (in press). Linking item response model parameters. *Psychometrika*, Manuscript accepted for publication.

Waller, N. G., & Feuerstahler, L. M. (2016). Bayesian modal estimation of the four-parameter item response model in real and simulated data sets. Manuscript submitted for publication.

Waller, N. G. & Jones, J. A. (2015). fungible: Fungible coefficients and Monte Carlo functions. R package version 1.3.

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

Wolfram Research, Inc. (2015, Version 10.2). [Computer software]. Champaign, IL: Wolfram Research, Inc.

Woods, C. M. (2007a). Empirical histograms in item response theory with ordinal data. *Educational Psychological Measurement*, *67*, 73–87.

Woods, C. M. (2007b). Ramsay curve IRT for Likert-type data. *Applied Psychological Measurement*, *31*, 195–212.

Woods, C. M. (2008a). Consequences of ignoring guessing when estimating the latent density in item response theory. *Applied Psychological Measurement*, *32*, 371–384.

Woods, C. M. (2008b). Ramsay-curve item response theory for the three-parameter logistic item response model. *Applied Psychological Measurement*, *32*, 447–465.

Woods, C. M. (2015). Estimating the latent density in unidimensional IRT to permit non-normality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 60–84). New York, NY: Routledge.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*, 102–117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*, 121–137.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

Yi, Q., Wang, T., & Ban, J. -C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, *38*, 267–292.

Appendix A

Linking Coefficients

Linear and nonlinear linking coefficients for the FMP model can be found using matrix algebra. Specifically, for item i,

$$\boldsymbol{b}_i^\star = \boldsymbol{W}\boldsymbol{b}_i,$$

where $\boldsymbol{b}_i$ is a vector of item parameters on the θ metric and $\boldsymbol{b}_i^\star$ is the desired vector of item parameters on the $\theta^\star$ metric. In this appendix, the $\boldsymbol{W}$ matrices are shown for various combinations of k_i and k_θ. Note that $\boldsymbol{W}$ is a matrix of dimension $(2k_i^\star + 2) \times (2k_i + 2)$ where $k_i^\star = 2k_ik_\theta + k_i + k_\theta$. Below, the $\boldsymbol{W}$ matrices are described in three sections corresponding to $k_\theta = \{0, 1, 2\}$. For $k_\theta = \{1, 2\}$, the $\boldsymbol{W}$ matrix is defined for $k_i = 2$. For $k_i < 2$, the appropriate $\boldsymbol{W}$ matrix consists of the first $2k_i^\star + 2$ rows and the first $2k_i + 2$ columns of $\boldsymbol{W}$.

A.0.1 Linear Metric Transformations ($k_\theta = 0$)

If $k_\theta = 0$, then $\boldsymbol{t} = (t_0, t_1)'$. If $k_i = 0$,

$$\boldsymbol{W} = \begin{bmatrix} 1 & t_0 \\ 0 & t_1 \end{bmatrix},$$

if $k_i = 1$,

$$\boldsymbol{W} = \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 \\ 0 & t_1 & 2t_0t_1 & 3t_0^2t_1 \\ 0 & 0 & t_1^2 & 3t_0t_1^2 \\ 0 & 0 & 0 & t_1^3 \end{bmatrix},$$

and if $k_i = 2$,

$$\boldsymbol{W} = \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 & t_0^4 & t_0^5 \\ 0 & t_1 & 2t_0t_1 & 3t_0^2t_1 & 4t_0^3t_1 & 5t_0^4t_1 \\ 0 & 0 & t_1^2 & 3t_0t_1^2 & 6t_0^2t_1^2 & 10t_0^3t_1^2 \\ 0 & 0 & 0 & t_1^3 & 4t_0t_1^3 & 10t_0^2t_1^3 \\ 0 & 0 & 0 & 0 & t_1^4 & 5t_0t_1^4 \\ 0 & 0 & 0 & 0 & 0 & t_1^5 \end{bmatrix}.$$

