# USING RESPONSE TIMES IN CAT

BY

JUSTIN L. KERN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Hua-Hua Chang, Chair
Professor Carolyn J. Anderson
Associate Professor Steven A. Culpepper
Professor Jeffrey A. Douglas
Associate Professor Jinming Zhang

# Abstract

Many areas of psychology and education place a high premium on measurement, using psychometric theory to measure constructs, such as cognitive ability, personality, and attitudes. Some of the more well-known measurement theories used are classical test theory (CTT), structural equation modeling (SEM), and item response theory (IRT). For the practical test construction needs of psychology and education, IRT is the most heavily used, and has been ever since Lord and Novick (1968) published their book, Statistical Theories of Mental Test Scores.

One of the biggest advances in IRT has been the advent of computerized adaptive testing (CAT). First introduced as tailored tests by Lord (1980), CATs have increasingly gained in popularity as the cost of computation has gone down. As suggested by the term "tailored tests," every person who takes an adaptive test takes a test form unique to the person. The test is constructed item-by-item by matching items' difficulty levels to the ability level of that particular person. The promise of CAT is that by constructing a test in this way items that do not contribute much to the overall effectiveness of the measurement are left out, which can shorten the test substantially while still maintaining a high level of measurement accuracy.

The efficiency of CAT has not gone unnoticed. The Armed Services Vocational Aptitude Test Battery (ASVAB), which is used for measuring vocationally relevant abilities was originally introduced as a paper-and-pencil test in 1968, and became operational as a CAT in 1996; the ASVAB was the first large-scale, high-stakes operational CAT. Numerous adaptive tests have gone into operational use for use in selection including the Graduate Management Admission Test (GMAT), the Adjustable Competence Evaluation (ACE), the Business Language Testing Service (BULATS) Computer Test, the IBM Selection Tests, among others. Additionally, many licensure exams currently in use—including the Uniform CPA Examination (for certified personal accountants), and the National Council Licensure Examinations (for nurses)—are adaptive. Furthermore, the re-

cently signed Every Student Succeeds Act has recommended a greater use of adaptive testing in the American educational system, allowing states to develop and administer CATs.

Computer-based tests, such as CATs, allow for easy collection of response times. With the abundance of essentially-free data, methods and applications for using response time data have become en vogue, though they are still in their infant stage. As such, no large-scale assessments are currently using response times as an active part of the test. Because the data is essentially-free, it is reasonable to believe that their use is simply the next step in the evolution of computer-based tests. Indeed, it only seems natural that CATs be modified to take advantage of response time information, especially since it is well-known that response accuracy and response time are related (Sternberg, 1999). Some applications include cheating detection (van der Linden, 2009a), shortening the time needed to take a test (Choe & Kern, 2014; Fan, Wang, Chang, & Douglas, 2012), and item selection (van der Linden, 2008).

The goal of this dissertation is to introduce CAT and some of the current issues surrounding its use, to introduce response times in measurement, and several new methods for using response times in adaptive testing. In the first chapter, a quick review of IRT will be given, including its historical roots, its assumptions, and some examples of commonly used IRT models. Following this will be a brief overview of the basic components of CAT, including item selection, ability (or trait) estimation, and item constraints. I will then discuss response times in measurement and their current role in CAT. In the second chapter, I will describe an already-completed study on estimating person ability and speededness jointly. In Chapter 3, I investigate the efficacy of using the MAP estimator developed in Chapter 2 when selecting items using a generalized time-weighted maximum information criterion (GMICT). In Chapter 4, I introduce a new item selection technique based on the ideas of Bayesian item selection that incorporates the response time model directly. A modified version of this criterion using the ideas from the GMICT is also investigated. In Chapter 5, I introduce a time-weighted Kullback-Leibler information technique and investigate its effectiveness. Finally, I conclude in Chapter 6 with some remarks about how these techniques fit in in the current literature on response times, scoring, and adaptive testing.

*To all the awesome people in my life. There are many. You know who you are.*

# Acknowledgments

This dissertation has been a long time coming. I have spent the better part of my young adult life living in Champaign, and I wouldn't trade my time spent here for anything; the University of Illinois has truly become my home. During my time here, I have come to learn many things and take on many projects, but more importantly, I have had the privilege to know many wonderful people. Those people include my dissertation committee members: Carolyn Anderson, Steve Culpepper, Jeff Douglas, and Jinming Zhang. Thank you for your thoughtful comments and help through this process. Two others who were not on my committee, but who have influenced me greatly, are Larry Hubert and Sungjin Hong: Larry for all of his wonderful stories and insight into the world of quantitative psychology, and Sungjin for serving as my initial advisor and confidant. I would like to acknowledge my current advisor, Hua-Hua Chang, who has not only become a great friend and mentor, but truly a second father to me. All of the lessons that I learned from him—about working hard, looking forward, having patience, and, above all, not giving up—I will take these with me from now until forever.

My career would not have made it to this point without the moral and intellectual support of my fellow quantoids: Chun Wang, Nate Helwig, Ehsan Bokhari, Chris Zwilling, Edison Choe, Susu Zhang, and others. In particular, I would like to acknowledge Chun for introducing me to the world of measurement and response times, Edison for being my closest colleague, and Ehsan for having been probably my best friend in grad school and for reinvigorating my love for baseball. Other grad students in Dr. Chang's group have also been wonderful friends and colleagues of mine: Annie Kang, Yi Zheng, Chanjin Zheng, and Shiyu Wang. I always look forward to seeing all of you at conferences, and will continue to do so in the future.

My time spent in Champaign would not have been so great if it were not for the wonderful people outside of the grad school world that I have met. Two in particular—Sarah Knight and

Claire Schreibfeder—have become two of my absolute closest friends. It's hard to imagine a world without them. I spent many nights with them, and I wouldn't be who I am without them. Another friend of mine was my longtime roommate/"common-law husband", Tom Galvin. Let me just say, I never thought I would, but I miss the gators, especially Ron. I would also like to acknowledge the friends back home who stuck by me through everything. Liz Rodgers, Dan Stratton, and Shaun Rea, I have known you guys for far too long; thank you for existing, it's rather handy knowing I can call whenever I want, even if I don't.

I would like to thank my family for the unwavering support that they have given me. Going to and graduating from college isn't exactly a familiar experience in my family, but despite the strange looks I get when I talk about my work, they love me and support me anyway. I would especially like to acknowledge my dad, Paul Kern, for teaching me how to be a man. He taught me to "always look out for number one, but along the way, don't step in number two." Those are true words of wisdom.

Finally, I would like to acknowledge my two loves. My darling Koko, you are loved. I promise you will get all the cuddles and catnip after this is over. Kayla, I am grateful for you. You truly are my better half. I love the hell out of you, and I can't wait for what the future holds for us. Whatever it is, as long as we get to play music together, I'm down.

# Contents

# Chapter 1

# Introduction

## 1.1 Item Response Theory (IRT)

### 1.1.1 History

Any good discussion of IRT starts with how it came to be. For most of the 20th century, the predominant theory of measurement in psychology was what is now termed classical test theory (CTT). CTT was a merger of three "simple" ideas that had been developed early in the 1900s: measurements exist in the presence of error, error can be described as a random variable, and the relations (or correlations) among variables can be quantified. These ideas came together when Charles Spearman (1904) published a paper wherein he found that psychological measurements contain error from one trial to the next; he called this error "accidental." Spearman developed a method for estimating the reliability of a measurement, and, in accounting for this accidental error, how to use this reliability coefficient to correct a correlation coefficient attenuation. This has been argued to be the beginning of what came to be CTT (Traub, 1997). For the next 60 years, and culminating in the treatment given by Lord and Novick's (1968) book, CTT would develop into the dominant theory of measurement, with much of the topics within measurement dominated by the topics of reliability and validity along with a focus on sum-scores.

As important as the Lord and Novick (1968) book was for its treatment of CTT, it may have been equally as important or more important to the history of measurement overall as the starting point for IRT. In it were four chapters on IRT that were written by Allan Birnbaum. The basis for these chapters was given in the decade prior with a paper by Lord (1953) and three U.S. Air Force technical reports written by Birnbaum (1957, 1958a, 1958b). The very fact that the classical treatment of CTT and the beginnings of IRT appear in the same place is an interesting continuity.

However, even though it's not discussed at length in Lord and Novick (1968), IRT is often seen as a response to many of the shortcomings of CTT.

The most important of these is that the examinees and tests are fundamentally intertwined in that the interpretation of examinees' scores cannot be done without the context of the test characteristics, and vice versa (Hambleton, Swaminathan, & Rogers, 1991). A desirable property would be that an individual's score on an underlying, or latent, trait to be the same regardless of the assessment, but that's not the case in CTT, because of the reliance on sum-scores. As an example of this, consider a high-ability and a medium-ability examinee; one would expect the high-ability examinee to have a higher overall score on a trait than the medium-ability examinee. However, if the assessment given to the examinees was extremely easy, both would have a high probability of answering the items correctly and receive high scores. Similarly, if the high-ability examinee took a very hard assessment while the medium-ability examinee took a very easy assessment, then the sum-score for the high-ability examinee would actually be worse than the medium-ability examinee, which is, of course, troublesome. Thus, if this was a hiring or admissions situation, it would be possible to choose a lower-ability applicant simply by virtue of the assessment. Thus, we can see that the scores on a test in CTT are test-dependent. Conversely, classical item indices are difficulty and discrimination, where difficulty is the proportion of examinees that got an item correct and discrimination is the point-biserial correlation (the correlation between the scores on an item and the overall score on the test). These indices are group-dependent. IRT, on the other hand, is invariant to the group and the assessment, so as long as the scaling of the items is done correctly, latent trait scores will be comparable (no dependence on group or test). This property is *extremely* important for CAT.

One other oft-cited shortcoming of CTT is the conceptualization of reliability and the standard error of measurement (Bock, 1997; Hambleton et al., 1991). Reliability, as it is normally defined, is the squared correlation between test scores and true scores. Unfortunately, true scores are unknown, and so while this definition is conceptually pleasing, it is not computationally useful. CTT shows us that the reliability coefficient is also equal to the correlation of test scores on parallel tests. The property of tests being parallel (see Lord, 1980) is very difficult to satisfy, so while this is more useful than the definition, it is still not computationally useful. The way this has been

handled in CTT is to compute lower bounds to reliability (Cronbach, 1951; ten Berge & Zegers, 1978), which can be quite unpleasing to some (using an estimate of a lower bound of reliability as an estimate to reliability, particularly when distributional properties are not known well, is an unsettling prospect). This becomes a much larger problem when it is noted that the precision of a measurement, the standard error of measurement, is a function of reliability; without a reasonable estimate to reliability, the precision for a measurement is difficult to know well. Furthermore, the standard error of measurement is also a function of the variance of the test scores, which is assumed constant in a group. However, the scores on an assessment are not equally precise across the entire spectrum of scores. With these problems taken together, clearly the concept of reliability in CTT is somewhat less than satisfying. IRT handles these issues with its concept of information. Every item contains information which is a function of the value of the latent trait. The sum of item information is test information, which is, by definition, also a function of the value of the latent trait. Taken together, it is possible to engineer an assessment that estimates particular areas of the scale well. This property of information is also very important for CAT.

### 1.1.2 Standard Assumptions of IRT

IRT can be seen as a set of very general assumptions from which important mathematical properties can be derived, or as a set of models with common properties, not just simply a set of disparate models (Lord, 1980). Essentially, if a model meets these assumptions, then the mathematical properties derived in IRT can be assumed to also apply. Extensions in IRT are often simply ways of either dealing with violations of assumptions or allowing for alternatives to the assumptions altogether. Two common assumptions made are unidimensionality and local independence (Hambleton et al., 1991).

Unidimensionality generally refers to the assessment only measuring one latent trait. While this is not, in general, met exactly by any set of items, the hope is that a single trait would have a greater presence than any others. While many of the standard IRT models make an assumption of unidimensionality, it is not always the case. A model that assumes that multiple latent traits are needed for an assessment is called multidimensional.

Local independence, on the other hand, refers to the examinee responses. Specifically, it means

that given a value for the latent trait, an examinee's responses to items are statistically independent of each other. One attractive interpretation is that the relationships among the items are entirely explained by the latent trait. This is an important property for estimation, because independent responses make maximum likelihood estimation mathematically tractable. An interesting point is that if the assumption of unidimensionality is met, then the assumption of local independence is also met, but not vice versa (Lord & Novick, 1968). The majority of IRT models make the local independence assumption, but usually when they don't, some form of "lower-order" local independence is generally assumed. For instance, items that are presented to an examinee following a reading passage are not assumed to be locally independent from each other. However, the response pattern of those items jointly are considered independent of the responses on all other items.

### 1.1.3 Examples of IRT Models

In general, IRT models are considered to be models of categorical data with a continuous latent variable. While the data could be considered dichotomous (having only two response categories, 0 or 1), or polytomous (having $k$ response categories, $0, 1, 2, \ldots, k-1$) with the possibility of polytomous data being ordinal, this section will only focus on dichotomous item response models. For an introduction to polytomous IRT models, see van der Linden and Hambleton (1997). Here, I will discuss the standard unidimensional logistic models, their multidimensional counterparts, and an unfolding model.

#### Logistic models

The most commonly used models in IRT are the so-called logistic models. The are the 1PL (one-parameter logistic, or Rasch model), 2PL (two-parameter logistic), and 3PL (three-parameter logistic), named for the number of item parameters in the model. These are unidimensional IRT models that model dichotomous responses.

Suppose that $x_{ij}$ is the $j$th examinee's response to item $i$. Then, since $x_{ij}$ is dichotomous, it can take on the values 1, if correct, or 0, if incorrect. In the case of a subjective response where the latent trait is psychological rather than ability, then an assignment of 1 and 0 would be arbitrary, but often would correspond agree and disagree, respectively. Then, according to the 3PL model,

the probability of answering correctly (or responding with agreement) to the $i$th item given the level of the latent trait $\theta$ of the $j$th examinee is

$$p(x_{ij} = 1|\theta_j) = P_i(\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}}, \tag{1.1}$$

where $a_i$, $b_i$, and $c_i$ are the discrimination, difficulty, and pseudo-guessing parameters for the $i$th item, respectively.

The item parameters have nice interpretations. The difficulty parameter $b_i$ is a location parameter; it is the point on the trait scale where the probability of responding positively is $\frac{1+c_i}{2}$. For a fixed value of $\theta$, as the value of $b$ increases, the probability of a positive response decreases, and vice versa, hence the difficulty interpretation. For a subjective response, a difficulty interpretation is less intuitive, and so the parameter is simply the location of the item. The discrimination parameter $a_i$ is proportional to the slope of the curve $P_i(\theta_j)$ at $b_i$ on the trait scale. Essentially, a larger value of $a_i$ means that the item distinguishes between levels of the trait near $b_i$ more easily; that is, a small increase in $\theta$ yields a larger increase in the probability of responding positively as $a_i$ increases. The pseudo-guessing parameter $c_i$ is simply the lower bound to the probability of a positive response, i.e., as the trait level $\theta$ decreases, the probability of responding positively approaches $c_i$. Two other interpretations of the discrimination parameter can also be given. First, consider the interpretation that $P_i$ is a mixture model of a "guessing" group (with probability $c_i$), and a "non-guessing" group (with probability $1 - c_i$). If an examinee is in the "non-guessing" group, then $\exp(a_i)$ is the increase in the odds ratio of answering correctly (or giving a positive response) for a one-unit increase in the value of $\theta$. Another interpretation is that $a_i$ is the slope on the logit scale (of the examinees in the "non-guessing" group).

The 1PL and 2PL models can both be considered as special cases of the 3PL model in (1.1). The 2PL model is the 3PL model where all values of $c_i$ are equal to zero. The 1PL (Rasch) model is the 2PL model where all values of $a_i$ are equal to each other. Without loss of generality, the discrimination parameters are usually all set to one. More details on the logistic models can be found in any introduction to IRT (Drasgow & Hulin, 1990; Embretson & Reise, 2000; Hambleton et al., 1991; Lord, 1980; Lord & Novick, 1968).

**Multidimensional models**

Multidimensional IRT (MIRT) models allow for the case where an assessment measures multiple latent traits simultaneously. As such, they do not have an assumption of unidimensionality, though they do assume that the number of dimensions in the model is correct for the assessment at hand. A personality assessment is a classic example of a multidimensional assessment. MIRT models come in two forms: compensatory and partially compensatory. Essentially, a compensatory model is compensatory in the sense that a lower value on one latent trait can be offset by a higher value on another latent trait, so that the probability of responding positively stays the same; this compensation property is a linear relationship. A partially compensatory model is a model where the latent trait compensation property is more complex than in a conpensatory model. Because of space constraints and the limited usefulness of the partially compensatory models, only the compensatory model will be discussed. For more information on partially compensatory models, as well as MIRT in general, see Reckase (2009). Further information can also be found in the related literature of item factor analysis (Bock, Gibbons, & Muraki, 1988; Gibbons & Hedeker, 1992; Knol & Berger, 1991).

The unidimensional logistic IRT models can be generalized very easily to the multidimensional case. In (1.1), the part in the exponent $a(\theta - b)$ can be expanded as $a\theta - ab = a\theta + d$, where $d = -ab$. Then, the exponent can be expressed in a simple slope-intercept form. To expand multidimensionally, the exponent can simply be expanded by including more latent traits and corresponding weights as follows:

$$a_1\theta_1 + a_2\theta_2 + \cdots + a_m\theta_m + d = \left(\sum_{k=1}^{m} a_k\theta_k\right) + d = \mathbf{a}'\boldsymbol{\theta} + d, \tag{1.2}$$

where $\mathbf{a}$ is an $m \times 1$ vector of the $a$-parameters, $\boldsymbol{\theta}$ is an $m \times 1$ vector of the latent trait values, and $d$ is a constant intercept term. Finally, by replacing the exponent in (1.1) with (1.2), we have the multidimensional version of the 3PL model:

$$p(x_{ij} = 1|\boldsymbol{\theta}_j) = P_i(\boldsymbol{\theta}_j) = c_i + (1 - c_i)\frac{1}{1 + \exp(-(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i))}. \tag{1.3}$$

Here, the $\mathbf{a}$-parameter is called the discrimination parameter, in much the same way that is was for

the unidimensional logistic models. For the $k$th dimension of the person-space, $a_k$ is proportional to the slope of the curve $P_i(\boldsymbol{\theta}_j)$ on the line $0 = \mathbf{a}_i'\boldsymbol{\theta}_j + d_i$ in the direction of the $k$th dimension. Equation 1.2 can be reformulated as

$$\begin{aligned}\left(\sum_{k=1}^{m} a_{ik}\theta_{jk}\right) + d_i &= a_{i1}(\theta_{j1} + {}^{d_i}/_{a_{i1}}) + a_{i2}(\theta_{j2} + {}^{d_i}/_{a_{i2}}) + \cdots + a_{im}(\theta_{im} + {}^{d_i}/_{a_{im}}) \\ &= \sum_{k=1}^{m} a_{ik}(\theta_{jk} + {}^{d_i}/_{a_{im}}) \\ &= \sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_{ik}),\end{aligned}$$

where $b_{ik} = -{}^{d_i}/_{a_{ik}}$. As we see here, this is the same form as the exponent part in (1.1), and so $-{}^{d_i}/_{a_{ik}}$ can be interpreted as the relative difficulty of item $i$ with respect to dimension $k$. The parameter $c_i$ is exactly the same for the multidimensional 3PL model as with the unidimensional 3PL model, and is used for modeling guessing in an cognitive assessment.

Multidimensional counterparts of the 1PL and 2PL models are special cases of the multidimensional 3PL model. To obtain multidimensional 2PL, the parameter $c_i$ in (1.3) is set to zero for all items. The multidimensional 1PL is obtained by further constraining the **a**-parameters to be the same across all items; the **a**-parameters can be set to **1** for all items, without loss of generality.

**Unfolding model**

An underlying assumption of the logistic IRT models is that as the level of the latent trait increases, the probability of responding positively on an item goes up. Mathematically, this is the same as assuming that the item response function $P_i(\theta_j)$ is monotonic increasing. As Drasgow, Chernyshenko, and Stark (2010) pointed out, this is tantamount to what Coombs (1964) described as the dominance process. Essentially, an individual can be said to "dominate" an item if the individual's latent trait score is higher than the difficulty of the item on the latent trait scale.

While the dominance process is intuitive for cognitive assessments where answers are right or wrong, it may not be appropriate for attitudinal questions. In this case, the ideal-point process described by Coombs (1964) may be more fitting. The ideal-point process envisions a person and an item on the same scale (the latent trait scale). A person's location on the scale is his "ideal

point" where he endorses an item with maximum probability. As the distance of the item from the person increases, the probability of that person endorsing that item decreases. The novel idea here is that there is no difference, probabilistically, between a person's location being to the left or to the right of the item on the scale, so long as the distance remains the same. As such, the form of the item response function $P_i(\theta_j)$ is not monotonic, but increases to a maximum and then decreases after that point. Furthermore, the item response function is symmetric.

The development of these so-called unfolding models is still in its infancy, and so there is no one dominant model. Examples of binary unfolding models are given by Andrich (1988), Hoijtinck (1990), and Andrich and Luo's (1993) hyperbolic cosine model. The following polytomous unfolding models are generalizations of the hyperbolic cosine model: the generalized hyperbolic cosine model (Andrich, 1996), the generalized unfolding model (GUM; Roberts, 1995), and the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 1998). Finally, a few multidimensional unfolding models have been proposed. One, the normal PDF model, was given by Maydeu-Olivares, Hernández, and McDonald (2006). Finally, the Tailored Adaptive Personality Assessment System (TAPAS) uses the GGUM along with the multidimensional pairwise preference model (MDPP; Stark, 2002) for use in Army selection and classification (Drasgow, Stark, Chernyshenko, Nye, & Hulin, 2012). While there is much work to be done on unfolding models, its potential utility for attitude measurement is great.

## 1.2 Computerized Adaptive Testing (CAT)

Computerized adaptive testing (CAT) was first introduced by Lord (1980) as tailored testing. The idea is that by capitalizing on the additivity of item information, the non-uniformity of the standard error of measurement (and item information) across the scale of the latent trait, and the invariance properties of the assessment to the group (and of the group to the assessment) the IRT provides, tests can be individually constructed to be highly efficient for any given examinee. This is in direct contrast to the standard linear test in which all examinees take the same form of the test, or a parallel form of the test. To accomplish this goal, two components are necessary: an ability estimation procedure, and an item selection mechanism. Furthermore, many assessments have requirements not intrinsic to the measurement itself (e.g., item types, content constraints,

item key sequence, etc.). These are called item constraints.

### 1.2.1 Ability Estimation

Ability estimation is simply the process of obtaining an estimate for the value of the latent trait $\theta$. Other names for ability estimation are latent trait estimation and scoring. The standard ability estimates are the maximum likelihood estimate (MLE), and the Bayesian estimates, expected a posteriori (EAP) and maximum a posteriori (MAP). Maximum likelihood estimation relies heavily on the assumption of local independence. Its use is well-established in the statistical literature.

First, the joint distribution of all the items given a value of the latent trait is determined. With local independence assumed, this is just

$$f(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

where $\mathbf{x}$ is an $n \times 1$ vector containing the item responses. Now, the function $f$ is considered from the perspective of the data being fixed, rather than the latent trait value being fixed. That is,

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

In this perspective, the function $L$ is to be maximized with respect to $\theta$. As a simplification, the natural logarithm of the function $L$ is usually taken:

$$\log L(\theta|\mathbf{x}) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

For an item response function, this amounts to finding the maximum of

$$\begin{aligned} \log L(\theta|\mathbf{x}) &= \log \prod_{i=1}^{n} P_i(\theta)^{x_i} Q_i(\theta)^{1-x_i} \\ &= \sum_{i=1}^{n} \left[ x_i \log P_i(\theta) + (1 - x_i) \log Q_i(\theta) \right], \end{aligned} \tag{1.4}$$

where $Q_i(\theta) \equiv 1 - P_i(\theta)$. That maximum is the MLE $\hat{\theta}$. Equivalently, the MLE $\hat{\theta}$ solves the

maximum likelihood estimating equation

$$U_n(\theta) = \frac{\partial}{\partial\theta} \log L_n(\theta|\mathbf{x}) = \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial\theta} \log \frac{P_i(\theta)}{Q_i(\theta)} \right\} [x_i - P_i(\theta)] = 0.$$

The gold standard in test scoring has tended to be using the maximum likelihood estimate (MLE). This can be attributed to two properties of the MLE. First, MLE is an unbiased estimator of ability for long tests (Lord, 1983). In testing, this is generally seen as a desirable property, because we would like the estimated scores to be as indicative of the true scores as possible, especially in cases when the scores are used to determine pass-fail, such as in licensure testing. Second, we know the asymptotic properties of the MLE; that is, it is distributed normally, with a mean $\theta$ (the true ability) and variance as the inverse of the Fisher information of $\theta$. Thus, with long tests, we know the behavior of estimator quite well. However, there are some downsides to the MLE as well. The most well-known is that in some cases, the likelihood may only have a maximum at a boundary. In this case, the MLE for $\theta$ does not exist. One well-known example of when this happens is when an examinee answers all items correctly or all items incorrectly, which is not an uncommon occurrence.