A.0.2 Cubic Polynomial Metric Transformations ($k_\theta = 1$)

If $k_\theta = 1$, then $\boldsymbol{t} = (t_0, t_1, t_2, t_3)'$. Let $[\boldsymbol{W}]_{r,c}$ denote the row r, column c element of $\boldsymbol{W}$. Elements of $\boldsymbol{W}$ that are *not* defined below equal 0. Possibly nonzero

elements of $\boldsymbol{W}$ equal

$$
\begin{aligned}
[\boldsymbol{W}]_{1,1} &= 1, \\
[\boldsymbol{W}]_{1,2} &= t_0, \\
[\boldsymbol{W}]_{2,2} &= t_1, \\
[\boldsymbol{W}]_{3,2} &= t_2, \\
[\boldsymbol{W}]_{4,2} &= t_3, \\
[\boldsymbol{W}]_{1,3} &= t_0^2, \\
[\boldsymbol{W}]_{2,3} &= 2t_0t_1, \\
[\boldsymbol{W}]_{3,3} &= 2t_0t_2 + t_1^2, \\
[\boldsymbol{W}]_{4,3} &= 2t_0t_3 + 2t_1t_2, \\
[\boldsymbol{W}]_{5,3} &= 2t_1t_3 + t_2^2, \\
[\boldsymbol{W}]_{6,3} &= 2t_2t_3, \\
[\boldsymbol{W}]_{7,3} &= t_3^2, \\
[\boldsymbol{W}]_{1,4} &= t_0^3, \\
[\boldsymbol{W}]_{2,4} &= 3t_0^3t_1, \\
[\boldsymbol{W}]_{3,4} &= 3t_0(t_1^2 + t_0t_2), \\
[\boldsymbol{W}]_{4,4} &= t_1^3 + 6t_0t_1t_2 + 3t_0^2t_3, \\
[\boldsymbol{W}]_{5,4} &= 3(t_1^2t_2 + t_0t_2^2 + 2t_0t_1t_3), \\
[\boldsymbol{W}]_{6,4} &= 3(t_1t_2^2 + t_1^2t_3 + 2t_0t_2t_3), \\
[\boldsymbol{W}]_{7,4} &= t_2^3 + 6t_1t_2t_3 + 3t_0t_3^2, \\
[\boldsymbol{W}]_{8,4} &= 3t_3(t_2^2 + t_1t_3), \\
[\boldsymbol{W}]_{9,4} &= 3t_2t_3^2, \\
[\boldsymbol{W}]_{10,4} &= t_3^3,
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{W}]_{1,5} &= t_0^4, \\
[\boldsymbol{W}]_{2,5} &= 4t_0^3 t_1, \\
[\boldsymbol{W}]_{3,5} &= 2t_0^2(3t_1^2 + 2t_0 t_2), \\
[\boldsymbol{W}]_{4,5} &= 4t_0(t_1^3 + 3t_0 t_1 t_2 + t_0^2 t_3), \\
[\boldsymbol{W}]_{5,5} &= t_1^4 + 1^2 t_0 t_1^2 t_2 + 6t_0^2 t_2^2 + 1^2 t_0^2 t_1 t_3, \\
[\boldsymbol{W}]_{6,5} &= 4(t_1^3 t_2 + 3t_0 t_1 t_2^2 + 3t_0 t_1^2 t_3 + 3t_0^2 t_2 t_3), \\
[\boldsymbol{W}]_{7,5} &= 6t_1^2 t_2^2 + 4t_0 t_2^3 + 4t_1^3 t_3 + 2^4 t_0 t_1 t_2 t_3 + 6t_0^2 t_3^2, \\
[\boldsymbol{W}]_{8,5} &= 4[3t_1^2 t_2 t_3 + 3t_0 t_2^2 t_3 + t_1(t_2^3 + 3t_0 t_3^2)], \\
[\boldsymbol{W}]_{9,5} &= t_2^4 + 1^2 t_1 t_2^2 t_3 + 6t_1^2 t_3^2 + 1^2 t_0 t_2 t_3^2, \\
[\boldsymbol{W}]_{10,5} &= 4t_3(t_2^3 + 3t_1 t_2 t_3 + t_0 t_3^2), \\
[\boldsymbol{W}]_{11,5} &= 2t_3^2(3t_2^2 + 2t_1 t_3), \\
[\boldsymbol{W}]_{12,5} &= 4t_2 t_3^3, \\
[\boldsymbol{W}]_{13,5} &= t_3^4, \\
[\boldsymbol{W}]_{1,6} &= t_0^5, \\
[\boldsymbol{W}]_{2,6} &= 5t_0^4 t_1, \\
[\boldsymbol{W}]_{3,6} &= 5t_0^3(2t_1^2 + t_0 t_2), \\
[\boldsymbol{W}]_{4,6} &= 5t_0^2(2t_1^3 + 4t_0 t_1 t_2 + t_0^2 t_3), \\
[\boldsymbol{W}]_{5,6} &= 5t_0(t_1^4 + 6t_0 t_1^2 t_2 + 2t_0^2 t_2^2 + 4t_0^2 t_1 t_3), \\
[\boldsymbol{W}]_{6,6} &= t_1^5 + 20t_0 t_1^3 t_2 + 30t_0^2 t_1 t_2^2 + 30t_0^2 t_1^2 t_3 + 20t_0^3 t_2 t_3, \\
[\boldsymbol{W}]_{7,6} &= 5[t_1^4 t_2 + 6t_0 t_1^2 t_2^2 + 4t_0 t_1^3 t_3 + 1^2 t_0^2 t_1 t_2 t_3 + 2t_0^2(t_2^3 + t_0 t_3^2)], \\
[\boldsymbol{W}]_{8,6} &= 5[2t_1^3 t_2^2 + t_1^4 t_3 + 1^2 t_0 t_1^2 t_2 t_3 + 6t_0^2 t_2^2 t_3 + 2t_0 t_1(2t_2^3 + 3t_0 t_3^2)], \\
[\boldsymbol{W}]_{9,6} &= 5[4t_1^3 t_2 t_3 + 1^2 t_0 t_1 t_2^2 t_3 + 2t_1^2(t_2^3 + 3t_0 t_3^2) + t_0 t_2(t_2^3 + 6t_0 t_3^2)], \\
[\boldsymbol{W}]_{10,6} &= 5[6t_1^2 t_2^2 t_3 + 2t_1^3 t_3^2 + 2t_0 t_3(2t_2^3 + t_0 t_3^2) + t_1(t_2^4 + 1^2 t_0 t_2 t_3^2)],
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{W}]_{11,6} &= t_2^5 + 20t_1t_2^3t_3 + 30t_1^2t_2t_3^2 + 30t_0t_2^2t_3^2 + 20t_0t_1t_3^3, \\
[\boldsymbol{W}]_{12,6} &= 5t_3(t_2^4 + 6t_1t_2^2t_3 + 2t_1^2t_3^2 + 4t_0t_2t_3^2), \\
[\boldsymbol{W}]_{13,6} &= 5t_3^2(2t_2^3 + 4t_1t_2t_3 + t_0t_3^2), \\
[\boldsymbol{W}]_{14,6} &= 5t_3^3(2t_2^2 + t_1t_3), \\
[\boldsymbol{W}]_{15,6} &= 5t_2t_3^4, \text{ and} \\
[\boldsymbol{W}]_{16,6} &= t_3^5.
\end{aligned}
$$

A.0.3 Quintic Polynomial Metric Transformations ($k_\theta = 2$)