One answer for this is to assume a known distribution $g(\theta)$ for the latent trait a priori; this distribution is known as the prior distribution. The prior distribution, along with the likelihood $f(\mathbf{x}|\theta)$ are used in the well-known Bayes' Theorem to obtain the posterior distribution as follows:

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)g(\theta)\ d\theta}. \tag{1.5}$$

An estimate for $\theta$ can then be obtained in one of two common ways from the posterior distribution $f(\theta|\mathbf{x})$. If one takes the expectation (i.e., the mean) of the posterior distribution as the estimate of $\theta$, then the estimate is known as the EAP. If on the other hand the maximum (i.e., the mode) of the posterior distribution is taken as the $\theta$-estimate, then the estimate is called the MAP. One argument for using the EAP over the MAP is that the MAP requires a proper continuous density function for the prior, whereas the EAP can use a discrete prior (Bock & Aitkin, 1981; Bock & Mislevy, 1982). Also, the EAP does not require the calculation of derivatives and is non-iterative in nature. The MAP estimator, on the other hand, has several properties that are important. First and foremost, the MLE can be considered a special case of the MAP when the prior distribution

is uniform. As such, with a non-uniform prior, it can be considered to be a "regularized" form of MLE, and is thus able to avoid the likelihood issues present in MLE. In some sense, the prior acts as an extra observation which will tend to make the likelihood behave. Secondly, the MAP is a convex combination of the MLE and the prior. Essentially, the MAP "shrinks" the estimate towards the prior; this can help the estimate to avoid overfitting to the data, which can be especially helpful for shorter tests. In comparison to the EAP, the MAP is very similar, but is simpler to compute, especially in multiple dimensions.

These are not the only ability estimation techniques, but these are the most standard. More information on ability estimation can be found in Baker and Kim (2004), Fox (2010), and Thissen and Wainer (2001).

### 1.2.2 Item Selection

Item selection in CAT refers to the mechanism used to select the next item to administer to the examinee in real time. The general flow of CAT is the the examinee receives an item, he responds to that item, the level of his latent trait is estimated, and based on the current estimate of his latent trait score, the next item is selected for administration. This process is continued until some stopping criterion is hit, usually either a specified number of items (a fixed-length test) or a prescribed level of the test information/measurement precision (a variable-length test). Clearly, the item selection procedure is very important to this process.

The two standard item selection procedures are the maximum information criterion and the maximum Kullback-Leibler (K-L) information criterion. Central to both of these is the concept of item information. Item information is defined as the Fisher information of the item; that is,

$$I_i(\theta_j) \equiv -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_j^2}\ln(P_i(\theta_j))\middle|\theta_j\right] = \frac{P_i'(\theta_j)^2}{P_i(\theta_j)(1-P_i(\theta_j))}. \tag{1.6}$$

Equation 1.6 is very general, and applies to all IRT models. One specific example is the item information function for the 3PL, which is given as

$$I_i(\theta_j) = \frac{a_i^2(1-c_i)}{[c_i+\exp(a_i(\theta_j-b_i))]\,[1+\exp(-a_i(\theta_j-b_i))]^2}. \tag{1.7}$$

The sum of the information for $n$ individual items is the test information, $I^{(n)}(\theta_j)$; that is,

$$I^{(n)}(\theta_j) = \sum_{i=1}^{n} I_i(\theta_j). \tag{1.8}$$

It can be shown that an ability estimate $\hat{\theta}$ is, under regularity conditions, asymptotically normally distributed with mean $\theta_0$ (the true ability value) and variance $I^{(n)}(\theta_0)^{-1}$ in both linear paper-and-pencil tests (Lord, 1980) and in CAT (Chang & Ying, 2009). Thus, as information increases, the variability in the estimate decreases, i.e., the precision of the estimate increases. It should be clear that to have as "good" an estimate as possible, information needs to be high. As $\theta_0$ cannot be known in reality, the value of $I^{(n)}(\hat{\theta})^{-1}$ is used as an estimate of the variance of the estimate $\hat{\theta}$.

The maximum information criterion says that the next item to administer should be chosen such that (1.8) is maximized for the current estimate of $\theta$. Since test information is additive, this simply amounts to finding the item where item information is highest. As argued by Chang and Ying (1996), the value of the item information function at $\theta$ can be considered as local information. Thus, the maximum information criterion makes most sense when the current estimate of the latent trait $\hat{\theta}$ is close to the actual value of $\theta$. But when it is not, the chosen item is not optimal for the examinee, and so the search for the correct value of $\theta$ is weakened. This is particularly true for the early stages of CAT when few items have been administered.

To remedy this situation, a "global" criterion was proposed by Chang and Ying (1996) using the K-L information criterion. They argue that the log-likelihood ratio is a best quantity that can be used for this purpose; correspondingly, the K-L information function is a "distance" function (more appropriately, a "divergence" function, because of its non-symmetricity) between the log-likelihoods of two values, $\theta$ and $\theta_0$. It is defined as follows:

$$K_i(\theta \| \theta_0) = \mathrm{E}\left[\log\left(\frac{L_i(\theta_0|X_i)}{L_i(\theta|X_i)}\right)\right] = P_i(\theta_0)\log\left(\frac{P_i(\theta_0)}{P_i(\theta)}\right) + Q_i(\theta_0)\log\left(\frac{Q_i(\theta_0)}{Q_i(\theta)}\right), \tag{1.9}$$

where $\theta_0$ is the true parameter, $\theta$ is an estimate of $\theta_0$, and $L_i(\theta|X_i)$ is the likelihood function for the $i$th item. Interestingly, we find that the second derivative of the K-L information function at $\theta_0$ is the item information function, and so, geometrically, item information is the curvature of K-L information at the point $\theta_0$. In practice, we do not know the actual value of a person's latent trait,

so to use this, we need an index. A simple index can be the average of the area under the curve $K$ of an appropriate interval about the estimate $\hat{\theta}$

$$K_i(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} K_i(\theta \| \hat{\theta}) \, d\theta \tag{1.10}$$

where $\delta$ is a control to determine the interval size about $\hat{\theta}$. The maximum K-L information criterion, then, is to simply choose the next item such that (1.10) is maximized. The downside to the K-L index is that it is more computationally intensive than item information. To deal with this in practice, K-L information is only used at early stages of the test when its use is most crucial, switching to item information in the later stages when the estimate of $\theta$ is closer to the true value.

A third strategy for choosing items is used when estimating $\theta$ in a Bayesian manner; that is, to use the prior in selecting items. This idea was first proposed by Owen (1969, 1975) where he used an approximate empirical Bayes procedure, estimating $\theta$ with the EAP with a normal prior. The next item is chosen to satisfy the criterion

$$|b_{i_k} - \mathrm{E}(\theta|\mathbf{x})| < \delta, \tag{1.11}$$

where $i_k$ is the index of the item in the administered as the $k$th item in the test, and $\delta > 0$. Note that $\mathrm{E}(\theta|\mathbf{x})$ is the EAP. It was proposed to stop this procedure when the posterior variance is smaller than some threshold. Owen also proposed minimizing the preposterior risk under a quadratic loss to select items. This is equivalent to choosing items that minimize the expected posterior variance. At the time, this was an infeasible method for item selection, due to computational constraints at the time.

Later, van der Linden (1998) proposed a series of Bayesian item selection criteria, reasoning that the computational constraints present during Owen's time were no longer a concern. He found that the minimum expected posterior variance (MEPV) criterion performed the best, outperforming (or having similar performance as) the other criteria he suggested (viz., maximum posterior-weighted information, maximum expected information, and maximum expected posterior-weighted information) and the standard maximum information criterion. The minimum expected posterior variance

criterion chooses the $k$th item as the one that minimizes the expected posterior variance. That,

$$i_k \equiv \min_j \left\{ \sum_{m=0}^{1} p_j(X_j = m|x_1, \ldots, x_{k-1}) \mathrm{Var}(\theta|x_1, \ldots, x_{k-1}, X_j = m) : j \in R_k \right\} \tag{1.12}$$

where $p_j(X_j = m|x_1, \ldots, x_{k-1})$ is the posterior predictive distribution

$$p_j(X_j = m|x_1, \ldots, x_{k-1}) = \int p_j(X_j = m|\theta) g(\theta|x_1, \ldots, x_{k-1}) \, d\theta, \tag{1.13}$$

and $R_k$ is the set indices of the items remaining in the item pool. It turns out that the MEPV criterion is especially effective for short tests. As the test length increases, the choice of the item selection criterion has little effect on test efficiency. This is to be expected, as the asymptotic variance of the Bayesian estimate is just the inverse of the test information (Chang, 1996; Chang & Stout, 1993). Of course, most adaptive tests are designed to be shorter than when an asymptotic result may reasonably be expected to hold.

### 1.2.3 Item Constraints

Item constraints affect the construction of the test by altering the selection of items that would otherwise be selected. The general issue comes from a need to satisfy the constraints by the completion of the assessment while maintaining a level of optimality in terms of measurement efficiency for the individual. While item constraints are fairly straightforward to incorporate in a linear test, because all items are selected simultaneously, its implementation in CAT requires some care. Two methods to handle item constraints will be discussed here.

The first method is called the shadow test method (van der Linden & Glas, 2010; van der Linden & Reese, 1998). The idea is quite straight-forward. For each item, an entire assessment is constructed such that all of the constraints are met while maximizing efficiency for the current latent trait estimate $\hat{\theta}$; this test is called a shadow test. The item in the shadow test with maximum information at $\hat{\theta}$ is administered. After the examinee responds to the item, the estimate of the latent is updated. A new shadow test with the administered item included is assembled; with every iteration, all previously administered items are included in the shadow test. This process is iterated until the required number of items have been administered. Underlying the construction of the

shadow tests is a linear integer programming algorithm. Integer programming allows for an exact solution to the problem if it exists, though an exact solution is not guaranteed; when no solution exists, the shadow test method will produce an error. Furthermore, integer programming can be computationally intensive and difficult to implement in real time.

Another method is the maximum priority index (MPI) method (Cheng & Chang, 2009). In essence, it achieves item constraints by modifying the item information index with a multiplier that incorporates item constraints to create a new "attractiveness" measure, the priority index. Suppose there are $K$ constraints to be considered in the assessment and $I$ items in the item pool. Clearly not all constraints are relevant to every item. An $I \times K$ matrix $\mathbf{C}$ can be constructed that contains the information of which constraints are relevant to which items by setting $c_{ik} = 1$ if the the $k$th constraint is relevant to the $i$th item, and $c_{ik} = 0$ otherwise. This constraint relevancy matrix can be constructed a priori by the appropriate content experts. For each constraint, there is an associated weight $w_k$ which denotes the relative importance of that particular constraint to the assessment. Finally, the proportion of items still needed to satisfy to the $k$th content constraints of the assessment is constructed as

$$f_k = \frac{X_k - x_k}{X_k},$$

where $X_k$ is the number of items required to satisfy constraint $k$ and $x_k$ is the number of items that have been administered to satisfy constraint $k$; $f_k$ is reconstructed following each adminstration. Then, the priority index is

$$\mathrm{PI}_i = I_i(\hat{\theta}) \prod_{k=1}^{K} (w_k f_k)^{c_{ik}}, \tag{1.14}$$

where $I_i(\hat{\theta})$ is the item information for item $i$ at the current latent trait estimate $\hat{\theta}$. The MPI method then states that the next item chosen for administration is the item that maximizes $\mathrm{PI}_i$. This method is able to handle a large number of constraints simultaneously with little extra computational load. Furthermore, unlike the shadow test method it does not, in general, offer exact solutions. However, it will always offer a solution. Also, since it does not require the use of a sophisticated integer programming software, it is more easily implementable in current operational

CATs than the shadow test method.

## 1.3 Issues and Extensions in CAT

CAT has come a long way over the past 30 years. Many of the advancements in CAT have come as advancements in IRT have been developed. Here, some issues and extensions in CAT will be discussed briefly. First, the issue of bias in measurement, also known as differential item functioning (DIF) and as measurement invariance, will be examined. This issue arises as an item has different properties for different groups. Two extensions of CAT will also be examined, building on the discussion of multidimensional and unfolding IRT models.

### 1.3.1 Differential Item Functioning (DIF)

DIF refers to the situation in which the probabilities of two different groups answering the same way are different when their latent trait values are the same. That is, there is a value $\theta_j$ such that

$$P_{1i}(\theta_j) \neq P_{2i}(\theta_j),$$

where the numbered subscript refers to the group assignment. In other words, the item parameters for item $i$ are not the same across the groups. The detection of DIF is a well-studied topic, with several tests having been proposed. For dichotomous data, tests include the standardization method (Dorans & Kulick, 1986), Mantel-Haenszel (Holland & Thayer, 1988), and SIBTEST (Shealy & Stout, 1993). For polytomous data, generalizations of the standardization method, Mantel-Haenszel, and SIBTEST exist: the standardized mean difference method (Dorans & Schmitt, 1993), the Mantel procedure (Zwick, Donoghue, & Grima, 1993), and Poly-SIBTEST (Chang, Mazzeo, & Roussos, 1996), respectively. A detailed discussion of their similarities and differences can be found in Roussos and Stout (1996) and Chang et al. (1996).

While DIF has been studied extensively in the context of linear tests, little work has been done on DIF in the context of CAT, even though, as argued by Makransky and Glas (2013), DIF is even more important when examinees receive different items on a test, like in CAT. This is because when items display DIF, the act of giving the items to different examinees creates bias within and

between groups, rather than just between groups in a linear test. Also, items that display DIF could affect the item selection procedure and, thus, be administered to examinees incorrectly and ruining the optimality of the CAT. Furthermore, the items selected are partially dependent upon the previous items (Mislevy & Chang, 2000), and so an item adminstered incorrectly due to DIF could potentially have disastrous effects. Additionally, the very fact that fewer items are given in CAT, any single item is more important to examinee scores than in linear tests, and, thus, any given item exhibiting DIF will have a more negative effect in CAT than in linear tests (Zwick, 2010). If items are not shown to be without DIF, then a CAT used for selection purposes would clearly not be legally defensible, and, thus, could leave any organization using the test open to lawsuits. This is, of course, also the case for linear tests, but is simply magnified in the CAT scenario. After investigating DIF, and DIF items have been identified, those items do not necessarily have to be thrown out of the item pool, but can simply have differing sets of item parameters for each group, essentially creating multiple "pseudo-items" out of a single item. It has been shown that this approach is not detrimental to the overall assessment (Makransky & Glas, 2013). Some methods for assessing DIF in CATs include the method by Zwick, Thayer, and Wingersky (ZTW; 1994), and by Zwick, Thayer, and Lewis (ZTL; 1999, 2000); the ZTW methods are based on a modification of the Mantel-Haenszel test, whereas the ZTL methods are based on an empirical Bayes-enhanced version of the Mantel-Haenszel test. An in-depth review of DIF methods can be found in Zwick (2010).

### 1.3.2 Test Security

One large concern in computerized testing that frequently comes up is the security (and confidentiality) of a test. For professional test-makers and test-users, this is a codified standard. According to the "Standards for Educational and Psychological Testing" (2014), "Test users have the responsibility of protecting the security of the test materials at all times" (Standard 5.7), and "Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means" (Standard 5.6). In the perspective of testing companies, a large importance is placed on minimizing the cost of test development. One way this can be done is with the efficient use of items. This concern is in line with the goals of

test security; if items are used inefficiently—that is, that some items are used more frequently than others—then the chance of these items being compromised, or stolen by thieves, increases (Yi, Zhang, & Chang, 2008). In the continuous testing mode of CAT, efficient use of items is particularly of concern, since the method of selecting items may not necessarily guarantee a uniform distribution of item usage.

To quantify test security, two indices are commonly used. One is the item exposure rate, and the other is the test overlap rate (Chang & Ying, 1999; Way, 1998; Yi et al., 2008). Intuitively, these make sense. As the exposure rate, defined as the ratio between the number of times an item is administered and the total number of examinees, for an item increases, the more likely that item is to be remembered by thieves and eventually compromised. Similarly, as the test overlap rate (which is the average number of overlapping items encountered by a group of examinees over a given test length) increases, the more likely that an item is to be remembered by a group of thieves, compromising the item.

Particularly for the exposure rate, it is the case that the best distribution of item usage is the uniform distribution, which can be achieved by random item selection. This is, of course, not a tenable option under an adaptive testing system, but it does serve as a baseline for measuring a best exposure rate for a test. For a test of length $L$ with an item bank of size $N$, the uniform rate is simply $L/N$. A simple value quantifying the differences between this baseline and the actual exposure rate for $n$ examinees

$$r_i = \frac{\text{number of times the } i\text{th item is used}}{n}$$

is just the scaled $\chi^2$ (Chang & Ying, 1999)

$$\chi^2 = \sum_{i=1}^{N} \frac{r_i - L/N}{L/N}. \tag{1.15}$$

Essentially, $\chi^2$ measures the amount of departure from the uniform distribution, or skewness, of item exposure. The higher the value, the more skewed the item exposure distribution is and the more likely those items are to be compromised.

For test overlap rate, a baseline can also be constructed. Chang and Zhang (2002) argue that a

reasonable benchmark for comparison is the lower bound for the item overlap rate, which they also refer to as *item sharing*, as higher item overlap rates serve as evidence of skewed item exposure rates. Furthermore, as test overlap rate is particularly sensitive to item selection method, ability estimation method, and exposure control method, a theoretical lower bound over all methods can serve as a useful benchmark. It turns out that this benchmark can be achieved like the exposure rate, that is, assuming a random item selection. In Chang and Zhang (2002), a recursive formula based on the hypergeometric distribution for this lower-bound is found. In the special case of two examinees, the lower bound is simply $\frac{L^2}{N}$, where $L$ is the test length and $N$ is the item bank size. They also define a related concept called *item pooling*. Item pooling refers to the information pooled from several former examinees; it can be indexed by the number of overlapping items encountered by an examinee with a number of other examinees who have already taken the test. They also find a lower-bound for the item pooling index. In the special case where an examinee pools information from just one former examinee, this is equivalent to the item overlap rate between two examinees, and the lower bound is, again, $\frac{L^2}{N}$.

Several methods for handling the problem of a skewed item bank have been proposed, such as the Sympson-Hetter method (Sympson & Hetter, 1985), $a$-stratification (Chang & Ying, 1999), $a$-stratification with $b$ blocking (Chang, Qian, & Ying, 2001), and the maximum priority index method (Cheng & Chang, 2009). In Sympson and Hetter's (1985) method, a probabilistic method for exposure control was used. Their approach tries to make sure that the probability that an item $i$ is administered is less than some given value $r_i$; that is, that $P(A_i) \leq r_i$. If the probability of an item being selected is denoted as $P(S_i)$, then we know that the $P(A_i) = P(A_i|S_i)P(S_i) \leq r_i$. Since the values of $P(S_i)$ depend on the item pool and the particular CAT algorithm used, they are fixed. Since $P(S_i)$ and $r_i$ are both fixed, then after determining $P(S_i)$ from simulation, $P(A_i|S_i)$ can be found so that the inequality is satisfied; that is, by setting $P(A_i|S_i) \leq r_i/P(S_i)$. This method has been found to acceptably handle the problem of over-exposure of items, but not the problem of under-exposure of items. Modifications of the Sympson-Hetter method have been proposed by Stocking and Lewis (1998, 2000).

The method of $a$-stratification (Chang & Ying, 1999) was developed as a response to the argument that it may be useful to consider selecting "less optimal" choices of items in the early stages

of a CAT (Chang & Ying, 1996), and also as a way to remedy high exposure rates. When using maximum information item selection in CAT, which is the standard, items are chosen so that Fisher information is as large as possible. For the 2PL model (and 3PL model), this is the case when true ability $\theta_0$ is as close as possible to $b_i$ and $a_i$ is maximized. However, since $\theta_0$ is not known, it must be estimated by $\hat{\theta}$. When there is not enough data for an accurate estimate, such as in earlier stages of a test, the estimate $\hat{\theta}$ and the true value $\theta_0$ of ability may be drastically different. For the 2PL model, the true item information is, then

$$I_i(\theta_0|a_i,\hat{\theta}) = a_i^2 \frac{\exp\Big[a_i\Big(\theta_0-\hat{\theta}\Big)\Big]}{\Big\{1+\exp\Big[a_i\Big(\theta_0-\hat{\theta}\Big)\Big]\Big\}^2}.$$

From this, we find that true information can be drastically less than expected information when $\hat{\theta} \neq \theta_0$ (as in the earlier stages of a test), and, in fact, approaches 0 as $a_i$ increases (Chang & Ying, 2009). Thus, highly discriminating items should be avoided in the early parts of a test. The method of $a$-stratification remedies this situation, which is described in the following. The items in the item bank are ordered in terms of their $a$-parameters and split into $K$ levels, and the test is split into $K$ corresponding stages. In the $k$th test stage, $n_k$ items are sequentially selected from the $k$th level of the item bank; they are selected based simply on the closeness of $b$ to $\hat{\theta}$. This is repeated for each of the $K$ stages. This method was found to have decreased test overlap rate and more even exposure rates, as well as comparable ability estimation accuracy as the Sympson-Hetter method.

A related method is the method of $a$-stratification with $b$ blocking (Chang et al., 2001). A very important assumption when using the $a$-stratification method is that $b$ parameters are evenly distributed across all levels of $a$; that is, that $\text{Corr}(a,b) = 0$. However, this is rare in practice, with $a$- and $b$-parameters often positively correlated (Lord, 1984). Thus, the $b$-parameters also need to be partitioned accordingly. First, the items in the item bank are ordered in order of increasing $b$-parameters and split into $M$ blocks. Importantly, all blocks should have roughly the same number of items, differing by at most one. Within each block, the items are ordered in order of increasing $a$-parameters. The items are then partitioned within block into $K$ roughly equal sized strata. For the $k$th stratum across each of the blocks are combined together to create $K$ levels in the item

pool. The test is also split into $K$ corresponding stages. For the $k$th stage of the test, $n_k$ items are sequentially selected from the $k$th level of the item bank according to closeness of $b$ to $\hat{\theta}$. This is repeated for all $K$ stages of the test. For a simulated test administration of a retired item pool from the GRE-Quant—which has a correlation between $a$ and $b$ within its item bank of 0.44—it was found that the $a$-stratification with $b$ blocking method outperformed the $a$-stratification method in overlap rate, exposure rate, and ability estimation accuracy.

One final method to be discussed is the use of the MPI to include exposure control as a special constraint (Cheng & Chang, 2009). As previously discussed, the MPI method seeks to choose items such that the priority index in (1.14) is maximized. Here, $w_k$ is the weight for the $k$th constraint and $f_{ik}$ is an appropriate multiplier quantifying the "quota left" on a particular constraint for item $i$ on the $k$th constraint. If we suppose that constraint $k'$ requires that the exposure rate for item $i$ be less than or equal to $r_i$, then a special constraint for item exposure can simply be

$$f_{ik'} = \frac{r_i - \frac{n_i}{N}}{r_i},$$

where $n_i$ is the number of examinees that have seen item $i$ out of $N$ total examinees. The flexibility to handle content constraints, as well as exposure rate controls as constraints, is a highly important property of the MPI.

### 1.3.3 Extensions of CAT

MIRT models and unfolding IRT models are very useful for the measurement of a multitude of constructs, most notably personality and attitudinal measures. As such, combining these models with CAT is a very natural extension of CAT. While both of these topics are still fairly recent, more has been done in the area of multidimensional-CAT (MCAT) than in unfolding-CAT. Indeed, as much as unfolding IRT models are in their infancy, unfolding-CAT is in its infancy. Notably, the TAPAS (Tailored Adaptive Personality Assessment System) uses an unfolding model along with a preference switching model, the MDPP (Stark, 2002), to assess personality for use in Army selection (Drasgow et al., 2012). The assessment gives an examinee two statements, and the examinee must choose one of them as the statement that best describes himself. While in research and diagnostic settings, examinees will generally provide truthful answers, in a high-stakes test, such

as job selection and placement, it is more likely that they will not be truthful; the forced choice mechanism reduces the likelihood of an examinee simply providing answers seen as socially desirable. The unfolding model is used, because the assessment is composed are attitudinal questions. CAT elements are incorporated to maximize efficiency. For more information on current attempts at unfolding-CAT, see Roberts, Lin, and Laughlin (2001) and W.-C. Wang, Liu, and Wu (2013).

The most well-known operational MCAT is the Armed Services Vocational Aptitude Test Battery (ASVAB), which is also one of the oldest operational CATs in general. It is used to measure multiple facets of vocational aptitude simultaneously for the purpose of selection and placement in the armed services, much as TAPAS is used. The difference is that the TAPAS measures personality, while the ASVAB is a cognitive assessment.

In MCAT, much of the usual CAT mechanism is the same, with the largest change in implementation coming from the item selection algorithm. In a standard CAT with unidimensional IRT, items are selected that give the most information about the latent trait. While this goal is still desirable in MIRT, there is more than one latent trait, so a delicate balance between measuring all traits simultaneously arises. This is done by use of the Fisher information matrix, which is just a generalization of (refeq:iteminfo),

$$[\mathbf{I}_i(\boldsymbol{\theta})]_{jk} \equiv -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ln P_i(\boldsymbol{\theta})\middle|\boldsymbol{\theta}\right], \tag{1.16}$$

and the multidimensional K-L information index (Veldkamp & van der Linden, 2002; C. Wang, Chang, & Boughton, 2011), a generalization of (1.10)

$$\mathrm{KI}_i(\boldsymbol{\theta}_0) = \int_{\theta_{10}-\delta}^{\theta_{10}+\delta}\cdots\int_{\theta_{p0}-\delta}^{\theta_{p0}+\delta}\mathrm{KL}_i(\boldsymbol{\theta}_0\|\boldsymbol{\theta})\;\partial\boldsymbol{\theta}. \tag{1.17}$$

To use the Fisher information matrix, the information must be aggregated to create a criterion. Two possibilities are D-optimality (Mulder & van der Linden, 2009; Segall, 1996) and A-optimality (Mulder & van der Linden, 2009; van der Linden, 1999). D-optimality serves to minimize the generalized variance of the ability estimate $(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$, where $\boldsymbol{\Sigma}$ is a covariance matrix. A-optimality, on the other hand, serves to minimize the sum of the variances of the ability dimensions, which is simply the trace of the inverse of the Fisher information matrix. Interestingly, while KI is

clearly superior to Fisher information in unidimensional-CAT, this is not the case in MCAT when the MIRT model is compensatory in nature. Rather, a different use of K-L distance, the mutual information method (Mulder & van der Linden, 2010; Weissman, 2007) is best, with D-optimality providing similar results.

## 1.4 Response Times

CAT is an area with great promise and potential, but much work is left to do. Some current trends in the testing industry have spurred research to move in new directions. One trend that will be discussed here is using response time information. First, I will discuss models of response times in testing. I will conclude by discussing the literature of how response times can be used in testing.

### 1.4.1 Response Time Models

Response time information has become en vogue with the rise of computer-based tests, both linear and adaptive, because response times are easily-collected, essentially-free data. However, interest in the analysis of response times is not new. That interest is borne from the hypothesis—which dates back to at least the dawn of psychology as a science in the 1860s—that different ways of processing information take differing amounts of time, and so response times can, in essence, allow a researcher a way to infer the structure of the mind through a series of experiments (Luce, 1986).

One of the most well-known relationships in psychology is that of speed and accuracy. In general, it has been found that with increasing speed of response, accuracy decreases. This phenomenon, known as the speed-accuracy trade-off, has been observed for a long time (e.g., Garrett, 1922; Wickelgren, 1977; Woodworth, 1899). In testing, a distinction was made between a power test (i.e., a test with a number of items of differing difficulties with no time constraint) and a speed test (i.e., a time-limited test with a large number of easy items) by Gulliksen (1950). In a power test, the focus is entirely on response accuracy, whereas speed tests try to measure how quickly examinees answer items. Additionally, Gulliksen created a version of CTT for speed tests that mirrors the standard CTT (which is applicable to power tests). In his view, a random variable

$E = N - X$—where $N$ is the total number of items, $X$ represents the total correct score, and $E$ represents the total error score—can be decomposed into $W$, the number of items which the examinee gives an incorrect answer, and $U$, the number of items that the examinee does not reach; that is, $E = W + U$. A pure speed test can be defined as a test where $P(W = 0) = 1$, and, conversely, a pure power test can be defined as a test where $P(U = 0) = 1$. In this view, it is easy to see that in reality, no practical test is a pure power test or a power speed test; most tests are constructed as power tests, but are time-limited. Indeed, for complex tasks, it is known that speed and accuracy are distinct factors, so a single dimension is not sufficient to explain this relationship (Schnipke & Scrams, 2002). Thus, it is important to be able to model both the response accuracy and response time aspects of testing, though this has not always been the method of choice. As Spearman (1927) wrote: "Now, if we desire any genuine measurement of cognitive ability, it is to these universal quantitative properties of clearness [accuracy] and speed that we are obliged to turn" (p. 245).

Models for response times can be described as those that include only response times and those that include response times and response accuracy. Furthermore, van der Linden (2009b) identified two general approaches to modeling response accuracy with response times: (1) using separate models for response times and response accuracy, and (2) incorporating response times in models for response accuracy (or vice versa).

The first set of response time models—those only including response times—would be most applicable to tasks that are primarily governed by speed, such as speed tests. These models primarily differ in the assumption for the distribution of response times. Scheiblechner (1979) proposed a model for response times in the vein of IRT models that decomposes the mean response time linearly into effects due to persons and due to items, assuming an exponential distribution for response times, called the psychometric latency model. It takes the form

$$f(t_{ij}) = (\theta_j + \varepsilon_i) \exp\left[(\theta_j + \varepsilon_i) t_{ij}\right], \tag{1.18}$$

where $\theta_j > 0$ is the person-effect on response time for person $j$, $\varepsilon_i \geq 0$ is the item-effect on response time for item $i$, and $t_{ij}$ is the response time of person $j$ on item $i$. He also suggested a further linear decomposition of the item-effect $\varepsilon_i$, allowing for a better description of the item-effect. This

is given as

$$\varepsilon_i = \sum_{k=1}^{m} a_{ik}\eta_k + c_0, \tag{1.19}$$

where $a_{ik}$ is the known value of observed covariate $k$ on item $i$, $\eta_k$ is the effect of covariate $k$ on the items, and $c_0$ is a normalizing constant. This idea is similar to the linear logistic latent trait model (LLTM; Fischer, 1973), where the item difficulty parameter in a logistic response model is decomposed by known item attributes to elucidate the structure of item difficulty in a test. Another model that uses this linear decomposition of parameters to create an additive model is the generalized gamma distribution model by Maris (1993). He proposed to use a gamma distribution—which can be seen as a two-parameter generalization of the exponential distribution—to model response times, allowing a linear decomposition of the parameters for the gamma distribution. An advantage of using this model is that a multiplicative decomposition of the parameters is also easily handled and fit. A more recent model was given by (Rouder, Sun, Speckman, Lu, & Zhou, 2003). In their model, a three-parameter Weibull distribution is specified for response times, taking the form

$$f(t_{ij}) = \frac{\beta_j(t_{ij}-\psi_j)^{\beta_j-1}}{\sigma_j^{\beta_j}} \exp\left[-\frac{(t_{ij}-\psi_j)^{\beta_j}}{\sigma_j^{\beta_j}}\right],\ t_{ij} > \psi_j, \tag{1.20}$$

where $\beta_j$, $\sigma_j$, and $\psi_j$ are shape, scale, and shift parameters for person $j$, respectively. Importantly, this model assumes that response times $t_{ij}$ within person $j$ are identically distributed across items, and, thus, that the cognitive process for that individual is captured by the shape, scale, and shift parameters. Tatsuoka and Tatsuoka (1980) also proposed using a Weibull model for modeling response times in a testing situation, but unlike Rouder et al. (2003), they assumed the shape and scale were item parameters rather than person parameters, and that the shift parameter $\psi$ represents "minus" the person speededness. Other response time only models were given by Pieters and van der Ven (Poisson-Erlang model; 1982) and van der Linden (log-normal model; 2006).

The second modeling approach uses separate models for response times and response accuracy. A well-known example of this was given by Rasch (1960). In this volume, he gives two models for reading: a model for misreadings (measuring accuracy), and a model for reading speed. In the misreadings model, he assumes that for a reading passage $i$ with $N_i$, that the number of misreadings

$a_{ij}$ for examinee $j$ follows a Poisson distribution:

$$p(a_{ij}|N_i) = \frac{\lambda_{ij}^{a_{ij}}}{a_{ij}!}\exp(-\lambda_{ij}), \tag{1.21}$$

where $\lambda_{ij} = N_i\theta_{ij}$ and $\theta_{ij} = \frac{\delta_i}{\xi_j}$. Rasch interpreted $\delta_i$ as the difficulty of a text passage and $\xi_j$ as the ability of a reader. Thus, for more difficult texts or less able readers, the probability of a misreading goes up. It is a standard result in statistics that if the number of events given a fixed time is Poisson distributed, as above, then the amount of time taken for a fixed number of events to occur is gamma distributed. Thus, Rasch (1960) also proposed a model for reading speed where the time $t_ij$ taken to read $N_i$ words in a text passage is assumed to be gamma distributed. This is given as

$$p(t_{ij}|N_i) = \nu_{ij}\exp(-\nu_{ij}t_{ij})\frac{(\nu_{ij}t_{ij})^{N_i-1}}{(N_i-1)!}, \tag{1.22}$$

where $\nu_{ij} = \frac{\tau_j}{\omega_i}$, and $\omega_i$ is the difficulty of a text $i$ and $\tau_j$ is the speededness of an examinee $j$. Thus, for a faster examinee and a less difficult text, the rate at which an examinee reads $N_i$ words goes up. Taken together, Rasch was able to simultaneously model ability and speededness of examinees. Extensions of Rasch's approach were given by Jansen (1986, 1997a, 1997b) and Jansen and van Duijn (1992). Another approach used is that of Gorin (2005), where used the LLTM to decompose item difficulty into item attributes (experimental predictors) and attribute weights, as well as regressing the log-response times onto item difficulties and those same experimental predictors. This was done to find how much experimental predictors were related to correctly solving a reading comprehension problem and how quickly it was solved. Similar procedures were done by Embretson (1998) for abstract reasoning problems. One further response time model was proposed by van der Linden, Scrams, and Schnipke (1999). They propose a linear model of the log-response times

$$\ln t_{ij} = \mu + \delta_i + \tau_j + \epsilon_{ij}, \tag{1.23}$$

where $\mu$ is a grand-mean response-time level, $\delta_i$ is a time-intensity parameter for item $i$, $\tau_j$ is a slowness parameter for examinee $j$, and $\epsilon_{ij}$ is a normally distributed residual term. In their paper, they propose to use this model in a shadow-test approach to item selection (van der Linden & Glas, 2010; van der Linden & Reese, 1998), by incorporating model-estimated response times into the

constraints to control for differential speededness of examinees. This model was also used to detect aberrant behaviors (van der Linden & van Krimpen-Stoop, 2003). Importantly, modeling response times and response accuracy separately assumes that they are fully independent, which may not be the case.

The final modeling approach uses incorporates response times into models for response accuracy (and vice versa). Examples of these models include those given by Thissen (1983), Roskam (1987), Verhelst, Verstralen, and Jansen (1997), and Rouder et al. (2003). In Thissen (1983), for instance, proposed a model, modified from Furneaux (1961) as

$$\ln t_{ij} = \mu + \delta_i + \tau_j - \rho z_{ij} + \epsilon_{ij}, \tag{1.24}$$

where $\mu$, $\delta_i$, and $\tau_j$ are defined similarly to the model in (1.23); $z_{ij} = a_i(\theta_j - b_i)$, as defined by the 2PL model (used, here, as the response accuracy model); $\rho$ is a regression parameter describing how examinee ability and item difficulty relate to log-response time; and $\epsilon_{ij}$ is a normally distributed residual term. In this model, it is assumed that as ability increases (or difficulty decreases), response time decreases; it does not assume that response time directly affects response accuracy. This is an example of a model incorporating ability into a response time model. Other models include response times into a model of response accuracy. Roskam (1987, 1997) proposed one such model. His model is a Rasch-type response accuracy model with response time as a predictor:

$$P_i(\theta_j) = [1 + \exp(\theta_j + \ln t_{ij} - b_i)]^{-1}. \tag{1.25}$$

An interesting feature of this model is that the speed-accuracy trade-off is directly incorporated; for a faster examinee response (i.e., $\ln t_{ij}$ decreases), the probability of a correct response decreases. One other feature of this model is that it assumes that given enough time, any examinee can increase the probability of a correct response to 1. A related model given by Verhelst et al. (1997) uses a latent variable of person-speededness $\tau_j$ instead of a directly using response time in the model

$$P_i(\theta_j, \tau_j) = [1 + \exp(\theta_j - \ln \tau_j - b_i)]^{-\nu_i}, \tag{1.26}$$

where $\nu_i$ is an item-dependent shape parameter; with $\nu_i = 1$, this is just a Rasch-type logistic

model. As in the Roskam (1997) model, the speed-accuracy trade-off is directly incorporated; however, speededness is assumed to be a person-parameter invariant to the item which may not be reasonable in many cases. T. Wang and Hanson (2005) give a third model in this mold where the speed-accuracy trade-off is incorporated into a 3PL model rather than a Rasch-type model. It is given as follows:

$$P_i(\theta_j) = c_i + (1 - c_i)\left\{1 + \exp\left[-a_i\left(\theta_j - \frac{\rho_j d_i}{t_{ij}} - b_i\right)\right]\right\}^{-1}, \tag{1.27}$$

where $\rho_j$ and $d_i$ are person- and item-parameters for speededness, respectively. An interesting property of this model is that as response time increases, this model approaches a standard 3PL model, implying that unlike the Roskam (1997) model, a given examinee can increase the probability of a correct response, but is still limited by his or her ability.

Other joint models for response time and accuracy follow a multilevel approach. These models have two levels; at the first level, a response accuracy and a response time model are given separately, and at the second level, the person parameters for ability and speededness across the population are modeled jointly with a multivariate distribution, such as the multivariate normal distribution. Within this set of models, two approaches have been taken for modeling the response times in the first level: using a parametric model (Fox, Klein Entink, & van der Linden, 2007; Klein Entink, Fox, & van der Linden, 2009; Klein Entink, Kuhn, Hornke, & Fox, 2009; Molenaar, Tuerlinckx, & van der Maas, 2015; van der Linden, 2007); or using a semiparametric model (C. Wang, Chang, & Douglas, 2013; C. Wang, Fan, Chang, & Douglas, 2013). Justification for these types of models is given by the recognition of two things: 1) ability and speededness are separate factors (Kennedy, 1930); and 2) for a given examinee and a fixed level of speededness, accuracy remains constant (Tate, 1948). These models reject the idea of incorporating the speed-accuracy trade-off directly in some tests, because, as van der Linden (2007) notes, if the test has a reasonable time limit, then an examinee should have a fixed level of speededness throughout the test. Thus, the speed-accuracy trade-off is a within-person constraint, not allowing the prediction of accuracy of one person from another (Klein Entink, Fox, & van der Linden, 2009).

Of the multilevel approach models, the most popular is van der Linden's (2007) hierarchical model. The first level models include the 3PL model, given in (1.1), as a response accuracy model,

and a lognormal model for the response time model. The intuition behind the lognormal model can be traced the concept of speed in physics (van der Linden, 2006).

The average speed of an object is simply the distance $D$ from the object's location at the first time point to the object's location at the second time point divided by the time $T$ it took to go from the first location to the second location. In other words,

$$\textit{average distance} = \frac{D}{T}.$$

Similarly, the speed of an examinee $\tau_j^*$ on a test item should be measured as the amount of labor $\beta_i^*$ needed to finish the item divided by the amount of time $t_{ij}$ needed to complete the item, which can be expressed as follows:

$$\tau_j^* = \frac{\beta_i^*}{t_{ij}}.$$

Clearly, then, by algebraic manipulation $t_{ij} = \beta_i^*/\tau_j^* \Rightarrow \ln t_{ij} = \beta_i - \tau_j$, where $\beta_i = \ln \beta_i^*$ and $\tau_j = \ln \tau_j^*$. If we allow for an item-specific error term, then the lognormal model is obtained. It is as follows:

$$\ln T_{ij} = \beta_i - \tau_j + \epsilon_i, \tag{1.28}$$

where $\beta_i$ is an time-intensity effect for item $i$, $\tau_j$ is speededness for person $j$, and $\epsilon_i$ is a normally-distributed error term with mean 0 and variance $\alpha_i^{-2}$. The $\alpha_i$ parameters can be interpreted as item discrimination parameters. Another way of stating this is that

$$T_{ij}^* \equiv \ln T_{ij} \sim N(\beta_i - \tau_j, \alpha_i^{-2}). \tag{1.29}$$

At the second level of the hierarchical model, the person parameters $\boldsymbol{\xi}_j = (\theta_j, \tau_j)$ take on a joint multivariate normal distribution with mean vector $\boldsymbol{\mu}_P$ and covariance matrix $\boldsymbol{\Sigma}_P$.

### 1.4.2 Using Response Times in Testing

With this wealth of information, it has fallen upon psychometricians to find ways in which response times and response time models can be used effectively for testing purposes. Currently, the most popular model in the literature is van der Linden's (2007) hierarchical model. Some applica-

tions of the use of response times that make use the hierarchical model include cheating detection (van der Linden, 2009a; van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003), item parameter estimation (van der Linden, Klein Entink, & Fox, 2010), and item selection (Choe & Kern, 2014; Fan et al., 2012; van der Linden, 2008, 2009c). Much of the future work in response time research revolves around parameter estimation—the current methods rely heavily on computationally intensive Markov chain Monte Carlo techniques—and newer applications.

One hot topic in measurement right now is detecting cheating and its cousin, aberrant responses. Aberrant responses can be any response pattern which is out of the ordinary, which can occur due to bad test items, ambiguous instructions, copying answers (cheating), students that need test accomodations, and so on (van der Linden & Guo, 2008). In particular, it seems especially fruitful to consider response time information for detecting cheating and aberrant responses, and several methods have been created for this task.

In van der Linden (2009a), the author focuses on detecting a specific kind of cheating: the collusion between two test-takers. To this end, the lognormal model for response times is generalized to a bivariate lognormal model to analyze the RTs between pairs of test takers. This model is as follows:

$$f(\ln t_{ij}, \ln t_{ik}|\tau_j, \tau_k) = \frac{\alpha_i^2}{2\pi\sqrt{1-\rho_{jk}^2}} \exp\left\{-\frac{1}{2(1-\rho_{jk}^2)}(\psi_{jk}^2 - 2\rho_{jk}\psi_{ij}\psi_{ik} + \psi_{ik}^2)\right\}, \tag{1.30}$$

where $\psi_{ij} = \alpha_i[\ln t_{ij} - (\beta_i - \tau_j)]$, $\rho_{jk}$ is the degree to which the response times for two fixed test takers agree, and $\alpha_i$, $\beta_i$, and $\tau_j$ are defined as in the univariate lognormal model. It was proposed to use a test of the hypothesis

$$H_0 : \rho_{jk} = 0,$$

against

$$H_1 : \rho_{jk} > 0,$$

as a way to detect whether or not two test-takers were colluding together. In van der Linden and Guo (2008), the goal is to detect, more generally, aberrant responses using the posterior predictive distribution of the log-response time on item $i$ (given the response times and responses on all

remaining items). It is shown that using the hierarchical model that this is given as

$$\begin{aligned} f(t^*_{ij}|\mathbf{t}^*_{j(-i)}, \mathbf{u}_{j(-i)}) &= \int f(t^*_{ij}|\tau_j) f(\tau_j|\mathbf{t}^*_{j(-i)}, \mathbf{u}_{j(-i)})\ d\tau_j \\ &= \iint \left[\prod_{k=1}^{n} f(t^*_{kj}|\tau_j)\right] f(\tau_j|\theta_j) f(\theta_j|\mathbf{u}_{j(-i)})\ d\tau_j\ d\theta_j \end{aligned} \tag{1.31}$$

where $t^* = \ln t$, $\mathbf{t}^*_{j(-i)} = (t^*_{1j}, \ldots, t^*_{(i-1),j}, t^*_{(i+1),j}, \ldots, t^*_{nj})$, and $\mathbf{u}_{j(-i)}$ is defined similarly to $\mathbf{t}^*_{j(-i)}$. Two types of aberrances for an item can occur: low response times and high response times. If interest is on low response times, then define $\pi_{ij} = P(T^*_{ij} \leq t^*_{ij}|\mathbf{t}^*_{j(-i)}, \mathbf{u}_{j(-i)})$, and if interest is on high response times, then define $\pi_{ij} = P(T^*_{ij} \geq t^*_{ij}|\mathbf{t}^*_{j(-i)}, \mathbf{u}_{j(-i)})$. Then, the probability of observing a response time pattern on $k$ items is

$$\pi_j^{pattern} = 1 - \prod_{i=1}^{k}(1 - \pi_{ij}).$$

An exceptionally small value of $\pi_j^{pattern}$ means that this response time pattern is aberrant, and might warrant further investigation.

In testing, one common application is to use collateral information—that is, information that is not of direct interest to the parameters of concern that is collected simultaneously with the direct information—to bolster the estimation accuracy of item parameters or the ability parameter. It should be clear that response times can be considered to be a form of collateral information. In van der Linden et al. (2010), two sources of collateral information are identified and exploited: the joint information in the responses and the response times, as summarized in second-level parameters; and the information in the posterior distribution of the response parameters given the response times. For this application, an additional assumption of the joint second-level model of the item parameters $a_i$, $b_i$, $c_i$, $\alpha_i$, and $\beta_i$ is needed; in this case, it is simply assumed that $(a, b, c, \alpha, \beta) \sim MVN(\boldsymbol{\mu}_I, \Sigma_I)$. If we take $\boldsymbol{\xi}_i = (a_i, b_i, c_i)$, then the authors show that

$$f(\boldsymbol{\xi}_i|\mathbf{x}_i, \mathbf{t}_i, \boldsymbol{\mu}_I, \Sigma_I) \propto f(\mathbf{x}_i|\boldsymbol{\xi}_i) f(\boldsymbol{\xi}_i|\mathbf{t}_i, \boldsymbol{\mu}_I, \Sigma_I). \tag{1.32}$$

Furthermore, it can be shown that

$$f(\theta_j|\mathbf{x}_j, \mathbf{t}_j, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) \propto f(\mathbf{x}_j|\theta_j) f(\theta_j|\mathbf{t}_j, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P). \tag{1.33}$$

An MCMC algorithm, such as the Gibbs sampler, can be used to estimate $\boldsymbol{\xi}_i$ and $\theta_j$, jointly. If only $\theta_j$ is of interest and the item parameters are assumed to be known, such as in a CAT, then $\theta_j$ can be estimated in this framework by either taking the mean of the posterior (i.e., the EAP estimator) or taking the maximum of the posterior (i.e., the MAP estimator). In van der Linden (2008), it is shown that the empirical prior for $\theta_j$ in (1.33), $f(\theta_j|\mathbf{t}_j)$, can be given as $\theta_j|\mathbf{t}_j \sim N(\mu_{\theta|\mathbf{t}_{k-1}}, \sigma^2_{\theta|\mathbf{t}_{k-1}})$, where

$$\mu_{\theta_j|\mathbf{t}_{k-1}} = \frac{\sigma_{\theta\tau} \sum_{i=1}^{k-1} \alpha_i^2 (\beta_i - t_{ij}^*)}{1 + \sigma_\tau^2 \sum_{i=1}^{k-1} \alpha_i^2}$$

and

$$\sigma^2_{\theta_j|\mathbf{t}_{k-1}} = 1 - \frac{\sigma^2_{\theta\tau}}{\sigma^2_\tau} + \frac{\sigma^2_{\theta\tau}}{1 + \sigma_\tau^2 \sum_{i=1}^{k-1} \alpha_i^2}.$$

One very natural application in CAT is to use response times for item selection. In thinking about item selection, several criteria can be formulated corresponding to different goals. One possibility is to control for differential speededness. Differential speededness refers to the differences in time intensity for an overall test between examinees that is due to a combination of the tailored nature of a CAT and the differences in speededness parameters $\tau$ between examinees. This can become a problem for time-limited tests. For instance, some examinees may have to work under high pressure, because they were administered many highly time-intensive items, whereas other examinees have plenty of time to finish, because they were given many low time-intensity items. This can give the impression of unfairness between the test-takers, and can potentially impact their scores. Controlling for differential speededness can mitigate this issue. As previously discussed, one way of controlling for differential speededness is to use the log-response time model in (1.23) as a constraint within a shadow-test approach (van der Linden et al., 1999). In van der Linden (2009c), it was proposed to instead use the one of two posterior predictive densities of log-time derived from the hierarchical model within a shadow-testing approach, instead.

The main goal of a CAT is to minimize the length of a test while maintaining the accuracy of ability estimation. When response times are involved, another possible goal of item selection

could be to minimize the total time taken to complete a test; that is, to minimize length in terms of real time. Of course, just blindly minimizing test time by itself would undoubtedly mean that most items would be too easy for the majority of examinees, which, in turn, means that the tests are inappropriate for finding the locations of most examinees on the ability scale. Thus, when minimizing test time, it should be done in conjunction with maximizing ability estimation accuracy. One method that does this in a simple way is the maximum information per time unit (MICT) method of Fan et al. (2012). In this method, the next item is chosen to maximize the function

$$\text{MICT}_i = \frac{I_i(\hat{\theta})}{\text{E}(T_i|\hat{\tau})}, \tag{1.34}$$

where $\hat{\theta}$ and $\hat{\tau}$ are the current estimates of ability and speededness. Since the response time is lognormally distributed, it is easy to show that

$$\text{E}(T_i|\hat{\tau}) = \exp\left(\beta_i - \hat{\tau} + \frac{1}{2\alpha_i^2}\right).$$

This method works great for minimizing total test time; it has been shown to work better than other competing item selection methods for minimizing total test time (Veldkamp, 2016). Unfortunately, this comes with the huge drawback of greatly skewing the item exposure. The method that Fan et al. (2012) take is to use $a$-stratification with $b$ blocking in conjunction with the information per time unit concept. This method adequately handles the item exposure issue, while still lessening total test time, though not as much as with MICT.

A final method for item selection using response times is a generalization based on the MICT, called the generalized time-weighted maximum information criterion (GMICT; Choe & Kern, 2014). This is as follows:

$$\text{IT}_i^G = \frac{I_i(\hat{\theta})}{|\text{E}(T_i|\hat{\tau}) - v|^w}. \tag{1.35}$$

This method generalizes MICT in two ways. First, it includes a centering value $v$ that allows for more control over the test time; the procedure will tend to choose items where the expected response time is close to $v$. This can help with the problem of perceived fairness and differential speededness. Second, it adds an exponent weighting parameter $w$ for control over how the deviance of the expected response time from $v$ is weighted. Notably, the standard maximum information

criterion ($w = 0$) and MICT ($w = 1$, $v = 0$) are special cases of GMICT. To assess the performance of the item selection algorithm, MSE of $\theta$, MSE of $\tau$, mean and (MTT) standard deviation (STT) of test times across persons, and the item exposure rate skewness $\chi^2$ given in (1.15). It is argued that standard deviation of test time is an important index of differential speededness; a lower STT means that there are fewer differences in test times between persons, implying fewer effects of differential speededness. When compared on these measures, with the appropriate $v$ parameter and $w = 1$, item skewness, MTT, and STT for GMICT are all smaller than for the maximum information criterion, while the MSE of $\theta$ is increased. Compared to the MICT, item skewness and STT are greatly decreased, while MTT and MSE of $\theta$ are increased slightly. In all cases, MSE of $\tau$ is fairly constant. Thus, we find a complex relationship between $\theta$ estimation accuracy, test security control, differential speededness control, and overall test time minimization.