If $k_\theta = 2$ then, $\boldsymbol{t} = (t_0, t_1, t_2, t_3, t_4, t_5)'$. Let $[\boldsymbol{W}]_{rc}$ denote the row r, column c element of $\boldsymbol{W}$. Elements of $\boldsymbol{W}$ that are *not* defined below equal 0. Possibly nonzero elements of $\boldsymbol{W}$ equal

$$
\begin{aligned}
[\boldsymbol{W}]_{1,1} &= 1, \\
[\boldsymbol{W}]_{1,2} &= t_0, \\
[\boldsymbol{W}]_{2,2} &= t_1, \\
[\boldsymbol{W}]_{3,2} &= t_2, \\
[\boldsymbol{W}]_{4,2} &= t_3, \\
[\boldsymbol{W}]_{5,2} &= t_4, \\
[\boldsymbol{W}]_{6,2} &= t_5, \\
[\boldsymbol{W}]_{1,3} &= t_0^2, \\
[\boldsymbol{W}]_{2,3} &= 2t_0t_1, \\
[\boldsymbol{W}]_{3,3} &= t_1^2 + 2t_0t_2, \\
[\boldsymbol{W}]_{4,3} &= 2(t_1t_2 + t_0t_3), \\
[\boldsymbol{W}]_{5,3} &= t_2^2 + 2t_1t_3 + 2t_0t_4,
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{W}]_{6,3} &= 2(t_2t_3 + t_1t_4 + t_0t_5),\\
[\boldsymbol{W}]_{7,3} &= t_3^2 + 2t_2t_4 + 2t_1t_5,\\
[\boldsymbol{W}]_{8,3} &= 2(t_3t_4 + t_2t_5),\\
[\boldsymbol{W}]_{9,3} &= t_4^2 + 2t_3t_5,\\
[\boldsymbol{W}]_{10,3} &= 2t_4t_5,\\
[\boldsymbol{W}]_{11,3} &= t_5^2,\\
[\boldsymbol{W}]_{1,4} &= t_0^3,\\
[\boldsymbol{W}]_{2,4} &= 3t_0^2t_1,\\
[\boldsymbol{W}]_{3,4} &= 3t_0(t_1^2 + t_0t_2),\\
[\boldsymbol{W}]_{4,4} &= t_1^3 + 6t_0t_1t_2 + 3t_0^2t_3,\\
[\boldsymbol{W}]_{5,4} &= 3[t_1^2t_2 + 2t_0t_1t_3 + t_0(t_2^2 + t_0t_4)],\\
[\boldsymbol{W}]_{6,4} &= 3[t_1^2t_3 + t_1(t_2^2 + 2t_0t_4) + t_0(2t_2t_3 + t_0t_5)],\\
[\boldsymbol{W}]_{7,4} &= t_2^3 + 6t_1t_2t_3 + 3t_0t_3^2 + 3t_1^2t_4 + 6t_0t_2t_4 + 6t_0t_1t_5,\\
[\boldsymbol{W}]_{8,4} &= 3[t_2^2t_3 + t_1t_3^2 + 2t_0t_3t_4 + t_1^2t_5 + 2t_2(t_1t_4 + t_0t_5)],\\
[\boldsymbol{W}]_{9,4} &= 3[t_2^2t_4 + 2t_1t_3t_4 + t_2(t_3^2 + 2t_1t_5) + t_0(t_4^2 + 2t_3t_5)],\\
[\boldsymbol{W}]_{10,4} &= t_3^3 + 6t_2t_3t_4 + 3t_1t_4^2 + 3t_2^2t_5 + 6t_1t_3t_5 + 6t_0t_4t_5,\\
[\boldsymbol{W}]_{11,4} &= 3[t_3^2t_4 + t_2t_4^2 + 2t_2t_3t_5 + t_5(2t_1t_4 + t_0t_5)],\\
[\boldsymbol{W}]_{12,4} &= 3[t_3t_4^2 + t_3^2t_5 + t_5(2t_2t_4 + t_1t_5)],\\
[\boldsymbol{W}]_{13,4} &= t_4^3 + 6t_3t_4t_5 + 3t_2t_5^2,\\