# Chapter 2

# A Method for Estimating Ability and Speededness Jointly

## 2.1 Introduction

Computerized adaptive testing (CAT) systems are particularly suited for situations where an accurate estimation of an examinee's true score is of utmost importance, such as large-scale, high-stakes admissions exams. Additionally, computer-based tests, such as CATs, allow for easy collection of response times. With this abundance of essentially-free data, methods and applications for using response time data have become en vogue. Some of the applications include cheating detection (van der Linden, 2009a), shortening the time needed to take a test (Fan et al., 2012), and item selection (van der Linden, 2008). It only seems natural that CATs be modified to take advantage of response time information, especially since it is well-known that response accuracy and response time are related.

To use response times to their greatest potential, it is necessary to create a model to make a connection between observed response times and a latent trait describing the speededness of the test-takers. Much research has been done on this topic; for a comprehensive review of response time modeling in testing, see either Schnipke and Scrams (2002) or Lee and Chen (2011). Some examples of response time models, as discussed in Section 1.4.1, were developed by Scheiblechner (1979), Tatsuoka and Tatsuoka (1980), Thissen (1983), Maris (1993), Verhelst et al. (1997), Douglas, Kosorok, and Chewning (1999), T. Wang and Hanson (2005), Meyer (2008), and Klein Entink, Fox, and van der Linden (2009). Several of these models try to model the time-accuracy trade-off directly. However, to model the usual time-accuracy trade-off directly would seem to run counter to the ideal of an exam score—particularly on a high-stakes exam—being based on only the examinee's problem-solving ability, and not the speed at which he or she finishes a problem.

To handle this modeling situation, van der Linden (2007) developed a hierarchical framework for

response time, in which he introduced a second latent variable for the speededness of the test-taker along with the usual latent variable for the test-taker's ability. Later, van der Linden (2008) used his model in a CAT. In his method, he estimated ability using a response time–based expected a posteriori (EAP)-type procedure, but did not directly estimate speededness as well. He argued that by having more accurate estimation of ability, even though items were selected using maximum information of ability without taking response times into account, item selection was improved. By similar reasoning, it can be argued that if one utilizes a method for selecting items that uses estimates for both ability and speededness, then similar gains can be gleaned if better estimates for both ability and speededness can be obtained by using response times. Here, a method for incorporating response times in a maximum a posteriori (MAP) type estimator will be pursued.

The gold standard in test scoring has tended to be using the maximum likelihood estimate (MLE). This can be attributed to two properties of the MLE. First, MLE is an unbiased estimator of ability for long tests (Lord, 1983). In testing, this is generally seen as a desirable property, because we would like the estimated scores to be as indicative of the true scores as possible, especially in cases when the scores are used to determine pass-fail, such as in licensure testing. Second, we know the asymptotic properties of the MLE; that is, it is distributed normally, with a mean $\theta$ (which is the true ability, so it is unbiased) and variance as the inverse of the Fisher information of $\theta$. Thus, with long tests, we know the behavior of the estimator quite well. However, there are some downsides to the MLE as well. The most well-known is that in some cases, the likelihood may only have a maximum at a boundary. This can happen, for instance, when an examinee answers all items correctly or all items incorrectly. Furthermore, for short tests, it can have large bias (Lord, 1983).

One answer for this is to use the MAP estimator. Instead of finding the maximum of the likelihood, the MAP is the maximum of the posterior distribution. This is accomplished by simply weighting the likelihood by an appropriate prior distribution. The MAP estimator has several properties that are important. First and foremost, the MLE can be considered a special case of the MAP when the prior distribution is uniform. As such, with a non-uniform prior (the standard is a normal prior), it can be considered to be a "regularized" form of MLE, and is thus able to avoid the likelihood issues present in MLE. In some sense, the prior acts as an extra observation

which will tend to make the likelihood behave. Secondly, the MAP is a convex combination of the MLE and the prior. Essentially, the MAP "shrinks" the estimate towards the prior; this can help the estimate to avoid overfitting to the data, which can be especially helpful for shorter tests. Furthermore, in comparison to the MLE, the MAP has a smaller variance, but tends to have a larger bias (C. Wang, 2015).

In the current study, van der Linden's (2007) hierarchical framework is used, and a MAP estimation method is developed for the joint estimation of the person parameters for ability and speededness. The reason for jointly estimating ability and speededness is that in a Bayesian framework, both latent variables should benefit from the borrowing of information from the other. Furthermore, it can be argued that for many applications, the values of both latent variables are of interest. Here, we assume that the items are already operational, so that their item parameters are known. We also assume that the covariance between the ability and speededness parameters is known. Furthermore, the joint posterior information matrix for the person parameters is derived. A simulation study comparing the performance of the new MAP estimation method using response times with the standard MLE method is done for a variety of relationships between ability and speededness to determine when the MAP and the MLE perform best. A second simulation study is also completed which compares the two estimation methods with a real item bank. This simulation study is further extended by comparing these methods with two different item selection methods: using the maximum information criterion (MIC) and using the maximum information per time unit criterion (MICT; Fan et al., 2012).

## 2.2 Model

A hierarchical modeling framework was proposed by van der Linden (2007) to model speed and accuracy on test items. One level consists of an IRT model and a RT model, while a second level accounts for interdependencies between the item and person parameters within the two models.

### 2.2.1 First-level models

Let the items be indexed by $i = 1, \ldots, I$ and the persons be indexed by $j = 1, \ldots, J$. For each examinee $j$ there is a vector of responses $\mathbf{X}_j = (X_{1j}, \ldots, X_{Ij})$ and a vector of response time

$\mathbf{T}_j = (T_{1j}, \ldots, T_{Ij})$. The responses are assumed to be modeled by a 3PL model. That is, given the $j$th person's ability parameter $\theta_j$, the probability of answering the $i$th item correctly is modeled as

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}}, \tag{2.1}$$

where $a_i$, $b_i$, and $c_i$ are the discrimination, difficulty, and guessing parameters of item $i$, respectively. Then, the likelihood of examinee $j$'s responses is given as

$$l_j(\mathbf{X}_j) = \prod_{i=1}^{I} P_i(\theta_j)^{x_{ij}}(1 - P_i(\theta_j))^{1-x_{ij}}. \tag{2.2}$$

The response times $T_{ij}$ for the $j$th person on the $i$th item are assumed to be modeled by the lognormal model proposed by van der Linden (2006), and so the log of the response times $\ln T_{ij}$ are normally distributed. This is expressed as follows:

$$\ln T_{ij} \sim f(\ln t_{ij}|\tau_j, \alpha_i, \beta_i) = N(\beta_i - \tau_j, \alpha_j^{-2}),$$

such that

$$f(\ln t_{ij}|\tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))\right]^2\right\}, \tag{2.3}$$

where $\tau_j$ is the speed parameter for examinee $j$, and $\alpha_i$ and $\beta_i$ are the time intensity and discriminating power parameters of item $i$, respectively.

Let $\boldsymbol{\xi_j} = (\theta_j, \tau_j)$ denote the vector of person parameters for person $j$, and let $\boldsymbol{\psi_i} = (a_i, b_i, c_i, \alpha_i, \beta_i)$ denote the vector of item parameters for item $i$. Here, we note that an important assumption, beyond the usual IRT assumptions, is that $X_{ij}$ is conditionally independent from $T_{ij}$ given $(\theta, \tau)$. Due to this conditional independence, the joint sampling distribution of $\mathbf{X}_j$ and $\mathbf{T}_j$, $j = 1, \ldots, J$, follows directly from Equations 2.2 and 2.3:

$$f(\mathbf{x}_j, \ln \mathbf{t}_j|\boldsymbol{\xi}_j, \boldsymbol{\psi}) = \prod_{i=1}^{I} l(x_{ij}|\theta_j, a_i, b_i, c_i) f(\ln t_{ij}|\tau_j, \alpha_i, \beta_i). \tag{2.4}$$

### 2.2.2 Second-level models

At the second level, we have one model that describes the joint distribution of the person parameters. We assume that the values of $\boldsymbol{\xi}_j$ come from a multivariate normal distribution

$$\boldsymbol{\xi}_j \sim N_2(\boldsymbol{\mu}_p, \Sigma_p) = f(\boldsymbol{\xi}_j|\boldsymbol{\mu}_p, \Sigma_p),$$

where the joint density is given as

$$f(\boldsymbol{\xi}_j|\boldsymbol{\mu}_p, \Sigma_p) = \frac{|\Sigma_p|^{-1/2}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_p)^T \Sigma_p^{-1}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_p)\right], \tag{2.5}$$

with a mean vector

$$\boldsymbol{\mu}_p = (\mu_\theta, \mu_\tau)$$

and a covariance matrix

$$\Sigma_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}$$

Finally, combining Equations 2.4 and 2.5, we find that the joint density for responses, response times, and person parameters for the $j$th examinee is given as

$$f\big(\mathbf{x}_j, \ln \mathbf{t}_j, \boldsymbol{\xi}_j|\boldsymbol{\psi}\big) = f\big(\mathbf{x}_j, \ln \mathbf{t}_j|\boldsymbol{\xi}_j, \boldsymbol{\psi}\big)\, f\big(\boldsymbol{\xi}_j|\boldsymbol{\mu}_p, \Sigma_p\big)\,. \tag{2.6}$$

For ease of notation, where the meaning is clear, the person index $j$ is dropped throughout the rest of the present chapter.

### 2.2.3 Maximum a posteriori (MAP) estimation of person parameters

To find the joint maximum a posteriori (MAP) estimate of $\theta$ and $\tau$, it is necessary to find the maximum of the posterior distribution of $\boldsymbol{\xi}$; that is, to find $\text{argmax}_{\theta,\tau}\, f(\boldsymbol{\xi}|\mathbf{x}, \ln \mathbf{t}, \Psi)$, where $\mathbf{x}$ and

$\ln \mathbf{t}$ are the given data. This problem can be simplified as follows:

$$\begin{aligned}\underset{\theta,\tau}{\operatorname{argmax}}\, f(\boldsymbol{\xi}|\mathbf{x}, \ln \mathbf{t}, \boldsymbol{\Psi}) &= \underset{\theta,\tau}{\operatorname{argmax}}\, \frac{f(\mathbf{x}, \ln \mathbf{t}, \boldsymbol{\xi}|\boldsymbol{\Psi})}{\int_{\boldsymbol{\xi}} f(\mathbf{x}, \ln \mathbf{t}, \boldsymbol{\xi}|\boldsymbol{\Psi})} = \underset{\theta,\tau}{\operatorname{argmax}}\, f(\mathbf{x}, \ln \mathbf{t}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \\ &= \underset{\theta,\tau}{\operatorname{argmax}}\ f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi})\, f\big(\boldsymbol{\xi}|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\big) \\ &= \underset{\theta,\tau}{\operatorname{argmax}}\ \ln f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi}) + \ln f\big(\boldsymbol{\xi}|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\big)\,.\end{aligned}$$

This is done by finding the zeros for the first derivatives of the log of the joint density in Equation 2.6 with respect to $\theta$ and $\tau$, $\frac{\partial \ln f}{\partial \tau}$ and $\frac{\partial \ln f}{\partial \theta}$, jointly. These are given as follows:

$$\begin{aligned}\frac{\partial \ln f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi})}{\partial \theta} &= \sum_{i=1}^{I} \left(\frac{a_i(x_i - P_i(\theta))}{(1 - c_i)}\right)\left(\frac{P_i(\theta) - c_i}{P_i(\theta)}\right) - \frac{\sigma_\tau^2(\theta - \mu_\theta) - \sigma_{\theta\tau}(\tau - \mu_\tau)}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\ \frac{\partial \ln f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi})}{\partial \tau} &= -\sum_{i=1}^{I} \alpha_i^2(\ln t_i - (\beta_i - \tau)) - \frac{\sigma_\theta^2(\tau - \mu_\tau) - \sigma_{\theta\tau}(\theta - \mu_\theta)}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}.\end{aligned} \tag{2.7}$$

A solution to this may be found by using a Newton-Raphson algorithm. For this, the second derivatives of the function to be maximized are needed. They are as follows:

$$\begin{aligned}\frac{\partial^2 \ln f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi})}{\partial \theta^2} &= \sum_{i=1}^{I} a_i^2 \left(\frac{P_i(\theta) - c_i}{(1 - c_i)^2}\right)\left(\frac{1 - P_i(\theta)}{P_i(\theta)}\right)\left(\frac{x_i c_i - P_i(\theta)^2}{P_i(\theta)}\right) - \frac{\sigma_\tau^2}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\ \frac{\partial^2 \ln f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi})}{\partial \tau^2} &= -\sum_{i=1}^{I} \alpha_i^2 - \frac{\sigma_\theta^2}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}, \quad \text{and} \\ \frac{\partial^2 \ln f(\mathbf{x}, \ln \mathbf{t}|\boldsymbol{\xi}, \boldsymbol{\psi})}{\partial \theta\ \partial \tau} &= \frac{\sigma_{\theta\tau}}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}.\end{aligned} \tag{2.8}$$

Finally, given close enough starting values for the person parameters $\theta_0$ and $\tau_0$, and using (2.7) and (2.8), the equation

$$\begin{pmatrix}\theta_{n+1} \\ \tau_{n+1}\end{pmatrix} = \begin{pmatrix}\theta_n \\ \tau_n\end{pmatrix} - \begin{pmatrix}\frac{\partial^2 \ln f}{\partial \theta^2} & \frac{\partial^2 \ln f}{\partial \theta \partial \tau} \\ \frac{\partial^2 \ln f}{\partial \theta \partial \tau} & \frac{\partial^2 \ln f}{\partial \tau^2}\end{pmatrix}^{-1} \begin{pmatrix}\frac{\partial \ln f}{\partial \theta} \\ \frac{\partial \ln f}{\partial \tau}\end{pmatrix} \tag{2.9}$$

can be solved iteratively until convergence. The resulting value is the joint MAP.

Note that for use in a CAT, reasonable estimates for $\sigma_\theta^2$, $\sigma_\tau^2$, and $\sigma_{\theta\tau}$ are needed. Also, $\mu_\theta$ and $\mu_\tau$

are needed. When estimating an item bank, these values are largely arbitrary, and need to be set for identification purposes. For instance, we could set $\mu_\theta = \mu_\tau = 0$, and $\sigma_\theta^2 = 1$. The other values, $\sigma_\tau^2$ and $\sigma_{\theta\tau}$, would be estimated. With a known, operational item bank, these values for a given population of test-takers are not arbitrary. They would need to be known (from past experience) ahead of time for use within a single examinee's test.

### 2.2.4 Information functions

Generally, the accuracy of the measurement in a Bayesian context is indexed by the posterior variance of the variables of interest, which is $\boldsymbol{\xi} = (\theta, \tau)'$ in this case. By Sorensen and Gianola (2002, pg. 331), we find that asymptotically this is the inverse of the information matrix of $\boldsymbol{\xi}$. This information matrix is calculated as the Fisher information matrix which, under certain regularity conditions, is

$$[I(\boldsymbol{\xi})]_{ij} = I(\theta,\tau)_{ij} = -\mathrm{E}\left[\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln f(\boldsymbol{\xi}|\mathbf{x},\ln\mathbf{t}) \;\middle|\; \boldsymbol{\xi}\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln f(\mathbf{x},\ln\mathbf{t},\boldsymbol{\xi}) \;\middle|\; \boldsymbol{\xi}\right]. \tag{2.10}$$

From (2.8), we find that

$$\begin{aligned}
I(\theta,\tau)_{11} &= \sum_{i=1}^{I} a_i^2 \left(\frac{P_i(\theta)-c_i}{1-c_i}\right)^2 \left(\frac{1-P_i(\theta)}{P_i(\theta)}\right) + \frac{\sigma_\tau^2}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\
I(\theta,\tau)_{12} &= -\frac{\sigma_{\theta\tau}}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2}, \text{ and} \\
I(\theta,\tau)_{22} &= \sum_{i=1}^{I} \alpha_i^2 + \frac{\sigma_\theta^2}{\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2},
\end{aligned} \tag{2.11}$$

so that

$$I(\theta,\tau) = \begin{pmatrix} I(\theta,\tau)_{11} & I(\theta,\tau)_{12} \\ I(\theta,\tau)_{12} & I(\theta,\tau)_{22} \end{pmatrix}.$$

Finally, we find that the measurement error variance matrix is given as

$$I(\theta,\tau)^{-1} = \frac{1}{I(\theta,\tau)_{11}I(\theta,\tau)_{22} - I(\theta,\tau)_{12}^2} \begin{pmatrix} I(\theta,\tau)_{22} & -I(\theta,\tau)_{12} \\ -I(\theta,\tau)_{12} & I(\theta,\tau)_{11} \end{pmatrix}. \tag{2.12}$$

Keeping in mind that the usual goal of CAT is to select items so that the number of items needed for an accurate measurement is minimized—that is, to select items that measure ability well for a given individual—a scant glance at Equation 2.12 suggests a new item selection technique: to select items that minimize

$$[I(\theta,\tau)^{-1}]_{11} = \frac{I(\theta,\tau)_{22}}{I(\theta,\tau)_{11}I(\theta,\tau)_{22} - I(\theta,\tau)_{12}^2}. \tag{2.13}$$

### 2.2.5 Maximum Information Per Time Unit Item Selection

The main goal of a CAT is to minimize the length of a test while maintaining the accuracy of ability estimation. Traditionally, this is done by simply choosing items at every step of the test so that the chosen item maximizes Fisher information given the current estimate of ability $\hat{\theta}$ (Lord, 1980). That is, the next item $i$ to be administered is the one that makes

$$I_i(\hat{\theta}) = \frac{\left[\frac{\partial}{\partial\hat{\theta}}P(\hat{\theta})\right]^2}{P_i(\hat{\theta})\left(1 - P_i(\hat{\theta})\right)} \tag{2.14}$$

as small as possible.

When response times are involved, another possible goal of item selection could be to minimize the total time taken to complete a test; that is, to minimize length in terms of real time. Of course, just blindly minimizing test time by itself would undoubtedly mean that most items would be too easy for the majority of examinees, which, in turn, means that the tests are inappropriate for finding the locations of most examinees on the ability scale. Thus, when minimizing test time, it should be done in conjunction with maximizing ability estimation accuracy. One method that does this in a simple way is the maximum information per time unit (MICT) method of Fan et al. (2012). In this method, the next item is chosen to maximize the function

$$\text{MICT}_i = \frac{I_i(\hat{\theta})}{\text{E}(T_i|\hat{\tau})}, \tag{2.15}$$

where $\hat{\theta}$ and $\hat{\tau}$ are the current estimates of ability and speededness. Since the response time is lognormally distributed, it is easy to show that

$$\mathrm{E}(T_i|\hat{\tau}) = \exp\left(\beta_i - \hat{\tau} + \frac{1}{2\alpha_i^2}\right).$$

This method works great for minimizing total test time; it has been shown to work better than other competing item selection methods for minimizing total test time (Veldkamp, 2016). Examination of Equation 2.15 shows that estimates for both ability, $\hat{\theta}$, and speededness, $\hat{\tau}$, are used in the item selection scheme for MICT. In their method, they used the maximum likelihood estimates for both of these.

## 2.3 Simulation 1: Simulated Item Bank and Examinee Populations

### 2.3.1 Method

Simulation studies are carried out to compare the new MAP estimator with the standard MLE estimator. To determine when using response times improves ability estimation over a standard CAT, several factors are manipulated. First, the person parameters are simulated as

$$\begin{pmatrix} \theta \\ \tau \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\theta\tau} & 1 \end{pmatrix}\right),$$

where $\rho_{\theta\tau}$ is either 0, .25, .50, or .75. For these four conditions, ability and speededness are estimated using the joint MAP. Additionally, a set of persons is simulated with $\theta \sim N(0,1)$, $\tau \sim N(0,1)$, and $\rho_{\theta\tau} = 0$. For this condition, ability (and speededness) is estimated using the MLE estimator. For each of these five conditions, tests have fixed lengths of either 10, 15, or 30 items. Thus, there are a total of $5 \times 3 = 15$ conditions.

A 1000-item bank is simulated with the item parameters $a$, $b$, $c$, $\alpha$, and $\beta$ as defined earlier. Parameters are generated as follows:

- $(a^*, b, \beta) \sim N_3\left(\begin{pmatrix} 0.3 \\ 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.10 & 0.16 & 0.00 \\ 0.16 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{pmatrix}\right)$
  where $a^* = \log a$;
- $c \sim beta(2, 10)$;
- $\alpha \sim unif(1, 4)$.

The choices of distributions here were chosen to simulate items that are commonly found in standardized testing. Because it is known that there is typically a relationship between item difficulty and discrimination parameters, the covariance matrix has a non-zero relationship between $a^*$ and $b$. Furthermore, in the same covariance matrix, the covariance between the item difficulty and time intensity parameters is chosen because it is believed there is a moderate association between these parameters.

Finally, for every factor combination, 2000 examinees are simulated with true ability parameters $\theta$ at evenly spaced increments of 0.10 from -3 to 3. This is done to better assess the performance for typically undersampled points along the distribution. The true speededness parameters $\tau$ are generated from the conditional distribution of $\tau|\theta$. Here, that is given as

$$\tau|\theta \sim N\left(\mu_\tau + \rho_{\theta\tau}\frac{\sigma_\tau}{\sigma_\theta}(\theta - \mu_\theta), \sigma_\tau^2 - \rho_{\theta\tau}^2\sigma_\tau^2\right).$$

The values for the population parameters of $\mu_\theta$, $\mu_\tau$, $\sigma_\theta$ and $\sigma_\tau$ were chosen to standardize the latent variables. They are as follows: $\mu_\theta = \mu_\tau = 0$ and $\sigma_\theta = \sigma_\tau = 1$.

Ability (and speededness) estimation is assessed using bias, RMSE, and correlation. Due to the oversampling at the extremes of the ability distribution, these statistics must be weighted to account for this. Thus, the weighted outcome statistics are calculated as follows:

- Weighted bias:
$$WBias(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J} f(\theta_j)(\hat{\theta}_j - \theta_j)}{\sum_{i=j}^{J} f(\theta_j)}.$$
- Weighted RMSE:
$$WRMSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J} f(\theta_j)(\hat{\theta}_j - \theta_j)^2}{\sum_{j=1}^{J} f(\theta_j)}.$$

- Weighted correlation:

$$WCor(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{WCov(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\left[WVar(\boldsymbol{\theta}, \boldsymbol{\theta})^{1/2}\right]\left[WVar(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})^{1/2}\right]},$$

where

$$WVar(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J} f(\theta_j)\left(\hat{\theta}_j - \frac{1}{J}\sum_{j=1}^{J}\hat{\theta}_j\right)^2}{\sum_{j=1}^{J} f(\theta_j)}$$

and

$$WCov(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J} f(\theta_j)\left(\hat{\theta}_j - \frac{1}{J}\sum_{j=1}^{J}\hat{\theta}_j\right)\left(\theta_j - \frac{1}{J}\sum_{j=1}^{J}\theta_j\right)}{\sum_{j=1}^{J} f(\theta_j)}.$$

Ability estimation is further assessed using bias and RMSE conditioned on the ability level; because these are conditional, these are not weighted.

### 2.3.2 Results

Results of the simulation are summarized in Table 2.1. Bias for $\hat{\theta}$ and $\hat{\tau}$ is very small for all conditions, though it does decrease slightly as the test length increases. As $\rho_{\theta\tau}$ increases for MAP, bias decreases. Also, overall, bias tends to be smaller for MLE than MAP. In contrast, RMSE is smaller for MAP than MLE in every case. As with bias, RMSE decreases as test length increases. Furthermore, as $\rho_{\theta\tau}$ increases, RMSE decreases. Correlation for ability estimation $\text{Corr}(\theta, \hat{\theta})$ follows a similar pattern to RMSE. Across cases, $\text{Corr}(\theta, \hat{\theta})$ is larger for MAP than MLE. Also, as test length increases, the $\text{Corr}(\theta, \hat{\theta})$ increases. Furthermore, as $\rho_{\theta\tau}$ increases, the $\text{Corr}(\theta, \hat{\theta})$ increases. Lastly, correlation for speededness estimation is nearly perfect for all cases.

To further compare how the MAP performs against the MLE, bias and RMSE conditioned on the level of $\theta$ are examined. Figures 2.1 and 2.2 show the conditional bias and RMSE for all conditions. We find that for all conditions, there is a positive bias for values of $\theta$ less than $\mu_\theta = 0$, and a negative bias for values of $\theta$ greater than $\mu_\theta = 0$. As the length of the test increases, bias is lessened across the scale of $\theta$. The MAP has greater bias than the MLE across $\theta$ for all levels of $\rho_{\theta\tau}$, but as $\rho_{\theta\tau}$ increases, this difference decreases. For any given level of $\theta$, RMSE decreases as the test length increases. Also, we find that in the middle of the $\theta$-scale, RMSE is smaller for MAP than for MLE; this is reversed for the ends of the $\theta$-scale. An examination of the locations where

| selection method | $\rho_{\theta\tau}$ | test length | RMSE($\hat{\theta}$) | RMSE($\hat{\tau}$) | Bias($\hat{\theta}$) | Bias($\hat{\tau}$) | Cor($\theta$,$\hat{\theta}$) | Cor($\tau$,$\hat{\tau}$) |
|---|---|---|---|---|---|---|---|---|
| MLE | 0.00 | 10 | 0.37 | 0.12 | 0.012 | 0.0002 | 0.93 | 0.99 |
| MLE | 0.00 | 15 | 0.30 | 0.10 | 0.005 | -0.0003 | 0.96 | 0.99 |
| MLE | 0.00 | 30 | 0.22 | 0.07 | 0.001 | -0.0002 | 0.98 | 0.99 |
| MAP | 0.00 | 10 | 0.35 | 0.12 | 0.028 | -0.0002 | 0.94 | 0.98 |
| MAP | 0.00 | 15 | 0.28 | 0.10 | 0.020 | -0.0000 | 0.96 | 0.99 |
| MAP | 0.00 | 30 | 0.21 | 0.07 | 0.011 | -0.0001 | 0.98 | 0.99 |
| MAP | 0.25 | 10 | 0.34 | 0.12 | 0.029 | 0.0002 | 0.94 | 0.99 |
| MAP | 0.25 | 15 | 0.28 | 0.10 | 0.020 | 0.0003 | 0.96 | 0.99 |
| MAP | 0.25 | 30 | 0.21 | 0.07 | 0.011 | 0.0001 | 0.98 | 0.99 |
| MAP | 0.50 | 10 | 0.33 | 0.12 | 0.025 | 0.0015 | 0.94 | 0.99 |
| MAP | 0.50 | 15 | 0.28 | 0.10 | 0.018 | 0.0005 | 0.96 | 0.99 |
| MAP | 0.50 | 30 | 0.21 | 0.07 | 0.011 | 0.0002 | 0.98 | 0.99 |
| MAP | 0.75 | 10 | 0.31 | 0.12 | 0.020 | 0.0017 | 0.95 | 0.99 |
| MAP | 0.75 | 15 | 0.27 | 0.10 | 0.015 | 0.0011 | 0.96 | 0.99 |
| MAP | 0.75 | 30 | 0.20 | 0.07 | 0.009 | 0.0006 | 0.98 | 0.99 |

**Table 2.1:** Results for Simulation 1.

RMSE is lower for MAP shows that for 95% of the population of test-takers, the MAP outperforms the MLE in terms of RMSE. This effect is especially true for short tests. Lastly, as $\rho_{\theta\tau}$ increases, RMSE decreases across the $\theta$-scale.

## 2.4 Simulation 2: Real Item Bank and Examinee Population

### 2.4.1 Method

Simulation studies with a real item bank are carried out to determine when using the joint MAP estimator improves ability estimation over the MLE. Several factors are manipulated. First, the estimation method is either the joint MAP or the MLE. Second, the item selection method is either the MIC or the MICT. Since both latent variables $\theta$ and $\tau$ are used in item selection for MICT, it would be interesting to see how having better estimates of both $\theta$ and $\tau$ affects the test. Third, the tests have fixed lengths of either 10, 15, or 30.

The item bank comes from a data set of a real high-stakes, large-scale standardized CAT. The data consist of raw responses and RTs from about 2000 examinees with an item pool containing about 500 multiple-choice items that were pre-calibrated according to 3PLM. The lognormal model item parameters $\alpha$ and $\beta$ were estimated using a modified version of van der Linden's (2007) MCMC routine that fixed the 3PLM item parameters to the pre-calibrated values, and the distribution of

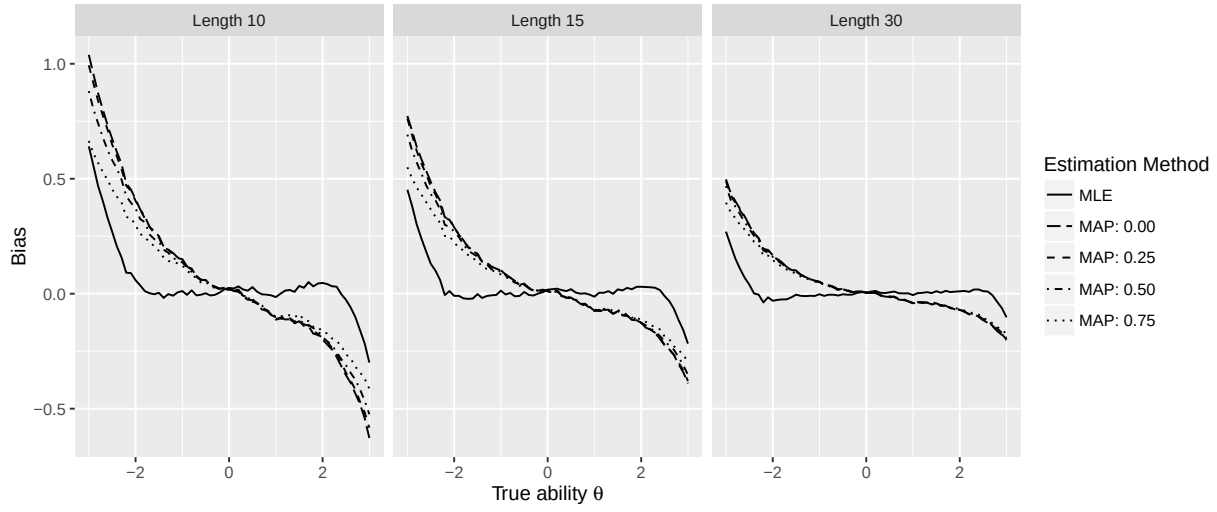


**Figure 2.1:** Plot of conditional bias of estimates by true ability for Simulation 1.

$\tau$ was set to have a mean of 0. All parameters appeared to converge using 10000 MCMC draws with a burn-in size of 5000.

Examinees are simulated as in Simulation 1, except the values for the population parameters of $\mu_\theta$, $\mu_\tau$, $\sigma_\theta$, $\sigma_\tau$, and $\rho_{\theta\tau}$ were chosen to reflect the population of test-takers from the real data set. They are as follows: $\mu_\theta = 0.5$, $\mu_\tau = 0$, $\sigma_\theta = 1$, $\sigma_\tau = 0.16$, and $\rho_{\theta\tau} = 0.76$.

Bias, RMSE, and correlation are used to assess ability and speededness estimation. As discussed previously, the oversampling at the extremes of the ability distribution must be taken into account in these statistics; they are weighted as in in Simulation 1. Furthermore, ability estimation is assessed with a combination of (unweighted) bias and RMSE conditioned on the ability level. Lastly, mean overall test time across individuals and the standard deviation of test times across individuals is assessed.

### 2.4.2 Results

Results of the simulation are summarized in Table 2.2. Bias for $\hat{\theta}$ and $\hat{\tau}$ is very small for all conditions—for $\hat{\theta}$ absolute bias is between 0.001 and 0.041, and for $\hat{\tau}$ absolute bias is between 0.0002 and 0.0068—though it does seem to decrease slightly as the test length increases. Interestingly, for MLE bias is *larger* for MICT than MIC, and for MAP bias is *smaller* for MICT than MIC; this effect is the case for both $\theta$ and $\tau$. Furthermore, overall, bias tends to be smaller for MLE

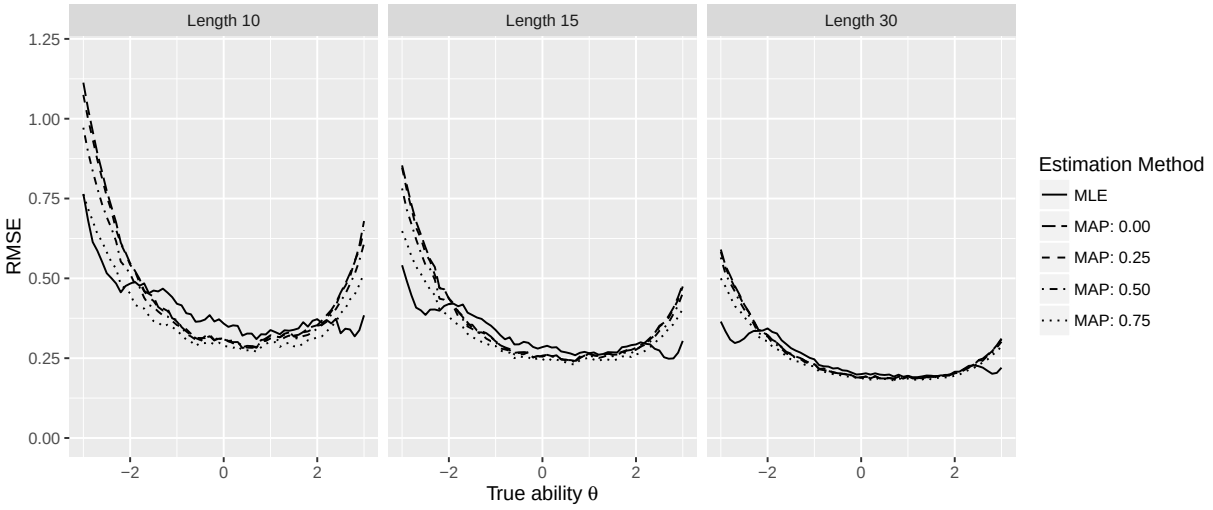


**Figure 2.2:** Plot of conditional RMSE of estimates by true ability for Simulation 1.

| selection method | estimation method | test length | RMSE($\hat{\theta}$) | RMSE($\hat{\tau}$) | Bias($\hat{\theta}$) | Bias($\hat{\tau}$) | Cor($\theta$,$\hat{\theta}$) | Cor($\tau$,$\hat{\tau}$) | mean test time | SD test time |
|---|---|---|---|---|---|---|---|---|---|---|
| MIC | MLE | 10 | 0.64 | 0.19 | 0.007 | 0.0002 | 0.83 | 0.63 | 18.59 | 5.93 |
| MIC | MLE | 15 | 0.53 | 0.15 | -0.004 | -0.0003 | 0.87 | 0.70 | 27.78 | 7.61 |
| MIC | MLE | 30 | 0.39 | 0.11 | -0.008 | 0.0028 | 0.93 | 0.82 | 57.32 | 12.15 |
| MIC | MAP | 10 | 0.49 | 0.09 | 0.041 | 0.0068 | 0.87 | 0.80 | 19.12 | 6.14 |
| MIC | MAP | 15 | 0.42 | 0.08 | 0.031 | 0.0042 | 0.91 | 0.83 | 28.47 | 7.81 |
| MIC | MAP | 30 | 0.32 | 0.07 | 0.018 | 0.0011 | 0.95 | 0.88 | 57.86 | 12.33 |
| MICT | MLE | 10 | 0.68 | 0.21 | -0.011 | 0.0007 | 0.81 | 0.59 | 12.27 | 4.04 |
| MICT | MLE | 15 | 0.57 | 0.17 | -0.018 | 0.0010 | 0.86 | 0.67 | 18.29 | 5.06 |
| MICT | MLE | 30 | 0.42 | 0.11 | -0.014 | 0.0006 | 0.92 | 0.80 | 37.90 | 8.24 |
| MICT | MAP | 10 | 0.52 | 0.09 | 0.023 | 0.0053 | 0.86 | 0.79 | 12.06 | 3.83 |
| MICT | MAP | 15 | 0.45 | 0.09 | 0.017 | 0.0038 | 0.90 | 0.82 | 17.90 | 4.80 |
| MICT | MAP | 30 | 0.34 | 0.07 | 0.013 | 0.0016 | 0.94 | 0.88 | 37.28 | 7.44 |

**Table 2.2:** Results for Simulation 2.

than MAP. In contrast, RMSE is smaller for MAP than MLE in every case. Unlike bias, RMSE is larger for MICT than MIC for both MLE and MAP, however the increase in RMSE from MIC to MICT for the MAP estimator is minimal at all test lengths. Furthermore, for all cases, RMSE decreases as test length increases. Lastly, correlation follows a similar pattern to RMSE. Across cases, correlation is smaller for MICT than MIC, and larger for MAP than MLE. Interestingly, the increase in correlation for MAP over MLE is fairly large, especially for speededness. For instance, for a small test length of 10 with items selected via MICT, the correlation of the true value and the estimate of speededness is 0.59 for MLE and 0.79 for MAP. Lastly, as test length increases, the correlation increases.

One of the main goals of using the MICT to select items is to decrease the test times for individuals. As such, it is important to investigate the mean test times across individuals. Furthermore, it is also important to compare the standard deviations of the test times across individuals as an

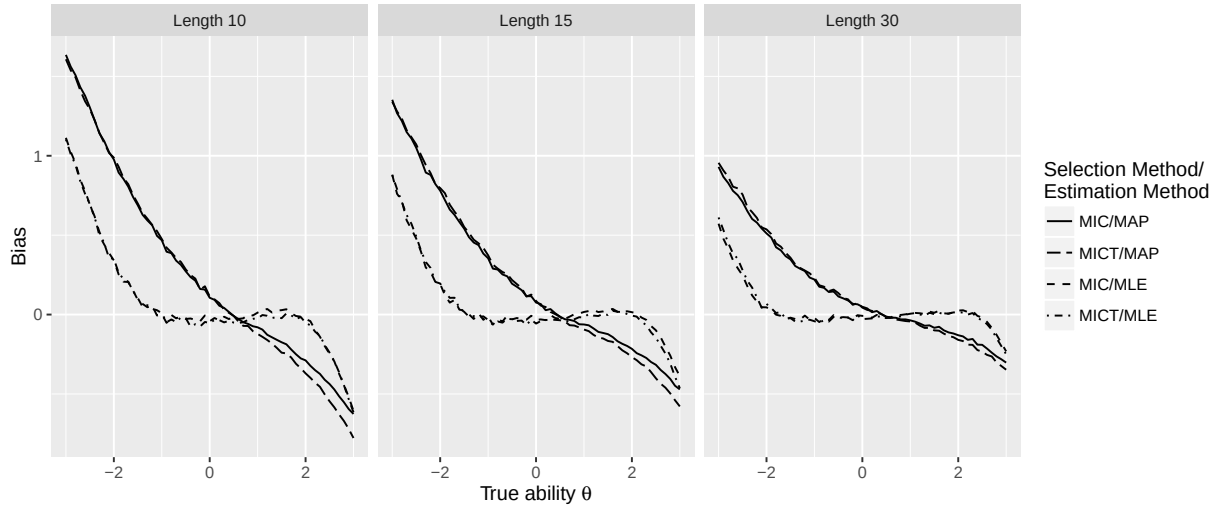


**Figure 2.3:** Plot of conditional bias of estimates by true ability for Simulation 2.

index of differential speededness. Obviously, as the number of items increases, the mean test time increases; the standard deviation of the test times also increases with the test length. As expected, MICT has much smaller mean test times than MIC for all conditions as well as smaller standard deviations of test times. Interestingly, within the MIC conditions, MLE has *smaller* mean test times and SD test times than MAP, whereas within the MICT conditions, mean test times and SD test times are *larger* for MLE than MAP.

To further compare how the MAP performs against the MLE, bias and RMSE conditioned on the level of $\theta$ are examined. Figures 2.3 and 2.4 show the conditional bias and RMSE for all conditions. We find that for all conditions, there is a positive bias for values of $\theta$ less than $\mu_\theta = 0.5$, and a negative bias for values of $\theta$ greater than $\mu_\theta = 0.5$. As the length of the test increases, bias is lessened across the scale of $\theta$. The MAP has greater bias than the MLE across $\theta$. The difference in conditional bias between the MIC and MICT for each estimation method is very small. For any given level of $\theta$, RMSE decreases as the test length increases. Also, we find that in the middle of the $\theta$-scale, RMSE is smaller for MAP than for MLE; this is reversed for the ends of the $\theta$-scale. This means that for 90% of the population of test-takers, the MAP outperforms the MLE. This effect is especially pronounced for short tests. Lastly, though the effect is small, RMSE is larger across the $\theta$-scale for MICT than for MIC.

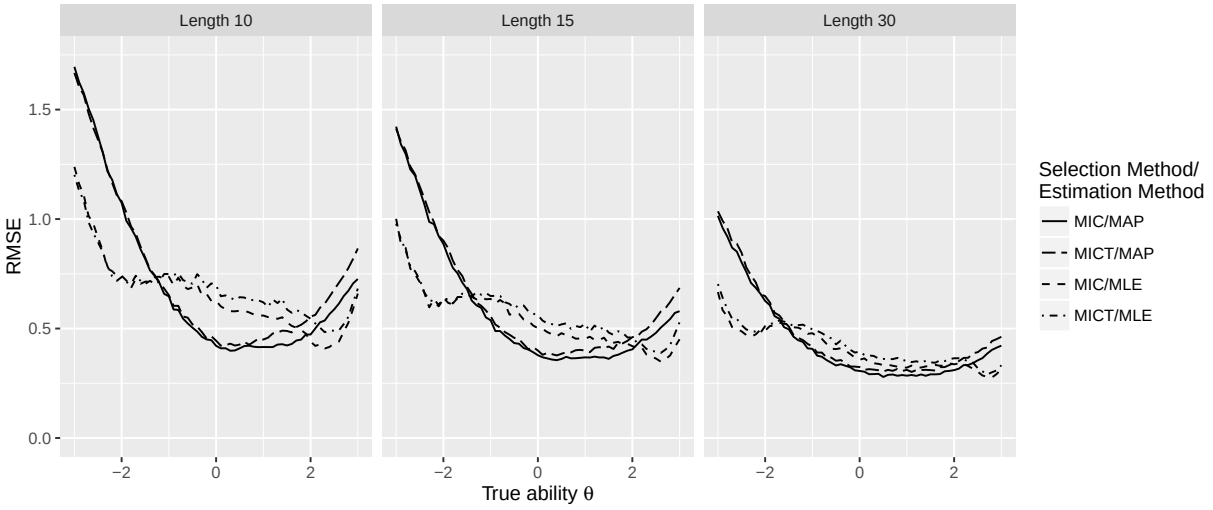


**Figure 2.4:** Plot of conditional RMSE of estimates by true ability for Simulation 2.

## 2.5 Discussion

Large-scale, high-stakes admissions exams, such as the GMAT, have one basic goal in mind: to accurately measure an examinee's true score with fewer items than needed for a paper-and-pencil test. It is also possible to decrease the amount of real time needed to finish a test. This study proposes a method that can improve on both tasks in a simple fashion, with minimal extra cost. Of course, this is all predicated on the assumption that ability and speededness of an individual are related in the context of a high-stakes exam, such as in our real data example; if they are, then that relationship can be exploited.

From the results, we see that using response time information improves estimation for both $\theta$ and $\tau$ with respect to RMSE. More to the point, as the relationship between $\theta$ and $\tau$ becomes stronger, the gains in efficiency with respect to RMSE increases; this is especially true for short tests. While the increase in performance is largely dependent on knowing the true correlation of ability and speededness, in an actual operational test, a good estimate of this correlation should emerge as the number of examinees increases. It could be recommended that a standard maximum likelihood estimation routine be used (that is, without using response times to estimate ability) while simultaneously estimating speededness independently of ability to calibrate the system and obtain a good estimate of their correlation.

The real data simulation demonstrates two things of particular interest to test creators. First,

if the goal is simply to have an accurate test, the MLE will result in a small RMSE and bias for $\theta$, with a fairly high correlation between $\theta$ and $\hat{\theta}$, particularly in conjunction with MIC. Of course, since absolute bias is very small for all conditions (at most 0.041), it is not necessary to rule out MAP in this situation. Indeed, at the expense of higher bias, MAP yields a smaller RMSE($\hat{\theta}$) for 10 items (RMSE = 0.49) than MLE does for 15 items (RMSE = 0.53). And even though mean testing time is not a consideration, certainly having fewer items would result in shorter test times; this is borne out in Table 2.2. On the other hand, if the dual objective of having an accurate ability estimate and decreasing overall test-taking time is desired, then using MICT clearly beats out using MIC for item selection, with lower test times across the board. Furthermore, using MAP for estimation alongside MICT is the best option. Not only does it have smaller RMSE for $\theta$ and shorter mean test times, but it also has similar levels of bias for $\theta$, negating the main advantage of using MLE. Furthermore, MAP-MICT has the lowest standard deviations of test times for all conditions, which signifies that there is less variability in test times amongst the examinees. One reason this is desirable is because there would be less perceived unfairness amongst the test-takers, which may allay the feelings of anxiety that they feel when someone finishes before themselves. From the test-maker's perspective, this is great for two interrelated reasons. First, the test-maker will have an easier time setting a reasonable time limit within which most test-takers will complete the test. Second, there should be less of an effect on the test scores due to examinees coming up against the test's time limit; if a reasonable time limit is set, then the examinee should not come dangerously close to the time limit and alter their test-taking behavior. Whenever this happens, the test score will reflect not only ability, but also their changing speededness towards the end of the test, reducing the effectiveness of the test scores as a measure of ability.

Lastly, while not the main focus, it is important to discuss the implications of using Equation 2.13 (minimum standard error; MinSE) to select items as opposed to either the MIC or MICT methods. I did a simulation to compare MIC with MinSE and found that the methods were nearly identical. Because the methods were so similar, the results of that simulation are not shown here. However, some knowledge may be gleaned from this. One thing to note is that the MinSE method was suggested by means of the asymptotic posterior variance of $\theta$ with response times.

# Chapter 3

# Using GMICT to Select Items With the MAP Estimator

## 3.1 Introduction

In Chapter 2, it was shown that ability $\theta$ can be more effectively estimated, when estimated jointly with speededness $\tau$ than by itself. Similarly, speededness is effectively estimated in this framework as well. These effects are more pronounced as the true correlation between $\theta$ and $\tau$ increases. Furthermore, it was shown that the MAP is especially effective in terms of increasing estimation accuracy and decreasing overall test time when items are selected via the MICT approach. It seems reasonable, then, given these results that any method that makes use of both of these should be more effective when estimated jointly.

In Section 1.4.2, the generalized time-weighted maximum information criterion (GMICT; Choe & Kern, 2014) was introduced. It is a generalization of the MICT method in (2.15) that was shown to effectively choose items so that $\theta$ estimation accuracy is high, test security is controlled, differential speededness is controlled, and overall test time is minimized (Choe & Kern, 2014). It is reproduced here as follows:

$$\mathrm{IT}_i^G = \frac{I_i(\hat{\theta})}{|\mathrm{E}(T_i|\hat{\tau}) - v|^w}. \tag{3.1}$$

This method generalizes MICT by including a centering value $v$ that allows for more control over the test time, and by also including an exponent weighting parameter $w$ for control over how the deviance of the expected response time from $v$ is weighted. Combined, this item selection criterion provides much more flexibility over how items are selected in the face of the competing goals of score reliability, overall test-taking time, and overall test security. Note that the standard MIC ($w = 0$) and MICT ($w = 1$, $v = 0$) both are special cases of GMICT.

As the effects are dependent on the estimated values for $\theta$ and $\tau$, it stands to reason that better

estimated latent traits will result in better control. Thus, in this chapter, the effectiveness of using the GMICT for item selection with the ability and speededness parameters estimated jointly is investigated using both a simulated item bank and a real item bank.

## 3.2 Simulation 1: Simulated Item Bank and Examinee Populations

### 3.2.1 Method

Simulation studies are carried out to compare the new MAP estimator with the standard MLE estimator. To determine when using response times improves ability estimation over a standard CAT, several factors are manipulated. First, the person parameters are simulated as

$$\begin{pmatrix} \theta \\ \tau \end{pmatrix} \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\theta\tau} & 1 \end{pmatrix} \right),$$

where $\rho_{\theta\tau}$ is either 0, .25, .50, or .75. In the CAT, ability and speededness are estimated either using the joint MAP or MLE. Tests have fixed lengths of either 10, 20, or 30 items. Finally, items are selected using several combinations of $v$ and $w$ in GMICT selection procedure; here $v$ ranges from 0 to 3 in 0.1 increments and $w$ is either 0.50, 0.75, or 1.00. For each factor combination, 10 replications of 2000 examinees are simulated.

A 500-item bank is simulated with the item parameters $a$, $b$, $c$, $\alpha$, and $\beta$ as defined earlier. Parameters are generated as follows:

- $(a^*, b, \beta) \sim N_3 \left( \begin{pmatrix} 0.3 \\ 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.10 & 0.16 & 0.00 \\ 0.16 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{pmatrix} \right)$
  where $a^* = \log a$;
- $c \sim beta(2, 10)$;
- $\alpha \sim unif(1, 4)$.

The choices of distributions here were chosen to simulate items that are commonly found in standardized testing. Because it is known that there is often a relationship between item difficulty and discrimination parameters, the covariance matrix has a non-zero relationship between $a^*$ and $b$. Furthermore, in the same covariance matrix, the covariance between the item difficulty and time intensity parameters is chosen because it is believed there is a moderate association between these parameters. If $J$ is the number of examinees, then ability estimation is assessed using bias, RMSE, and correlation:

- Bias:
$$Bias(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J},$$
- RMSE:
$$RMSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J},$$
- Correlation:
$$Cor(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{Cov(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\left[Var(\boldsymbol{\theta})^{1/2}\right]\left[Var(\hat{\boldsymbol{\theta}})^{1/2}\right]}.$$

To further assess the effectiveness of the test, mean test time (MTT), standard deviation of test time (STT), and the test overlap rate.

### 3.2.2 Results

The results of the simulation on estimation bias, RMSE, and correlation of ability are given in Figures 3.1, 3.2, and 3.3, respectively. Several conclusions can be made. First, it is clear that when choosing items using GMICT all combinations of $w$ and $v$, absolute bias increases, RMSE increases, and the correlation between estimated $\theta$ and true $\theta$ decreases compared to choosing items using MIC; that is, using GMICT necessarily worsens the accuracy of estimating ability. Conversely, the test overlap drops, MTT decreases, and STT decreases for GMICT compared to MIC for most values of $v$. Across values of $w$, the general trend that using GMICT over MIC has on all measures is present, but to varying degrees. Essentially, just as it is designed to do, as $w$ increases, GMICT places more weight on selecting items according to the absolute deviation of expected response time from $v$ relative to Fisher information of $\hat{\theta}$, meaning that items are selected with increasing

importance on minimizing response times. Thus, as $w$ increases, estimation of ability worsens, but MTT, STT, and test overlap all decrease. Simulation results on test overlap rate, MTT, and STT are given in Figures 3.4, 3.5, and 3.6, respectively.

For all levels of $\rho_{\theta\tau}$ and test lengths, the general effect of selection method is true, though the effects may be moderated. For instance, as test length increases, the effect of item selection method on ability estimation decreases. Furthermore, increasing test length has an obvious impact of increasing MTT and STT, while an increasing $\rho_{\theta\tau}$ very slightly decreases MTT and STT. On the other hand, test length very slightly increases test overlap, while $\rho_{\theta\tau}$ has no discernible effect on overlap rates.