[\boldsymbol{W}]_{14,4} &= 3t_5(t_4^2 + t_3t_5),\\
[\boldsymbol{W}]_{15,4} &= 3t_4t_5^2,\\
[\boldsymbol{W}]_{16,4} &= t_5^3,\\
[\boldsymbol{W}]_{1,5} &= t_0^4,
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{W}]_{2,5} &= 4t_0^3t_1, \\
[\boldsymbol{W}]_{3,5} &= 2t_0^2(3t_1^2+2t_0t_2), \\
[\boldsymbol{W}]_{4,5} &= 4t_0(t_1^3+3t_0t_1t_2+t_0^2t_3), \\
[\boldsymbol{W}]_{5,5} &= t_1^4+1^2t_0t_1^2t_2+6t_0^2t_2^2+1^2t_0^2t_1t_3+4t_0^3t_4, \\
[\boldsymbol{W}]_{6,5} &= 4[t_1^3t_2+3t_0t_1^2t_3+3t_0t_1(t_2^2+t_0t_4)+t_0^2(3t_2t_3+t_0t_5)], \\
[\boldsymbol{W}]_{7,5} &= 2[2t_1^3t_3+3t_1^2(t_2^2+2t_0t_4)+t_0(2t_2^3+3t_0t_3^2+6t_0t_2t_4) \\
&\quad +6t_0t_1(2t_2t_3+t_0t_5)], \\
[\boldsymbol{W}]_{8,5} &= 4[t_1^3t_4+t_1(t_2^3+3t_0t_3^2+6t_0t_2t_4)+3t_1^2(t_2t_3+t_0t_5) \\
&\quad +3t_0(t_2^2t_3+t_0t_3t_4+t_0t_2t_5)], \\
[\boldsymbol{W}]_{9,5} &= t_2^4+6t_1^2t_3^2+2^4t_0t_1t_3t_4+1^2t_2^2(t_1t_3+t_0t_4)+4t_1^3t_5+6t_0^2(t_4^2+2t_3t_5) \\
&\quad +1^2t_2[t_1^2t_4+t_0(t_3^2+2t_1t_5)], \\
[\boldsymbol{W}]_{10,5} &= 4[t_2^3t_3+3t_1^2t_3t_4+3t_0^2t_4t_5+3t_2^2(t_1t_4+t_0t_5)+3t_2(t_1t_3^2+2t_0t_3t_4+t_1^2t_5) \\
&\quad +t_0(t_3^3+3t_1t_4^2+6t_1t_3t_5)], \\
[\boldsymbol{W}]_{11,5} &= 2\{2t_2^3t_4+3t_2^2(t_3^2+2t_1t_5)+3t_1^2(t_4^2+2t_3t_5)+2t_1(t_3^3+6t_0t_4t_5) \\
&\quad +3t_0(2t_3^2t_4+t_0t_5^2)+6t_2[2t_1t_3t_4+t_0(t_4^2+2t_3t_5)]\}, \\
[\boldsymbol{W}]_{12,5} &= 4\{3t_2^2t_3t_4+t_2^3t_5+t_2[t_3^3+6t_1t_3t_5+3t_4(t_1t_4+2t_0t_5)]+3[t_1^2t_4t_5 \\
&\quad +t_0t_3(t_4^2+t_3t_5)+t_1(t_3^2t_4+t_0t_5^2)]\}, \\
[\boldsymbol{W}]_{13,5} &= t_3^4+6t_2^2t_4^2+4t_0t_4^3+6t_1^2t_5^2+1^2t_2t_5(2t_1t_4+t_0t_5)+1^2t_3^2(t_2t_4+t_1t_5) \\
&\quad +1^2t_3[t_1t_4^2+(t_2^2+2t_0t_4)t_5], \\
[\boldsymbol{W}]_{14,5} &= 4\{t_3^3t_4+3t_2t_3^2t_5+3t_4(t_2^2+t_0t_4)t_5+t_1(t_4^3+3t_2t_5^2) \\
&\quad +3t_3[t_2t_4^2+t_5(2t_1t_4+t_0t_5)]\}, \\
[\boldsymbol{W}]_{15,5} &= 2[3t_3^2t_4^2+2t_2t_4^3+2t_3^3t_5+3t_2^2t_5^2+6t_4t_5(t_1t_4+t_0t_5)+6t_3t_5(2t_2t_4+t_1t_5)], \\
[\boldsymbol{W}]_{16,5} &= 4\{3t_3^2t_4t_5+t_3(t_4^3+3t_2t_5^2)+t_5[3t_2t_4^2+t_5(3t_1t_4+t_0t_5)]\},