Lastly, and most importantly for this study, the choice of estimator has a noticeable impact on the outcomes of interest. First, independently of selection method, using the MAP decreases RMSE over using the MLE while increasing bias. Additionally, the correlation between true and estimated ability is higher for MAP. This effect is mitigated by test length, as well as by $\rho_{\theta\tau}$; as test length increases, the difference in RMSE, bias, and correlation between the estimation methods lessens, but as the $\rho_{\theta\tau}$ increases, the effect gets larger. Thus, the methods have the largest difference when the strength of relationship between ability and speededness is largest and test length is short, and have the smallest difference when ability and speededness are unrelated and the test is long. Interestingly, MTT and test overlap do not seem to be affected by estimator. For STT, on the other hand, the effect of estimator is unclear. For some combinations of test length and $\rho_{\theta\tau}$ (length of 30, $\rho_{\theta\tau} = 0.25$ and 0.50), STT noticeably decreases, but for all others the effect is negligible.

## 3.3 Simulation 2: Real Item Bank and Examinee Population

### 3.3.1 Method

Simulation studies with a real item bank are carried out to determine how the joint MAP estimator performs against the MLE estimator when selecting items using the GMICT in a situation closer to a real operational setting. Several factors are manipulated. First, estimation method is either the joint MAP or the MLE. Second, items are selected using several combinations of $v$ (0 to 3 in 0.1 increments) and $w$ (0.50, 0.75, or 1.00) in GMICT selection procedure. Third, the tests

have fixed lengths of either 10 or 30.

The item bank comes from a data set of a real high-stakes, large-scale standardized CAT. The data consists of raw responses and RTs from about 2000 examinees with an item pool containing about 500 multiple-choice items that were pre-calibrated according to 3PLM. The lognormal model item parameters $\alpha$ and $\beta$ were estimated using a modified version of van der Linden's (2007) MCMC routine that fixed the 3PLM item parameters to the pre-calibrated values, and the distribution of $\tau$ was set to have a mean of 0. A check of trace plots showed that all parameters appeared to converge using 10000 MCMC draws with a burn-in size of 5000.

Examinee ability and speededness parameters are simulated from a multivariate normal distribution with population parameters of $\mu_\theta$, $\mu_\tau$, $\sigma_\theta$, $\sigma_\tau$, and $\rho_{\theta\tau}$ chosen to reflect the population of test-takers from the real data set. They are as follows: $\mu_\theta = 0.5$, $\mu_\tau = 0$, $\sigma_\theta = 1$, $\sigma_\tau = 0.16$, and $\rho_{\theta\tau} = 0.76$. Ten replications of 1000 examinees were carried out for each condition.

Bias, RMSE, and correlation are used to assess ability estimation. Furthermore, test performance is assessed using mean overall test time across individuals (MTT), the standard deviation of test times across individuals (STT), and the test overlap rate.

### 3.3.2 Results

The results of the simulation on estimation bias, RMSE, and correlation of ability are given in Figures 3.7, 3.8, and 3.9, respectively. Many of the conclusions from Section 3.2 are similar, which is promising. In particular, using GMICT increases bias, increases RMSE, and decreases the correlation in estimating ability over MIC; this is true for all values of $w$ and $v$. Thus, using GMICT makes ability estimation worse compared to MIC. On the other hand, using GMICT makes the test overlap rate drop for large enough values of $v$, and MTT and STT decrease for small enough values of $v$. The results of these are given in Figures 3.10, 3.11, and 3.12 for test overlap, MTT, and STT, respectively. As can be seen, these effects are true for all values of $w$. That said, the magnitude of the results are moderated by $w$; larger values for $w$ makes the difference between GMICT and MIC larger.

The length of the test has the expected effect on the measures. First, the estimation of ability is more accurate for a longer test—bias decreases, RMSE decreases, and correlation increases.

Furthermore, the overlap rate increases just a little. Lastly, the mean test time increases and the SD of test time increases. None of these results are surprising.

The effect of the estimator is very similar to those in Section 3.2. First, bias increases, RMSE decreases, and correlation increases when using MAP over MLE. Thus, while estimation accuracy gets worse in one way, it gets better in other ways. This is pretty typical effect for Bayesian methods. Furthermore, MTT and STT are incredibly similar for the two estimators, with MTT being nearly identical. STT has a little more variability in the results, so it is hard to say with certainty, but it looks like for a small value of $v$ in GMICT, MAP has smaller STT, and for a large value of $v$, MAP has a larger STT. For the MIC, STT is smaller for MAP. The biggest difference from Simulation 1 is that unlike previous results, the test overlap rate is larger for the MAP than the MLE.

## 3.4 Discussion

The main question for this chapter is whether the joint MAP estimator should be used in conjunction with a method that uses ability and speededness parameters in item selection. The answer to this question largely depends on two things: is the trade-off of larger bias and smaller RMSE worth it; and is estimation accuracy, test security, or controlling differential speededness/testing time more important.

Let us first tackle the first question: is the trade-off of larger bias and smaller RMSE worth it? To answer this, it is important to understand where this comes from. As mentioned earlier, this is a pretty standard effect of using a Bayesian method vs. a maximum likelihood method. It is indicative of a bias–variance trade-off. Basically, mean-squared error can be decomposed into two additive parts, squared bias and variance; that is,

$$\text{MSE}(\hat{\theta}) = \text{bias}(\hat{\theta}, \theta)^2 + \text{Var}(\hat{\theta}),$$

meaning that

$$\text{Var}(\hat{\theta}) = \text{MSE}(\hat{\theta}) - \text{bias}(\hat{\theta}, \theta)^2.$$

Thus, if bias increases, but RMSE decreases, then variance must necessarily have decreased. The

question becomes, then, is the overall increase in precision (i.e., (R)MSE) worth the increase in bias. Here, I would say, yes, since the absolute magnitude of bias is so small—it is only about 0.06 at its highest—then decreasing variability, in the end, will have a greater effect on overall outcomes. Essentially, while the estimator is "wrong on average" (bias is larger), it will be "close to right" (bias is only slightly larger) most of the time (variance is smaller). This is captured by the fact that the MSE for the biased estimator (MAP) is smaller than the MSE of the unbiased estimator (MLE). Since we only have one shot at estimating a person's ability level, being close to truth more often than a small bit closer to truth less often is desirable.

The answer to the second question is largely dependent on the purpose of the test. On one hand, it is always desirable to have more accurate estimation results. On the other, there are practical global test constraints: time-limits and test security. To some major extent, time-limits, while practically necessary, can have serious negative effects on measurements. For instance, as an examinee closes in on a time-limit, oftentimes he will speed up to try to complete the test, which will, due to the well-known speed–accuracy trade-off effect, cause the person to get more items wrong. This, in turn, increases measurement error, thereby decreasing the accuracy of determining the ability-level of the examinee. Decreasing the overall mean test time and standard deviation of test times for examinees will mean fewer examinees come against the time-limit. Thus, the effect due to time-limits is minimized, which increases accuracy of determining ability-levels of examinees. At the same time, test security is an important concern. One important way to help maintain test security is to make the distribution of test items across examinees as uniform as possible; this can be done by making the test overlap between examinees small.

What has been shown is that the choice of estimator has little effect on mean test times and the variability in test times given a particular selection routine (GMICT or MIC). Test overlap, on the other hand, seems to be dependent on estimator: MAP has higher test overlap rates than MLE does in the real item bank case. Curiously, this same effect was not present for the simulated item bank. While not investigated here, possible reasons should be discussed. First, investigation of the real item bank shows that a large positive correlation between $\ln \alpha$ and $\beta$ exists ($\rho_{\ln \alpha, \beta} = 0.53$). This correlation was not modeled in the simulated item bank. This doesn't seem terribly likely to have caused this difference. The other possibility, on the other hand, does. It was found that

the distributions of $\theta$ and $\beta$ do not match in the real item bank simulation ($\mu_\theta = 0.5$, $\mu_\beta = 0$, $\sigma_\theta = 1$, $\sigma_\beta = 1.11$). Thus, there are fewer possible items to match with levels of ability across the scale of $\theta$. This may be exacerbated by the values of $\theta$ being pulled toward the prior distribution of $\theta$ in the MAP estimator, causing the increase in test overlap. With this in mind, test overlap is still greatly reduced for GMICT compared to MIC for either estimator. Since the accuracy of ability estimation is of primary concern, it seems that using MAP with GMICT is the best of all competing worlds; it has great accuracy (nearly as good as with MIC), much lower test overlap, low mean test times, and low variability in test times. Thus, it should be recommended that if GMICT is to be used, it is best to use the MAP estimator in conjunction with it.

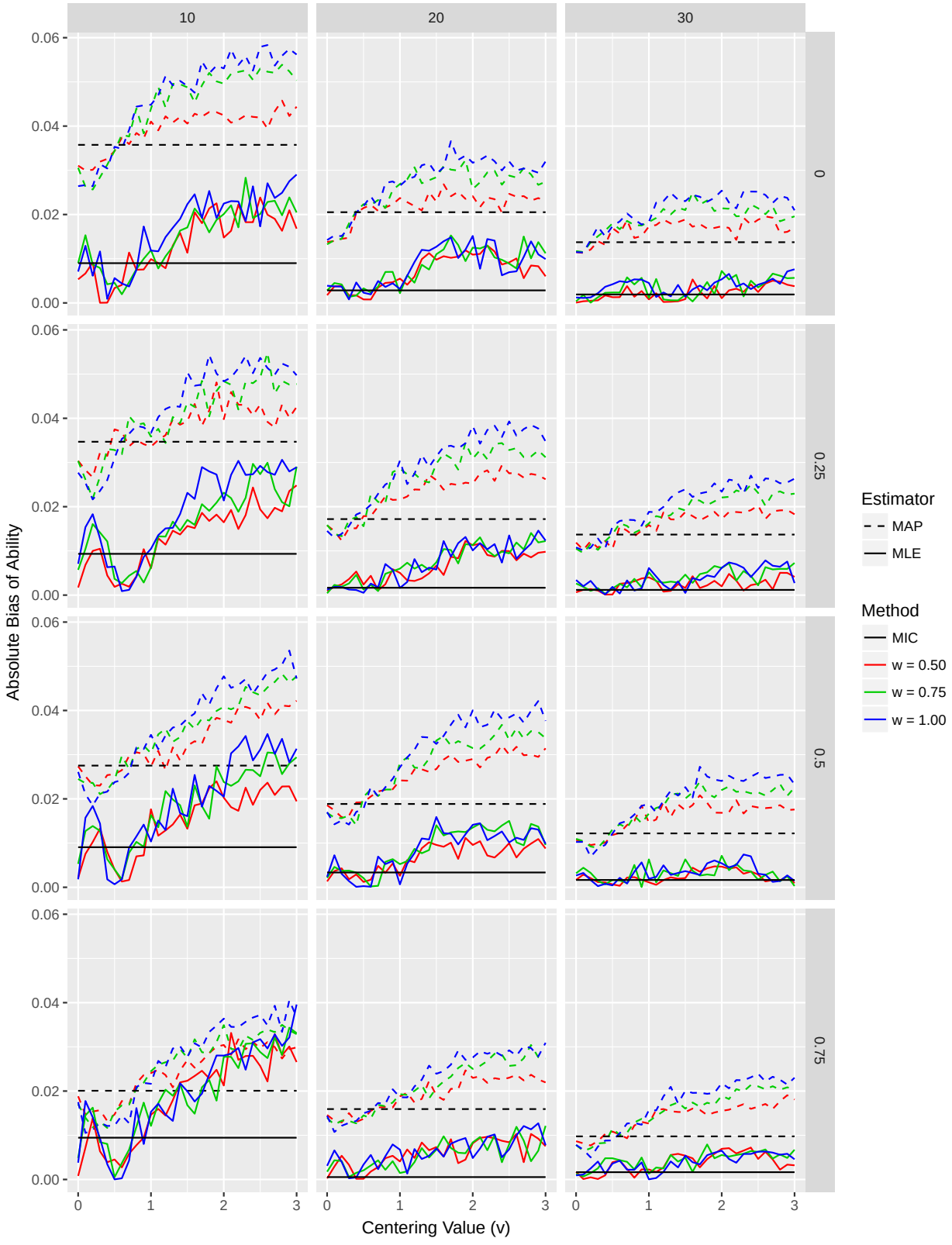


**Figure 3.1:** Plot of bias of ability estimates for Simulation 1.

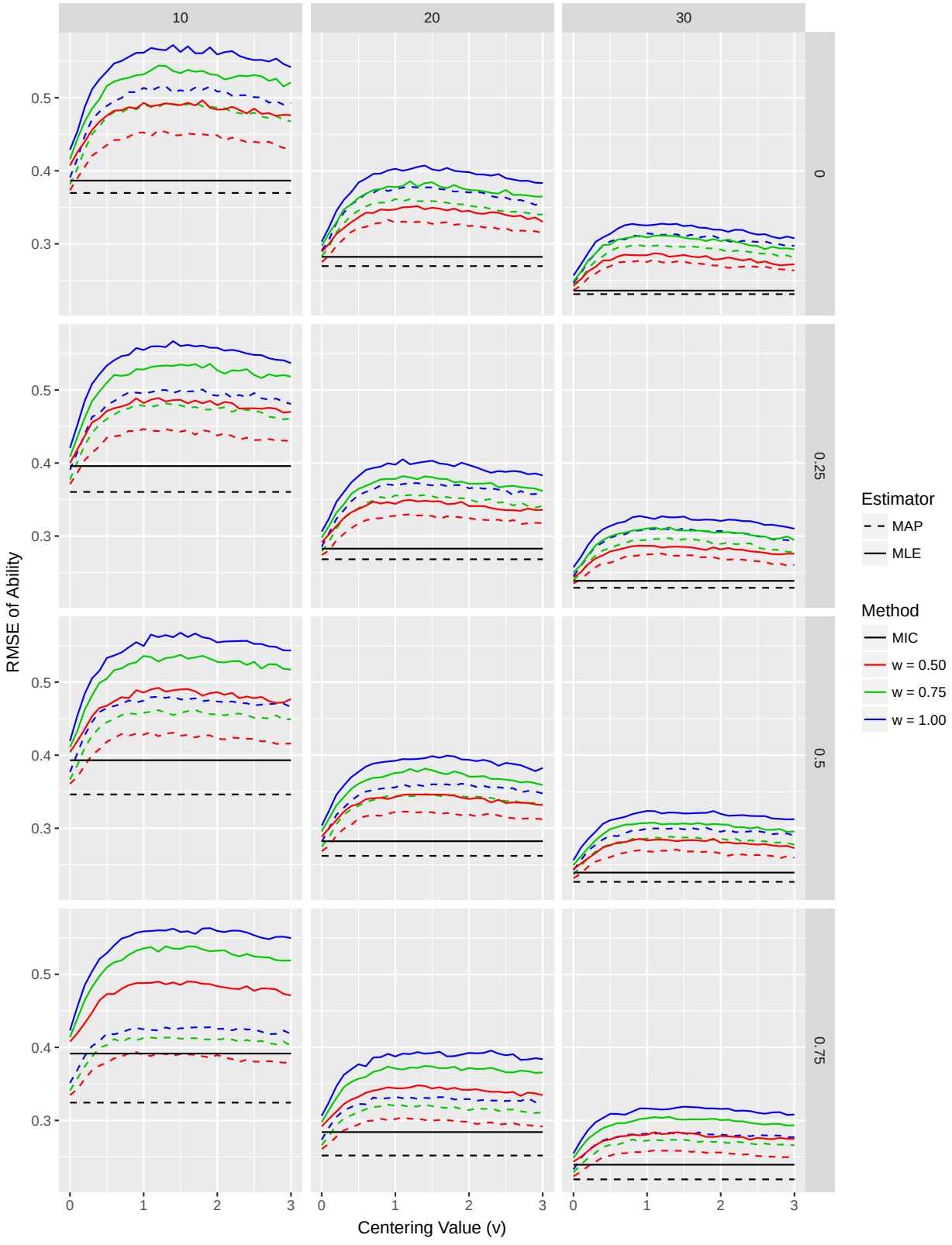


**Figure 3.2:** Plot of RMSE of ability estimates for Simulation 1.

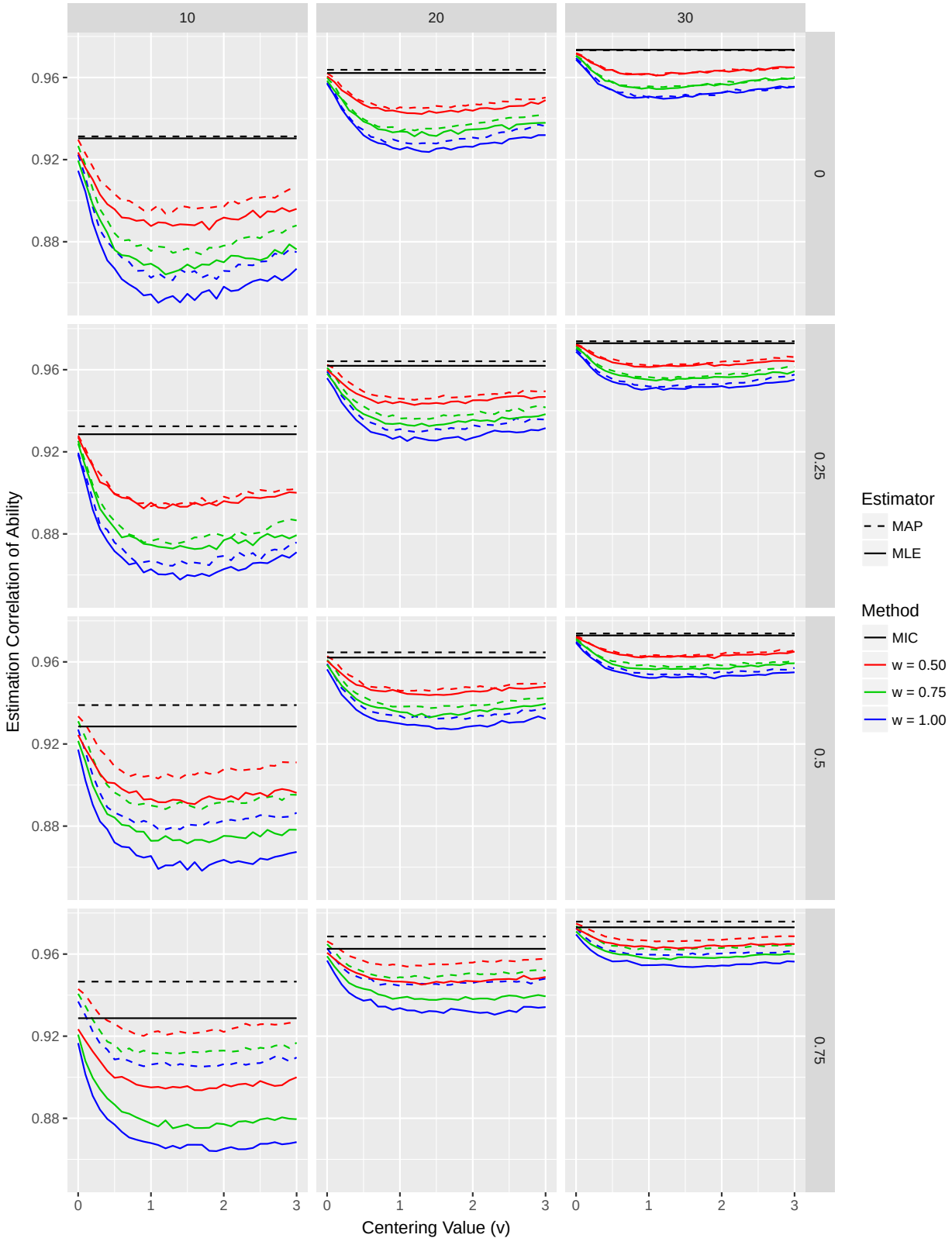


**Figure 3.3:** Plot of correlation of ability estimates with true values for Simulation 1.

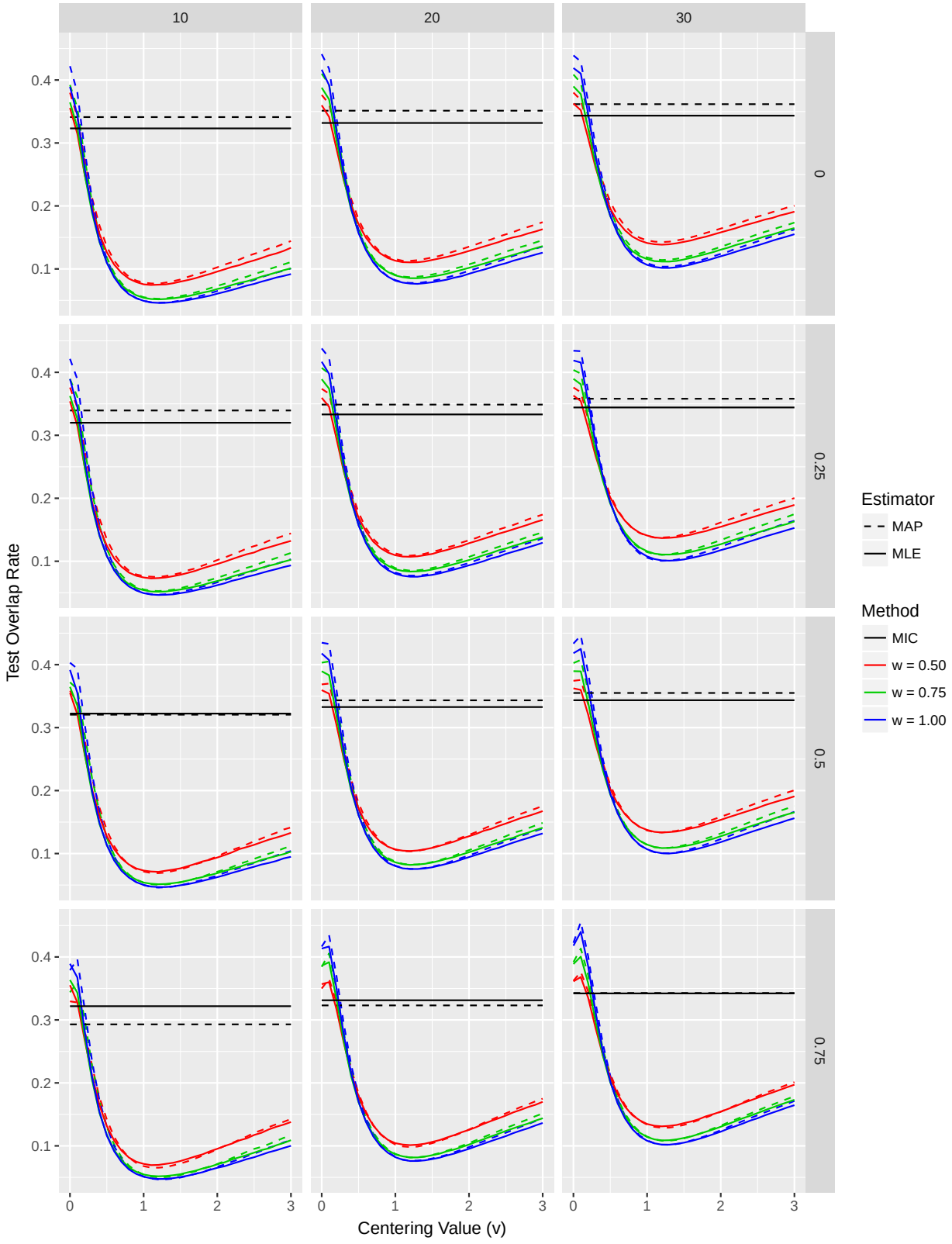


**Figure 3.4:** Plot of test overlap rate for Simulation 1.

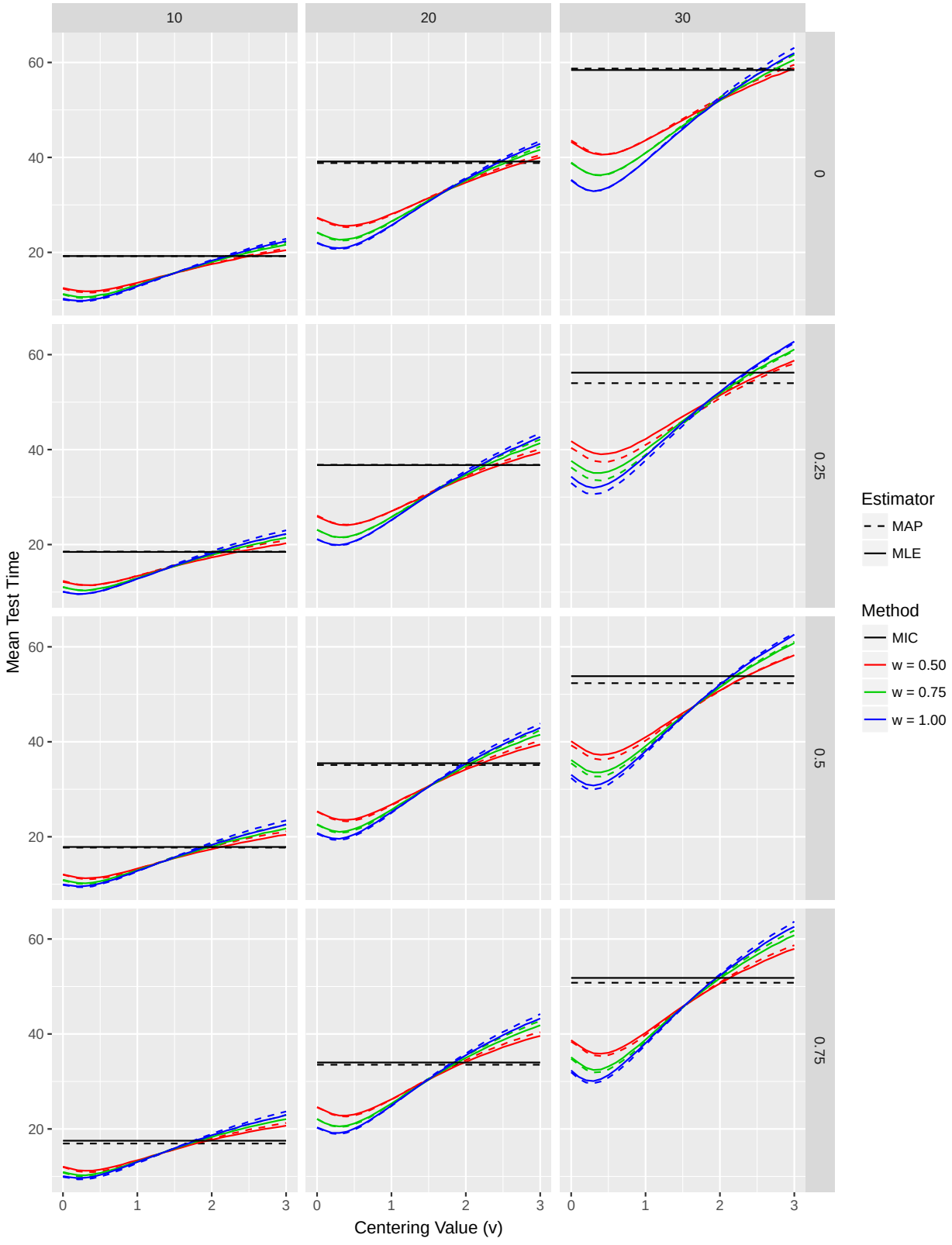


**Figure 3.5:** Plot of mean test time for Simulation 1.

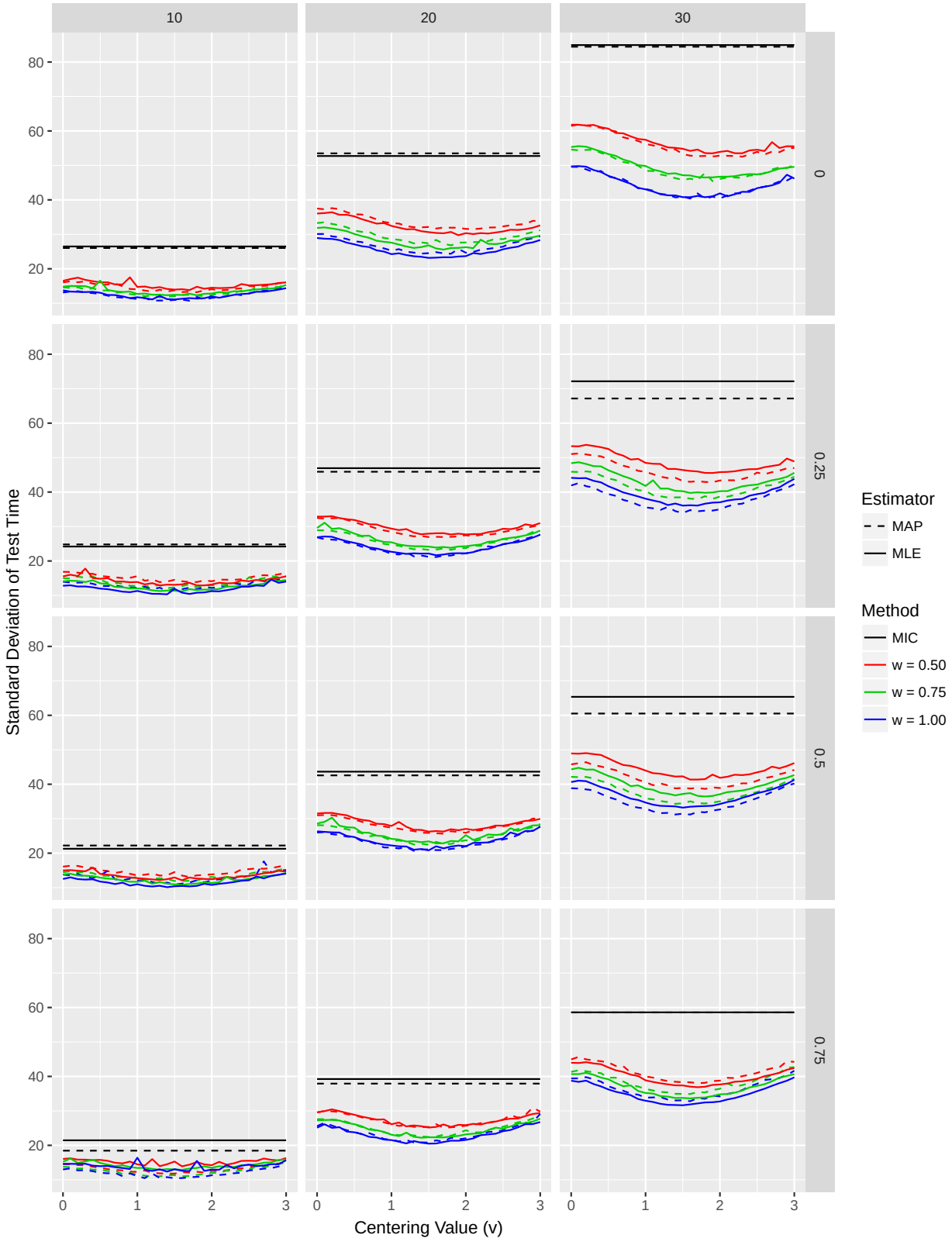


**Figure 3.6:** Plot of SD of test time for Simulation 1.

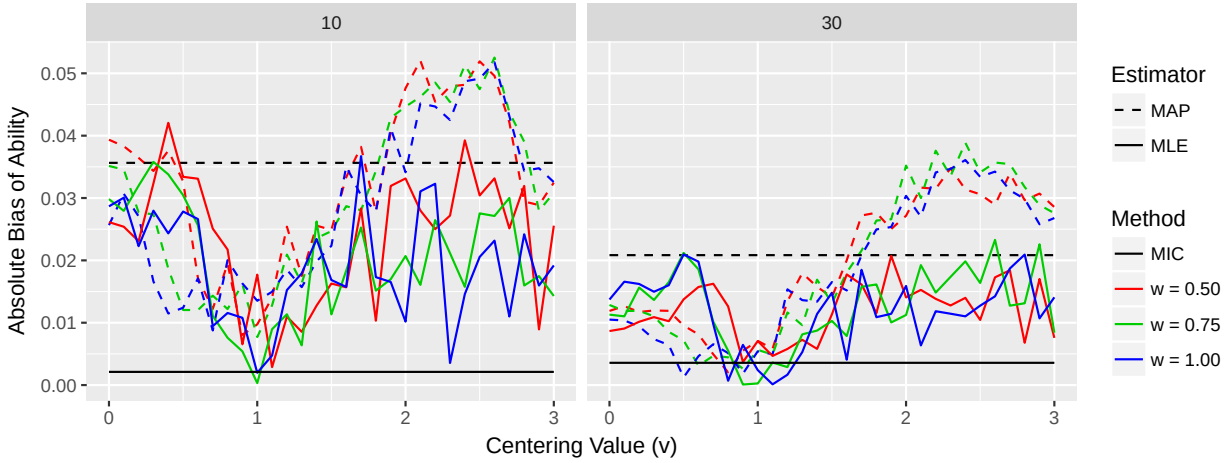


**Figure 3.7:** Plot of bias of ability estimates for Simulation 2.

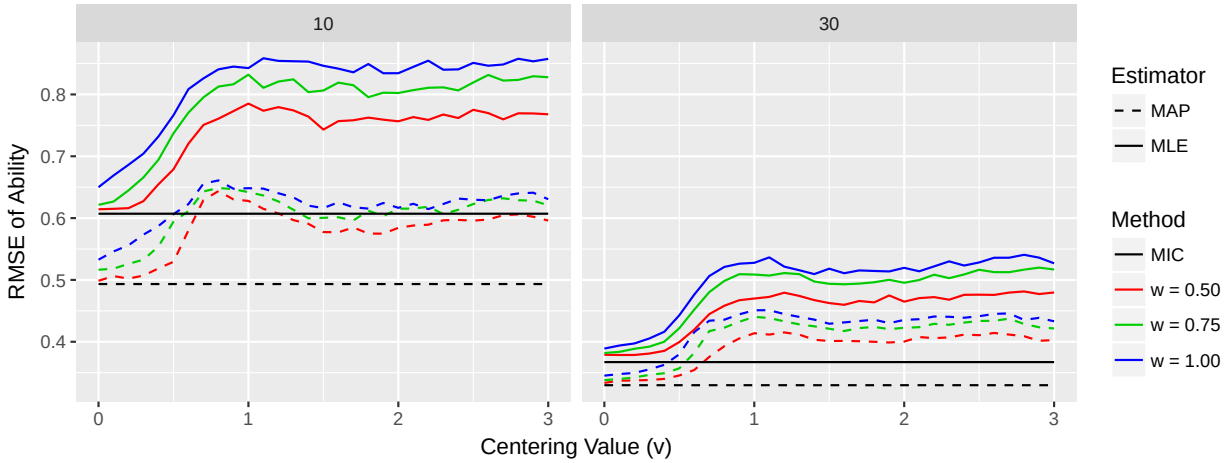


**Figure 3.8:** Plot of RMSE of ability estimates for Simulation 2.

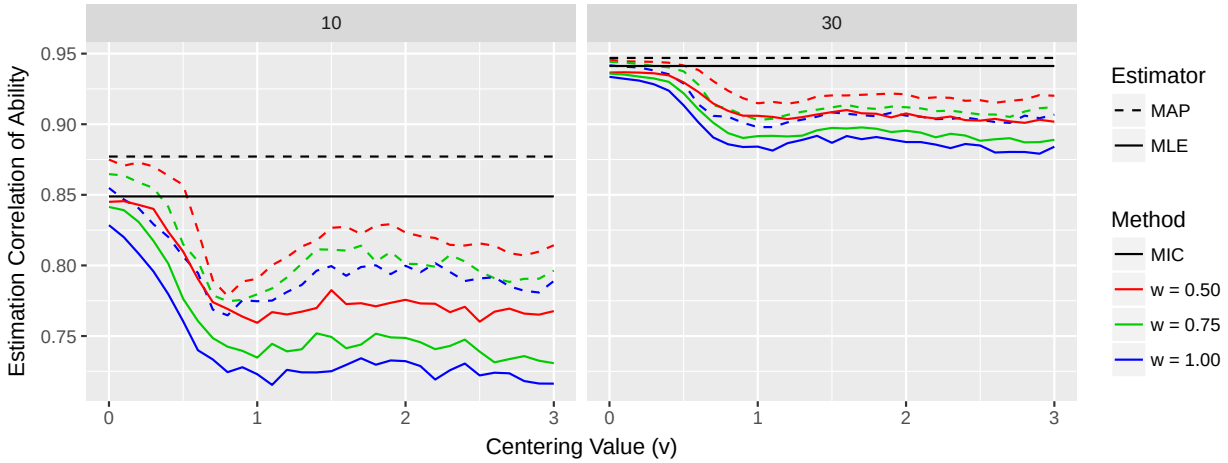


**Figure 3.9:** Plot of correlation of ability estimates with true values for Simulation 2.

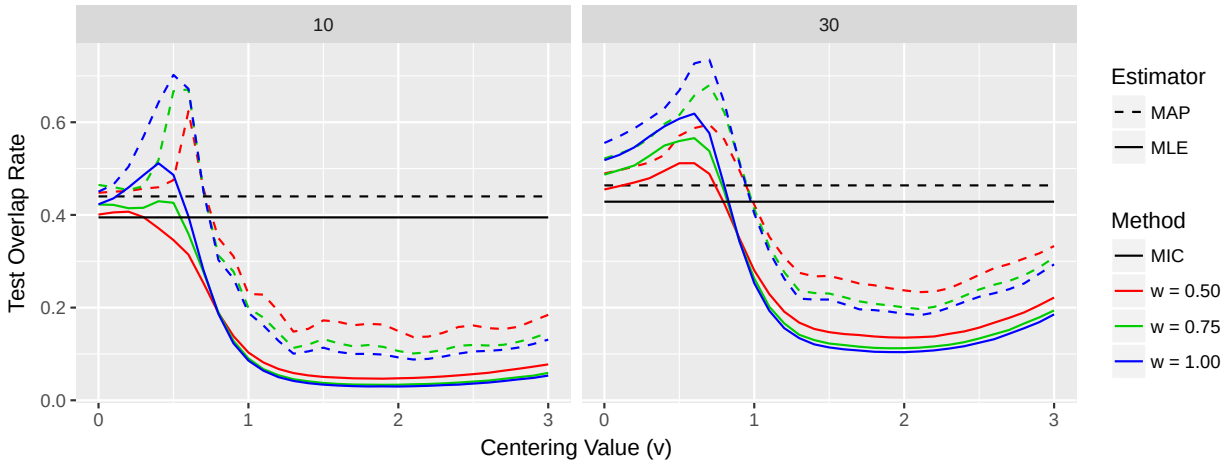


**Figure 3.10:** Plot of test overlap rate for Simulation 2.

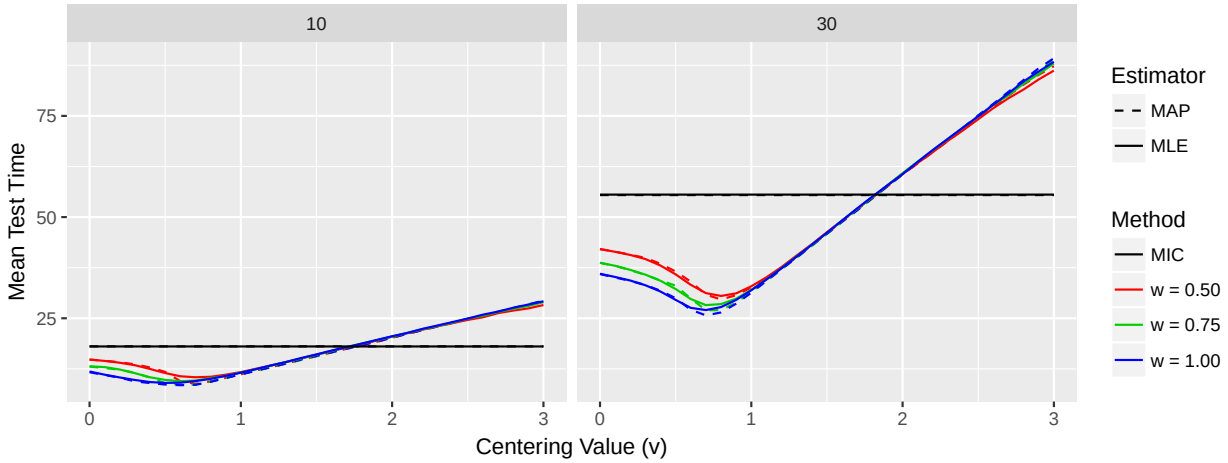


**Figure 3.11:** Plot of mean test time for Simulation 2.

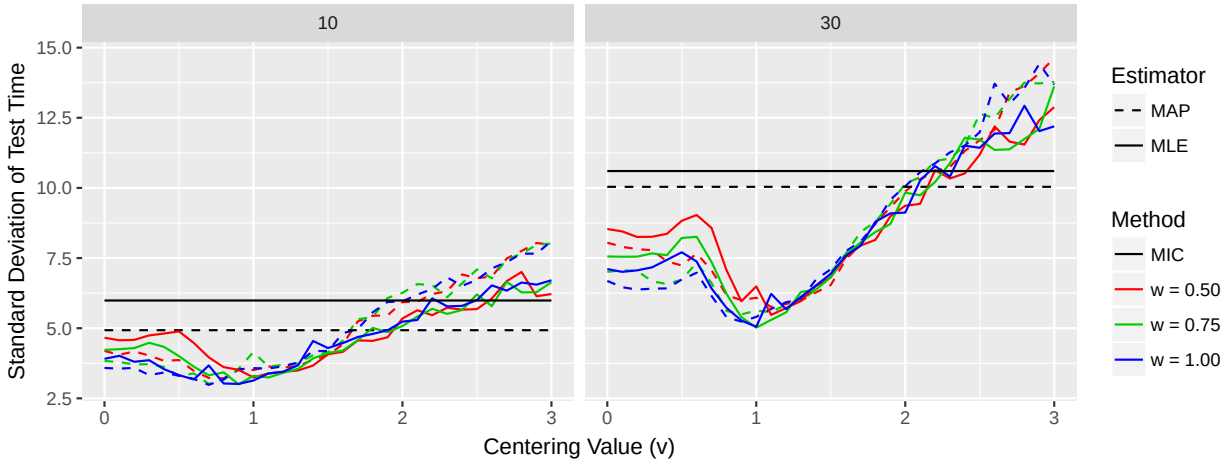


**Figure 3.12:** Plot of SD of test time for Simulation 2.

# Chapter 4

# Expected Posterior Variance

## 4.1 Introduction

As noted earlier, the new item selection technique suggested by the information matrix of the MAP estimator in (2.12) is indistinguishable from the standard maximum information method. One reason for this may that the item selection technique discussed is based on the asymptotic posterior variance of $\theta$. By using the non-asymptotic posterior variance, there may be a larger gain in efficiency (van der Linden, 1998) when using response times. Furthermore, it is important to note that the areas of largest gains when using the non-asymptotic posterior variance are the same areas where there is most uncertainty in choosing efficient items, that is, at the beginning of test. Therefore, it is very possible that the slight effect that using response times for selecting items has is actually a function of that inherent uncertainty. As such, item selection methods that take uncertainty of the estimate of $\theta$ into account at the beginning of the test, such as a Bayesian selection method (van der Linden, 1998) or the KL-information method (Chang & Ying, 1996) would be beneficial and calls for an investigation.

The item selection method can also be investigated for two other possibilities: minimizing overall test-taking time, and effects on item exposure. Previous studies have shown trade-offs between minimizing test time and maximizing spread of item exposure (Choe & Kern, 2014; Fan et al., 2012). Indeed, a more efficient estimator of both $\theta$ and $\tau$, such as the maximum a posteriori (MAP) estimator in Chapter 2, would help with item selection for these trade-offs. In this chapter, the Bayesian selection method will be investigated. The next chapter will similarly investigate the KL-information method.

## 4.2 Expected Posterior Variance Using Response Times

Using van der Linden's (2007) hierarchical framework to model the between-persons relationship of ability and speededness, a joint estimation technique using a MAP estimator of the parameters for ability $\theta$ and speededness $\tau$ was developed in the Chapter 2. It was showed that this estimator improves the estimation of ability when the covariance between $\theta$ and $\tau$ ($\sigma_{\theta\tau}$) is moderate to high, and its estimate $\hat{\sigma}_{\theta\tau}$ to used in estimation is fairly accurate. Furthermore, it was shown that selecting items using a minimum asymptotic posterior variance of $\theta$ (with response times accounted for) vs. using the standard maximum information method made little to no difference in the accuracy of $\theta$. However, according to Ranger (2013), response times do contain some information on ability. This suggests that the use of a large sample selection method may not have been appropriate.

Van der Linden (1998) proposed several Bayesian item selection criteria appropriate for use with Bayesian estimation methods, such as the MAP. Of the criteria he proposed, the best performing method was the minimum expected posterior variance (MEPV) where the selected item is one chosen according to (1.12). This method is especially effective for small sample (i.e., short) tests. A similar selection method is developed here using response times based on van der Linden's (2007) model. Items will be selected according to a generalization of the MEPV, called the MEPVT:

$$\text{EPVT}_i = \sum_{x=0}^{1} \int p_i(X_i = x, T_i = t|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)\text{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x, \mathbf{t}_{(-i)}, T_i = t, \tau)\ dt_i, \qquad (4.1)$$

where $\mathbf{x}_{(-i)} = (x_1, \ldots, x_{i-1})$, $\mathbf{t}_{(-i)} = (t_1, \ldots, t_{i-1})$, $p_i(X_i = x, T_i = t|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)$ is the posterior predictive distribution after responses and response times to the $i-1$ previous items, and $\text{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x, \mathbf{t}_{(-i)}, T_i = t, \tau)$ is the posterior variance of $\theta$. It is shown in Appendix A that

the posterior predictive distribution is

$$\begin{aligned} p_i(X_i = x, T_i = t|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) &= \int f(x|\theta) f(t|\tau) f(\theta|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)\ d\theta \\ &= \frac{\int f(x|\theta) f(t|\tau) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta}{\int f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta} \\ &= \frac{f(t|\tau) \int P_i(\theta)^x Q_i(\theta)^{1-x} \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}{\int \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}. \end{aligned} \tag{4.2}$$

Thus, the posterior predictive distribution is independent of the previous response times. This makes sense, as $\tau$ and $\mathbf{t}_{(-i)}$ contain the same information. Therefore, $p_i(X_i = x, T_i = t|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) = p_i(X_i = x, T_i = t|\mathbf{x}_{(-i)}, \tau)$.

In Appendix B, it is shown that the posterior variance is

$$\begin{aligned} \mathrm{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x, \mathbf{t}_{(-i)}, T_i = t, \tau) &= \mathrm{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x, \tau) \\ &= \mathrm{E}(\theta^2|\mathbf{x}_{(-i)}, X_i = x, \tau) - [\mathrm{E}(\theta|\mathbf{x}_{(-i)}, X_i = x, \tau)]^2, \end{aligned} \tag{4.3}$$

where

$$\mathrm{E}(\theta^2|\mathbf{x}_{(-i)}, X_i = x, \tau) = \frac{\int \theta^2 P_i(\theta)^x Q_i(\theta)^{1-x} \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}{\int P_i(\theta)^x Q_i(\theta)^{1-x} \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}$$

and

$$\mathrm{E}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \tau) = \frac{\int \theta P_i(\theta)^x Q_i(\theta)^{1-x} \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}{\int P_i(\theta)^x Q_i(\theta)^{1-x} \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}.$$

By plugging (4.2) and (4.3) into (4.1), we find that the expected posterior variance using responses/response times is

$$\begin{aligned} \mathrm{EPVT}_i = &\frac{\int f(t|\tau)\ dt}{\int f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta} \\ &\times \sum_{x=0}^{1} \left[\int \theta^2 f(x|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta - \frac{\left[\int \theta f(x|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta\right]^2}{\int f(x|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta}\right]. \end{aligned} \tag{4.4}$$

However, since $\int f(t|\tau)\ dt = 1$ and $\int f(\mathbf{x}_{(-i)}|\theta)f(\theta|\tau)\ d\theta$ is a constant that does not depend on the next item, the first term in (4.4) is constant across all choices of item $i$. Thus, the MEPVT would choose as the next item, the item that minimizes:

$$\mathrm{PV}_i = \sum_{x=0}^{1}\left[\int \theta^2 f(x|\theta)f(\mathbf{x}_{(-i)}|\theta)f(\theta|\tau)\ d\theta - \frac{\left[\int \theta f(x|\theta)f(\mathbf{x}_{(-i)}|\theta)f(\theta|\tau)\ d\theta\right]^2}{\int f(x|\theta)f(\mathbf{x}_{(-i)}|\theta)f(\theta|\tau)\ d\theta}\right]. \tag{4.5}$$

This method can be further modified for use in a GMICT-type selection method. To do this, the inverse of $\mathrm{PV}_i$ in (4.5) can be used in place of the item Fisher information in the GMICT. Equivalently, items could be chosen so that the expected posterior variance modified GMICT (PV-GMICT)

$$\mathrm{IT}_i^{PV} = \mathrm{PV}_i \times |\mathrm{E}(T_i|\hat{\tau}) - v|^w \tag{4.6}$$

is minimized.

As the expected posterior variance is a small sample Bayesian technique, it is expected that the performance of ability estimation should be increased over maximum information item selection, especially for shorter tests. In the first simulation, MEPVT will be compared against MIC with a simulation using the simulated item bank. In the second simulation, PV-GMICT will be compared to both the MIC and the GMICT, with a simulation using the real item bank. Estimation of traits will be done using both the MLE and the MAP.

## 4.3 Simulation 1: MIC vs. MEPVT

### 4.3.1 Method

A simulation study is carried out to compare the MIC with the MEPVT. To determine when the MEPVT improves ability estimation over a standard CAT, several factors are manipulated. First, the person parameters are simulated from the multivariate normal distribution, where $\mu_\theta = 0$, $\mu_\tau = 0$, $\sigma_\theta^2 = 1$, $\sigma_\tau^2 = 1$, and $\rho_{\theta\tau}$ is either 0, .25, .50, or .75. In the CAT, ability and speededness are estimated using either the joint MAP from Chapter 2 or the standard MLE. Tests have fixed lengths of either 5, 10, 20, or 30 items. For each factor combination, 2000 examinees are simulated. Ten replications of all factor combinations were done.

A 1000-item bank is simulated with the item parameters $a$, $b$, $c$, $\alpha$, and $\beta$ as defined earlier. Parameters are generated as follows:

- $(a^*, b, \beta) \sim N_3\left(\begin{pmatrix} 0.3 \\ 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.10 & 0.16 & 0.00 \\ 0.16 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{pmatrix}\right)$
  where $a^* = \log a$;
- $c \sim beta(2, 10)$;
- $\alpha \sim unif(1, 4)$.