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{W}]_{17,5} &= t_4^4 + 1^2 t_3 t_4^2 t_5 + 6 t_3^2 t_5^2 + 1^2 t_2 t_4 t_5^2 + 4 t_1 t_5^3, \\
[\boldsymbol{W}]_{18,5} &= 4 t_5 (t_4^3 + 3 t_3 t_4 t_5 + t_2 t_5^2), \\
[\boldsymbol{W}]_{19,5} &= 2 t_5^2 (3 t_4^2 + 2 t_3 t_5), \\
[\boldsymbol{W}]_{20,5} &= 4 t_4 t_5^3, \\
[\boldsymbol{W}]_{21,5} &= t_5^4, \\
[\boldsymbol{W}]_{1,6} &= t_0^5, \\
[\boldsymbol{W}]_{2,6} &= 5 t_0^4 t_1, \\
[\boldsymbol{W}]_{3,6} &= 5 t_0^3 (2 t_1^2 + t_0 t_2), \\
[\boldsymbol{W}]_{4,6} &= 5 t_0^2 (2 t_1^3 + 4 t_0 t_1 t_2 + t_0^2 t_3), \\
[\boldsymbol{W}]_{5,6} &= 5 t_0 [t_1^4 + 6 t_0 t_1^2 t_2 + 4 t_0^2 t_1 t_3 + t_0^2 (2 t_2^2 + t_0 t_4)], \\
[\boldsymbol{W}]_{6,6} &= t_1^5 + 20 t_0 t_1^3 t_2 + 30 t_0^2 t_1^2 t_3 + 10 t_0^2 t_1 (3 t_2^2 + 2 t_0 t_4) + 5 t_0^3 (4 t_2 t_3 + t_0 t_5), \\
[\boldsymbol{W}]_{7,6} &= 5 [t_1^4 t_2 + 4 t_0 t_1^3 t_3 + 6 t_0 t_1^2 (t_2^2 + t_0 t_4) + 2 t_0^2 (t_2^3 + t_0 t_3^2 + 2 t_0 t_2 t_4) \\
&\quad + 4 t_0^2 t_1 (3 t_2 t_3 + t_0 t_5)], \\
[\boldsymbol{W}]_{8,6} &= 5 [t_1^4 t_3 + 2 t_1^3 (t_2^2 + 2 t_0 t_4) + 2 t_0 t_1 (2 t_2^3 + 3 t_0 t_3^2 + 6 t_0 t_2 t_4) \\
&\quad + 6 t_0 t_1^2 (2 t_2 t_3 + t_0 t_5) + 2 t_0^2 (3 t_2^2 t_3 + 2 t_0 t_3 t_4 + 2 t_0 t_2 t_5)], \\
[\boldsymbol{W}]_{9,6} &= 5 \{ t_1^4 t_4 + 2 t_1^2 (t_2^3 + 3 t_0 t_3^2 + 6 t_0 t_2 t_4) + 4 t_1^3 (t_2 t_3 + t_0 t_5) \\
&\quad + 1^2 t_0 t_1 (t_2^2 t_3 + t_0 t_3 t_4 + t_0 t_2 t_5) + t_0 [t_2^4 + 6 t_0 t_2 t_3^2 + 6 t_0 t_2^2 t_4 \\
&\quad + 2 t_0^2 (t_4^2 + 2 t_3 t_5)] \}, \\
[\boldsymbol{W}]_{10,6} &= 5 \{ 2 t_1^3 (t_3^2 + 2 t_2 t_4) + t_1^4 t_5 + 6 t_1^2 (t_2^2 t_3 + 2 t_0 t_3 t_4 + 2 t_0 t_2 t_5) + t_1 [t_2^4 + 1^2 t_0 t_2 t_3^2 \\
&\quad + 1^2 t_0 t_2^2 t_4 + 6 t_0^2 (t_4^2 + 2 t_3 t_5)] + 2 t_0 [2 t_2^3 t_3 + 6 t_0 t_2 t_3 t_4 + 3 t_0 t_2^2 t_5 \\
&\quad + t_0 (t_3^3 + 2 t_0 t_4 t_5)] \}, \\
[\boldsymbol{W}]_{11,6} &= t_2^5 + 20 t_2^3 (t_1 t_3 + t_0 t_4) + 30 t_2^2 [t_1^2 t_4 + t_0 (t_3^2 + 2 t_1 t_5)] + 10 t_2 [3 t_1^2 t_3^2 \\
&\quad + 1^2 t_0 t_1 t_3 t_4 + 2 t_1^3 t_5 + 3 t_0^2 (t_4^2 + 2 t_3 t_5)] + 10 [2 t_1^3 t_3 t_4 + t_0^3 t_5^2
\end{aligned}
$$

$$
\begin{aligned}
& +3t_0^2t_4(t_3^2+2t_1t_5)+t_0t_1(2t_3^3+3t_1t_4^2+6t_1t_3t_5)],\\
[\boldsymbol{W}]_{12,6} &= 5\{t_2^4t_3+4t_2^3(t_1t_4+t_0t_5)+6t_2^2(t_1t_3^2+2t_0t_3t_4+t_1^2t_5)\\
& +4t_2[3t_1^2t_3t_4+3t_0^2t_4t_5+t_0(t_3^3+3t_1t_4^2+6t_1t_3t_5)]+2[3t_0^2t_3(t_4^2+t_3t_5)\\
& +t_1^3(t_4^2+2t_3t_5)+t_1^2(t_3^3+6t_0t_4t_5)+3t_0t_1(2t_3^2t_4+t_0t_5^2)]\},\\
[\boldsymbol{W}]_{13,6} &= 5\{t_2^4t_4+2t_2^3(t_3^2+2t_1t_5)+2t_1^2t_4(3t_3^2+2t_1t_5)+2t_0^2(t_4^3+6t_3t_4t_5)\\
& +t_0(t_3^4+1^2t_1t_3t_4^2+1^2t_1t_3^2t_5+6t_1^2t_5^2)+6t_2^2[2t_1t_3t_4+t_0(t_4^2+2t_3t_5)]\\
& +2t_2[3t_1^2(t_4^2+2t_3t_5)+2t_1(t_3^3+6t_0t_4t_5)+3t_0(2t_3^2t_4+t_0t_5^2)]\},\\
[\boldsymbol{W}]_{14,6} &= 5\{4t_2^3t_3t_4+t_2^4t_5+2t_1^3t_5^2+6t_1^2t_3(t_4^2+t_3t_5)+t_1(t_3^4+4t_0t_4^3+2^4t_0t_3t_4t_5)\\
& +2t_0(2t_3^3t_4+3t_0t_4^2t_5+3t_0t_3t_5^2)+2t_2^2[t_3^3+6t_1t_3t_5+3t_4(t_1t_4+2t_0t_5)]\\
& +1^2t_2[t_1^2t_4t_5+t_0t_3(t_4^2+t_3t_5)+t_1(t_3^2t_4+t_0t_5^2)]\},\\
[\boldsymbol{W}]_{15,6} &= 5\{2t_2^3(t_4^2+2t_3t_5)+6t_2^2[t_3^2t_4+t_5(2t_1t_4+t_0t_5)]+t_2[t_3^4+4t_0t_4^3+1^2t_1t_3^2t_5\\
& +6t_1^2t_5^2+1^2t_3t_4(t_1t_4+2t_0t_5)]+2[t_1^2(t_4^3+6t_3t_4t_5)+2t_1(t_3^3t_4+3t_0t_4^2t_5\\
& +3t_0t_3t_5^2)+t_0(3t_3^2t_4^2+2t_3^3t_5+3t_0t_4t_5^2)]\},\\
[\boldsymbol{W}]_{16,6} &= t_3^5+20t_3^3(t_2t_4+t_1t_5)+30t_3^2[t_1t_4^2+(t_2^2+2t_0t_4)t_5]+10(2t_1t_2t_4^3+2t_2^3t_4t_5\\
& +3t_1^2t_4^2t_5+6t_0t_2t_4^2t_5+3t_1t_2^2t_5^2+6t_0t_1t_4t_5^2+t_0^2t_5^3)+10t_3[3t_2^2t_4^2+2t_0t_4^3\\
& +3t_1^2t_5^2+6t_2t_5(2t_1t_4+t_0t_5)],\\
[\boldsymbol{W}]_{17,6} &= 5\{t_3^4t_4+2t_2^2t_4^3+t_0t_4^4+4t_2t_3^3t_5+2t_2^3t_5^2+6t_1^2t_4t_5^2+4t_0t_1t_5^3\\
& +1^2t_2t_4t_5(t_1t_4+t_0t_5)+6t_3^2[t_2t_4^2+t_5(2t_1t_4+t_0t_5)]+4t_3[3t_4(t_2^2+t_0t_4)t_5\\
& +t_1(t_4^3+3t_2t_5^2)]\}\mathrm{m}\\
[\boldsymbol{W}]_{18,6} &= 5\{2t_3^3t_4^2+t_3^4t_5+6t_2^2t_4^2t_5+4t_0t_4^3t_5+2t_1^2t_5^3+4t_0t_2t_5^3+6t_3^2t_5(2t_2t_4+t_1t_5)\\
& +t_1(t_4^4+1^2t_2t_4t_5^2)+2t_3[2t_2t_4^3+3t_2^2t_5^2+6t_4t_5(t_1t_4+t_0t_5)]\},\\
[\boldsymbol{W}]_{19,6} &= 5\{4t_3^3t_4t_5+6t_2^2t_4t_5^2+2t_4^2t_5(2t_1t_4+3t_0t_5)+2t_3^2(t_4^3+3t_2t_5^2)\\
& +t_2(t_4^4+4t_1t_5^3)+4t_3t_5[3t_2t_4^2+t_5(3t_1t_4+t_0t_5)]\},
\end{aligned}
$$

$$
\begin{aligned}
[\boldsymbol{W}]_{20,6} &= 5\{6t_3^2t_4^2t_5 + 2t_3^3t_5^2 + t_3(t_4^4 + 1^2t_2t_4t_5^2 + 4t_1t_5^3) + 2t_5[2t_2t_4^3 + t_2^2t_5^2 \\
&\quad + t_4t_5(3t_1t_4 + 2t_0t_5)]\}, \\
[\boldsymbol{W}]_{21,6} &= t_4^5 + 20t_3t_4^3t_5 + 30t_2t_4^2t_5^2 + 5t_5^3(4t_2t_3 + t_0t_5) + 10t_4t_5^2(3t_3^2 + 2t_1t_5), \\
[\boldsymbol{W}]_{22,6} &= 5t_5[t_4^4 + 6t_3t_4^2t_5 + 4t_2t_4t_5^2 + t_5^2(2t_3^2 + t_1t_5)], \\
[\boldsymbol{W}]_{23,6} &= 5t_5^2(2t_4^3 + 4t_3t_4t_5 + t_2t_5^2), \\
[\boldsymbol{W}]_{24,6} &= 5t_5^3(2t_4^2 + t_3t_5), \\
[\boldsymbol{W}]_{25,6} &= 5t_4t_5^4, \text{ and} \\
[\boldsymbol{W}]_{26,6} &= t_5^5.
\end{aligned}
$$