The choices of distributions here were chosen to simulate items that are commonly found in standardized testing. Because it is known that there is often a relationship between item difficulty and discrimination parameters, the covariance matrix has a non-zero relationship between $a^*$ and $b$. Furthermore, in the same covariance matrix, the covariance between the item difficulty and time intensity parameters is chosen because it is believed there is a moderate association between these parameters. Ability estimation is assessed using bias, RMSE, and correlation:

- Bias:
$$Bias(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J},$$
- RMSE:
$$RMSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J},$$
- Correlation:
$$Cor(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{Cov(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\left[Var(\boldsymbol{\theta})^{1/2}\right]\left[Var(\hat{\boldsymbol{\theta}})^{1/2}\right]}.$$

To further assess the effectiveness of the test, mean test time (MTT), standard deviation of test time (STT), and the test overlap rate.

### 4.3.2 Results

The results of the simulation are given in Tables 4.1 and 4.2. As expected, as test length increases ability estimation accuracy increases: RMSE decreases, bias decreases, and correlation

| Test Length | Cor($\theta$, $\tau$) | Selection Method | RMSE($\hat{\theta}$) | Absolute Bias($\hat{\theta}$) | Cor($\theta$,$\hat{\theta}$) | Mean Test Time | SD Test Time | Overlap rate |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.00 | MIC | 0.54 | 0.001 | 0.86 | 9.10 | 13.49 | 0.33 |
| 5 | 0.00 | MEPVT | 0.52 | 0.013 | 0.87 | 8.85 | 12.58 | 0.42 |
| 5 | 0.25 | MIC | 0.54 | 0.001 | 0.86 | 8.86 | 12.77 | 0.34 |
| 5 | 0.25 | MEPVT | 0.52 | 0.010 | 0.87 | 8.82 | 12.08 | 0.42 |
| 5 | 0.50 | MIC | 0.53 | 0.000 | 0.86 | 8.68 | 12.12 | 0.33 |
| 5 | 0.50 | MEPVT | 0.51 | 0.006 | 0.87 | 8.62 | 11.46 | 0.37 |
| 5 | 0.75 | MIC | 0.54 | 0.006 | 0.86 | 8.62 | 11.65 | 0.33 |
| 5 | 0.75 | MEPVT | 0.51 | 0.003 | 0.88 | 8.37 | 10.65 | 0.29 |
| 10 | 0.00 | MIC | 0.39 | 0.009 | 0.93 | 19.23 | 26.48 | 0.32 |
| 10 | 0.00 | MEPVT | 0.39 | 0.003 | 0.93 | 18.46 | 24.40 | 0.38 |
| 10 | 0.25 | MIC | 0.40 | 0.009 | 0.93 | 18.45 | 24.19 | 0.32 |
| 10 | 0.25 | MEPVT | 0.38 | 0.002 | 0.93 | 17.99 | 23.14 | 0.38 |
| 10 | 0.50 | MIC | 0.39 | 0.009 | 0.93 | 17.85 | 21.28 | 0.32 |
| 10 | 0.50 | MEPVT | 0.38 | 0.003 | 0.93 | 17.15 | 19.79 | 0.36 |
| 10 | 0.75 | MIC | 0.39 | 0.010 | 0.93 | 17.52 | 21.46 | 0.32 |
| 10 | 0.75 | MEPVT | 0.38 | 0.003 | 0.93 | 16.48 | 18.47 | 0.31 |
| 20 | 0.00 | MIC | 0.28 | 0.003 | 0.96 | 39.14 | 52.74 | 0.33 |
| 20 | 0.00 | MEPVT | 0.28 | 0.001 | 0.96 | 37.86 | 51.95 | 0.37 |
| 20 | 0.25 | MIC | 0.28 | 0.002 | 0.96 | 36.73 | 46.93 | 0.33 |
| 20 | 0.25 | MEPVT | 0.29 | 0.002 | 0.96 | 36.81 | 47.30 | 0.37 |
| 20 | 0.50 | MIC | 0.28 | 0.003 | 0.96 | 35.48 | 43.65 | 0.33 |
| 20 | 0.50 | MEPVT | 0.28 | 0.002 | 0.96 | 34.62 | 42.20 | 0.37 |
| 20 | 0.75 | MIC | 0.28 | 0.001 | 0.96 | 34.00 | 39.24 | 0.33 |
| 20 | 0.75 | MEPVT | 0.28 | 0.001 | 0.96 | 33.62 | 39.03 | 0.34 |
| 30 | 0.00 | MIC | 0.24 | 0.002 | 0.97 | 58.40 | 84.93 | 0.34 |
| 30 | 0.00 | MEPVT | 0.24 | 0.001 | 0.97 | 57.96 | 81.90 | 0.37 |
| 30 | 0.25 | MIC | 0.24 | 0.001 | 0.97 | 56.21 | 72.14 | 0.34 |
| 30 | 0.25 | MEPVT | 0.24 | 0.000 | 0.97 | 53.45 | 67.14 | 0.37 |
| 30 | 0.50 | MIC | 0.24 | 0.002 | 0.97 | 53.82 | 65.37 | 0.34 |
| 30 | 0.50 | MEPVT | 0.24 | 0.001 | 0.97 | 52.10 | 60.75 | 0.37 |
| 30 | 0.75 | MIC | 0.24 | 0.002 | 0.97 | 51.81 | 58.60 | 0.34 |
| 30 | 0.75 | MEPVT | 0.24 | 0.002 | 0.97 | 49.82 | 56.90 | 0.36 |

**Table 4.1:** Results for Simulation 1 (MLE).

increases. Furthermore, MTT and STT both increase. Also, overall, test overlap rate increases. Similarly, the effect of $\rho_{\theta\tau}$ on ability estimation is as expected as well; as $\rho_{\theta\tau}$ increases, RMSE decreases, bias decreases, and correlation increases. Additionally, as $\rho_{\theta\tau}$ increases, MTT, STT, and test overlap all decrease. Finally, the effect of using the MAP estimator over the MIC estimator is consistent with earlier findings. First, with respect to ability estimation accuracy, RMSE decreases, bias increases, and correlation increases. Also, MTT and STT both decrease, whereas overlap rate increases.

Inspection of the results shows that there is very little difference between the selection methods. In general, ability estimation accuracy is very slightly better using MEPVT over MIC; RMSE for

| Test Length | Cor($\theta$, $\tau$) | Selection Method | RMSE($\hat{\theta}$) | Absolute Bias($\hat{\theta}$) | Cor($\theta$,$\hat{\theta}$) | Mean Test Time | SD Test Time | Overlap rate |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.00 | MIC | 0.50 | 0.047 | 0.86 | 8.74 | 11.89 | 0.35 |
| 5 | 0.00 | MEPVT | 0.49 | 0.031 | 0.87 | 8.65 | 11.53 | 0.42 |
| 5 | 0.25 | MIC | 0.50 | 0.044 | 0.87 | 8.95 | 13.22 | 0.35 |
| 5 | 0.25 | MEPVT | 0.49 | 0.029 | 0.87 | 8.87 | 12.19 | 0.42 |
| 5 | 0.50 | MIC | 0.47 | 0.035 | 0.88 | 8.74 | 11.65 | 0.31 |
| 5 | 0.50 | MEPVT | 0.47 | 0.028 | 0.89 | 8.67 | 11.61 | 0.38 |
| 5 | 0.75 | MIC | 0.42 | 0.030 | 0.91 | 8.56 | 10.84 | 0.27 |
| 5 | 0.75 | MEPVT | 0.42 | 0.018 | 0.91 | 8.40 | 19.29 | 0.30 |
| 10 | 0.00 | MIC | 0.37 | 0.036 | 0.93 | 19.21 | 26.03 | 0.34 |
| 10 | 0.00 | MEPVT | 0.36 | 0.029 | 0.93 | 18.87 | 26.07 | 0.38 |
| 10 | 0.25 | MIC | 0.36 | 0.035 | 0.93 | 18.51 | 24.80 | 0.34 |
| 10 | 0.25 | MEPVT | 0.36 | 0.024 | 0.93 | 18.07 | 23.38 | 0.38 |
| 10 | 0.50 | MIC | 0.35 | 0.028 | 0.94 | 17.72 | 22.22 | 0.32 |
| 10 | 0.50 | MEPVT | 0.35 | 0.025 | 0.94 | 17.28 | 20.76 | 0.36 |
| 10 | 0.75 | MIC | 0.32 | 0.020 | 0.95 | 16.93 | 18.46 | 0.29 |
| 10 | 0.75 | MEPVT | 0.33 | 0.020 | 0.95 | 16.64 | 19.10 | 0.32 |
| 20 | 0.00 | MIC | 0.27 | 0.021 | 0.96 | 38.82 | 53.50 | 0.35 |
| 20 | 0.00 | MEPVT | 0.27 | 0.018 | 0.96 | 37.84 | 50.89 | 0.37 |
| 20 | 0.25 | MIC | 0.27 | 0.017 | 0.96 | 36.79 | 45.90 | 0.35 |
| 20 | 0.25 | MEPVT | 0.27 | 0.014 | 0.96 | 36.18 | 47.65 | 0.37 |
| 20 | 0.50 | MIC | 0.26 | 0.019 | 0.96 | 35.11 | 42.58 | 0.34 |
| 20 | 0.50 | MEPVT | 0.26 | 0.012 | 0.97 | 34.25 | 39.55 | 0.36 |
| 20 | 0.75 | MIC | 0.25 | 0.016 | 0.97 | 33.52 | 37.92 | 0.32 |
| 20 | 0.75 | MEPVT | 0.25 | 0.016 | 0.97 | 33.14 | 36.61 | 0.35 |
| 30 | 0.00 | MIC | 0.23 | 0.014 | 0.97 | 58.71 | 84.45 | 0.36 |
| 30 | 0.00 | MEPVT | 0.23 | 0.015 | 0.97 | 57.10 | 75.04 | 0.37 |
| 30 | 0.25 | MIC | 0.23 | 0.014 | 0.97 | 53.99 | 67.16 | 0.36 |
| 30 | 0.25 | MEPVT | 0.23 | 0.015 | 0.97 | 54.06 | 69.63 | 0.37 |
| 30 | 0.50 | MIC | 0.23 | 0.012 | 0.97 | 52.34 | 60.52 | 0.36 |
| 30 | 0.50 | MEPVT | 0.23 | 0.012 | 0.97 | 52.12 | 59.42 | 0.37 |
| 30 | 0.75 | MIC | 0.22 | 0.010 | 0.98 | 50.79 | 58.57 | 0.34 |
| 30 | 0.75 | MEPVT | 0.22 | 0.012 | 0.98 | 50.08 | 56.74 | 0.36 |

**Table 4.2:** Results for Simulation 1 (MAP).

ability is lower and Cor($\theta, \hat{\theta}$) are higher. Bias is a little more complicated; for the shortest test length (test length 5) when using MLE, bias for MEPVT is larger than MIC, whereas it is smaller for all other conditions. These effects on estimation accuracy become increasingly small as test length increases. Finally, MTT and STT are smaller for MEPVT, whereas the test overlap rate is larger. For the test overlap rate, interestingly, as test length increases, the effect of selection method choice lessens.

## 4.4 Simulation 2: GMICT vs. PV-GMICT

### 4.4.1 Method

Simulation studies with a real item bank are carried out to determine how the MIC, GMICT, MEPVT, and PV-GMICT compare with each other in item selection. Several factors are manipulated. First, items are selected using either the MIC, GMICT, MEPVT, or PV-GMICT. Second, for the GMICT and PV-GMICT methods, items are selected using several combinations of $v$ (0 to 3 in 0.1 increments) and $w$ (0.50, 0.75, or 1.00). Third, the tests have fixed lengths of either 10 or 30.

The item bank comes from a data set of a real high-stakes, large-scale standardized CAT. The data consists of raw responses and RTs from about 2000 examinees with an item pool containing about 500 multiple-choice items that were pre-calibrated according to 3PLM. The lognormal model item parameters $\alpha$ and $\beta$ were estimated using a modified version of van der Linden's (2007) MCMC routine that fixed the 3PLM item parameters to the pre-calibrated values, and the distribution of $\tau$ was set to have a mean of 0. A check of trace plots showed that all parameters appeared to converge using 10000 MCMC draws with a burn-in size of 5000.

To better reflect the population of test-takers from the real data set, examinee ability and speededness parameters are simulated from a multivariate normal distribution with population parameters of $\mu_\theta$, $\mu_\tau$, $\sigma_\theta$, $\sigma_\tau$, and $\rho_{\theta\tau}$. They are as follows: $\mu_\theta = 0.5$, $\mu_\tau = 0$, $\sigma_\theta = 1$, $\sigma_\tau = 0.16$, and $\rho_{\theta\tau} = 0.76$. Examinees were simulated in ten replications of 1000 examinees each.

Bias, RMSE, and correlation are used to assess ability and speededness estimation. Furthermore, test performance is assessed using mean overall test time across individuals (MTT), the standard deviation of test times across individuals (STT), and the test overlap rate.

### 4.4.2 Results

As usual, increases in test length are consistent with increases in ability estimation accuracy, as shown in Figures 4.1, 4.2, and 4.3 for RMSE, bias, and correlation, respectively. Specifically, RMSE decreases, bias decreases, and correlation increases. Furthermore, the test overlap rate, MTT, and STT all increase as well. These are shown in Figures 3.10, 3.11, and 3.12 for test overlap, MTT,

and STT, respectively.

Investigation of the results with respect to selection method gives some surprising results. In particular, PV-GMICT performs significantly worse than GMICT on estimation accuracy (RMSE is larger and correlation is lower), and performs no better than GMICT on lowering mean test times and variability in test times. The only areas that PV-GMICT improves upon GMICT is with marginally smaller test overlap rates and smaller bias. Unfortunately, the improvement in bias disappears with a longer test. In comparison with MIC and MEPVT, for lower values of $v$, bias, MTT, and STT are lower for GMICT and PV-GMICT. Conversely, for high enough values of $v$, test overlap rate is lower for GMICT and PV-GMICT. This means that reasonably chosen values of $v$ are necessary for using these methods to their best ability. One thing to note is that $w$ has very little effect on any of the outcome variables.

## 4.5 Discussion

This chapter proposed a new method for item selection using a generalization of the minimum expected posterior variance selection method (van der Linden, 2007) using the hierarchical model for response times and accuracies (van der Linden, 1998); this method is called MEPVT. It also proposed a second method—called PV-GMICT—that weights MEPVT by response times in the same way as the GMICT. These methods were introduced as methods for incorporating response time information into item selection using a non-asymptotic variance, with the goal of capitalizing on that information in short tests to improve estimation of ability.

In a similar argument given in the previous chapter, a smaller RMSE can be a result of three different possibilities: smaller bias, smaller variance, or smaller bias and smaller variance. It is clear that in many conditions, bias has decreased, so the question becomes "did variance decrease as well?" Simple calculation shows that in most cases, yes, it did. Of course, as test length increases, the differences between the selection methods vanish.

While MEPVT is minimally successful at this goal, it comes at the expense of a higher test overlap rate. Whether a higher test overlap is an expense that the test developer is willing to spend will depend on the overall goals of the test developer. One possibility for longer tests is to use MEPVT at the beginning of a test (maybe for just the first 5 or 10 items), when there is

less certainty in the estimate, and switch to MIC later on. That said, unless the test is only 5 to 10 items in length, the gain may be so small that by the end of the test, the effect on estimation accuracy of using MEPVT would probably be a wash. While it does decrease time spent on a test, that decrease is not much considering that other methods do a much better job at this.

When comparing PV-GMICT with GMICT, the decrease in estimation accuracy is surprising; PV-GMICT does much worse in terms of RMSE and correlation, though it does have a smaller bias. Unfortunately, the scale of bias is small, so the gain due to bias is incredibly slight. While PV-GMICT does make the test overlap rate decrease marginally, it does no better than GMICT on mean and standard deviation of test times. By all appearances, it seems that PV-GMICT is not particularly well-suited for its job, especially when GMICT already does it so well, but one possibility that was not considered before is that the $w$ parameter may simply be too high for PV-GMICT, giving too much weight to controlling time. This seems to be corroborated by the fact that the three values of $w$ are barely distinguished between each other, meaning that the chosen values of $w$ may be toward the extreme side of the balance between choosing items according to posterior variance and the deviation of expected response times from $v$. This needs to be investigated further, but at this time it seems that if suitable values of $w$ are able to be determined, that gains compared to GMICT may be minimal, just as they are for MEPVT over MIC.

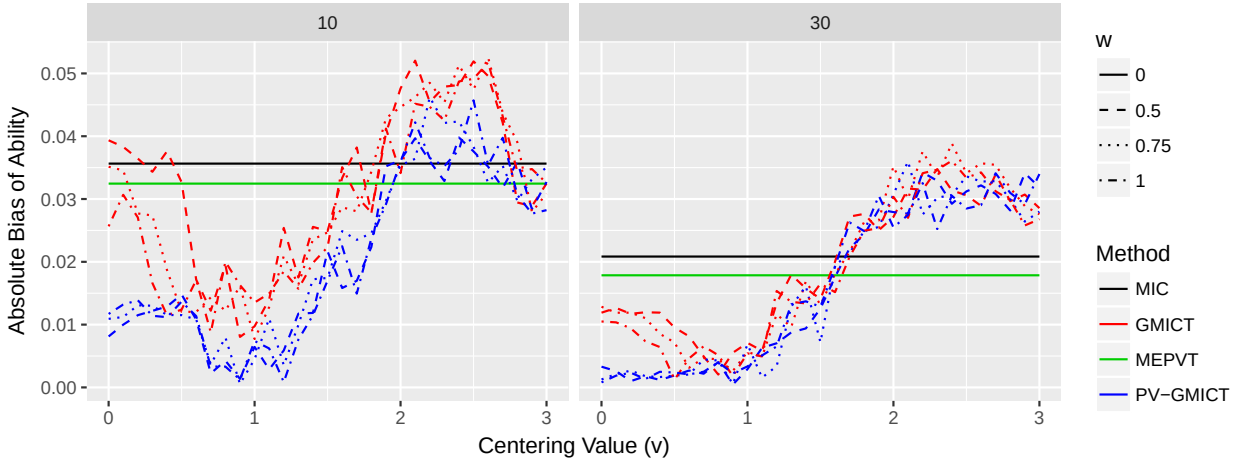


**Figure 4.1:** Plot of bias of ability estimates for Simulation 2.

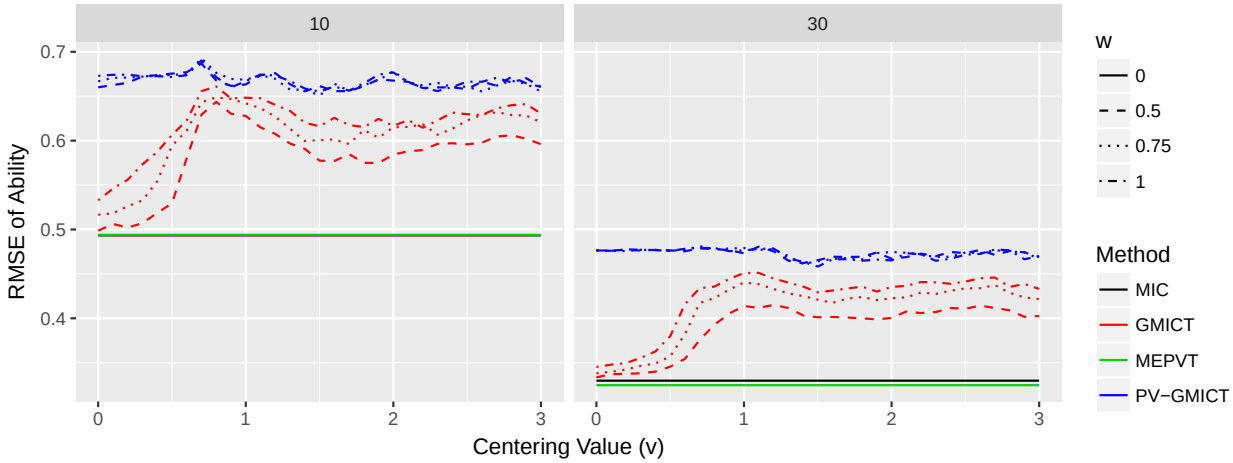


**Figure 4.2:** Plot of RMSE of ability estimates for Simulation 2.

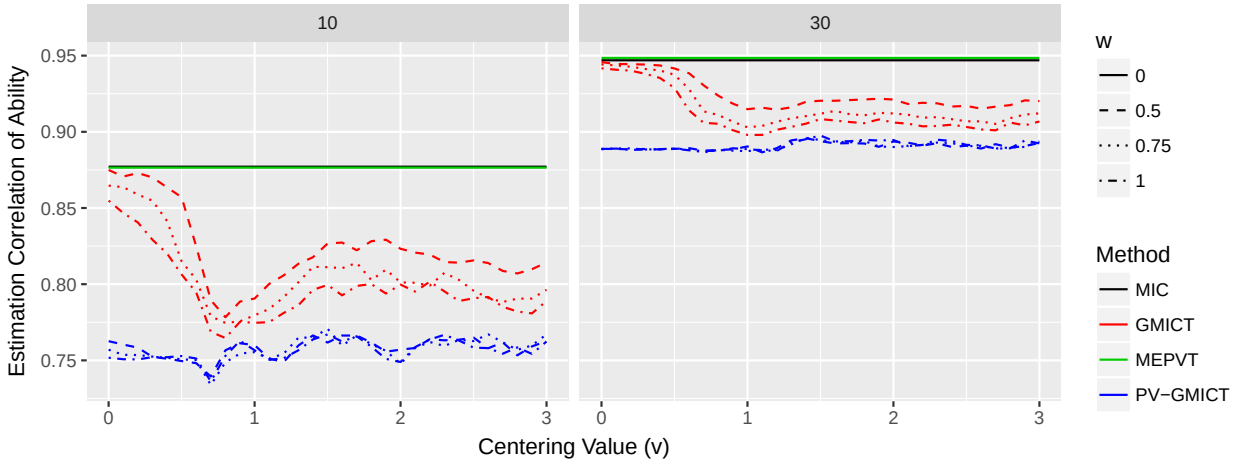


**Figure 4.3:** Plot of correlation of ability estimates with true values for Simulation 2.

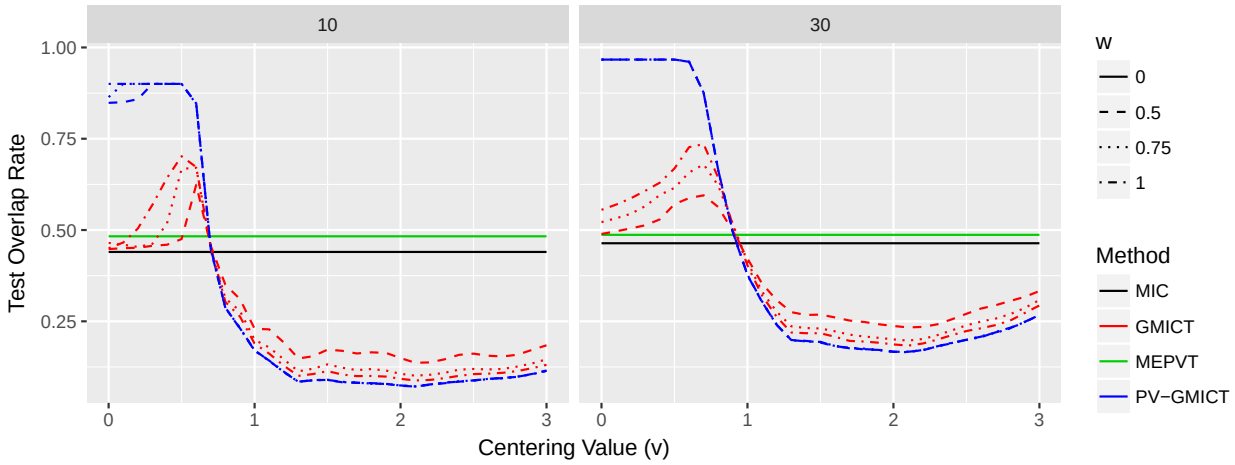


**Figure 4.4:** Plot of test overlap rate for Simulation 2.

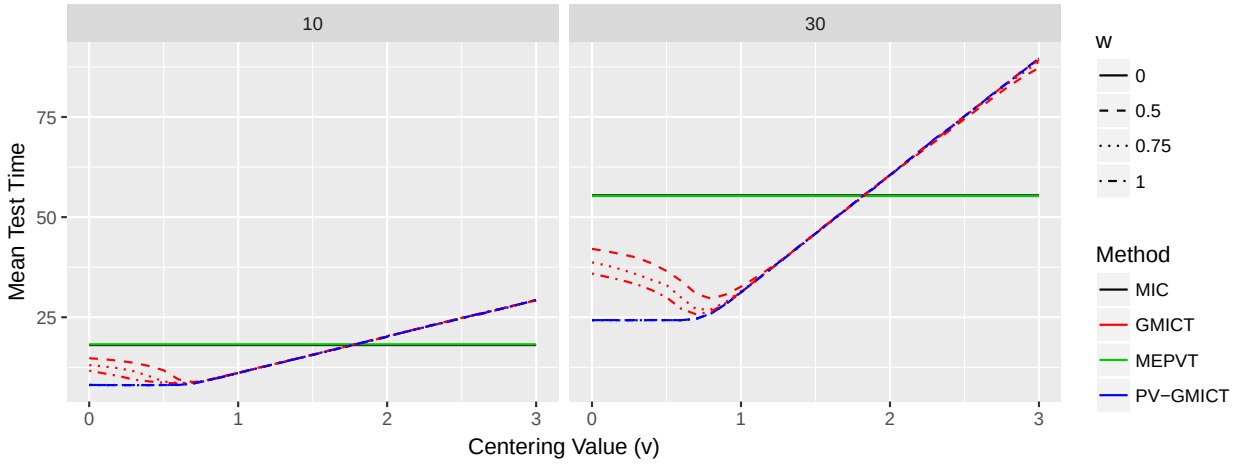


**Figure 4.5:** Plot of mean test time for Simulation 2.

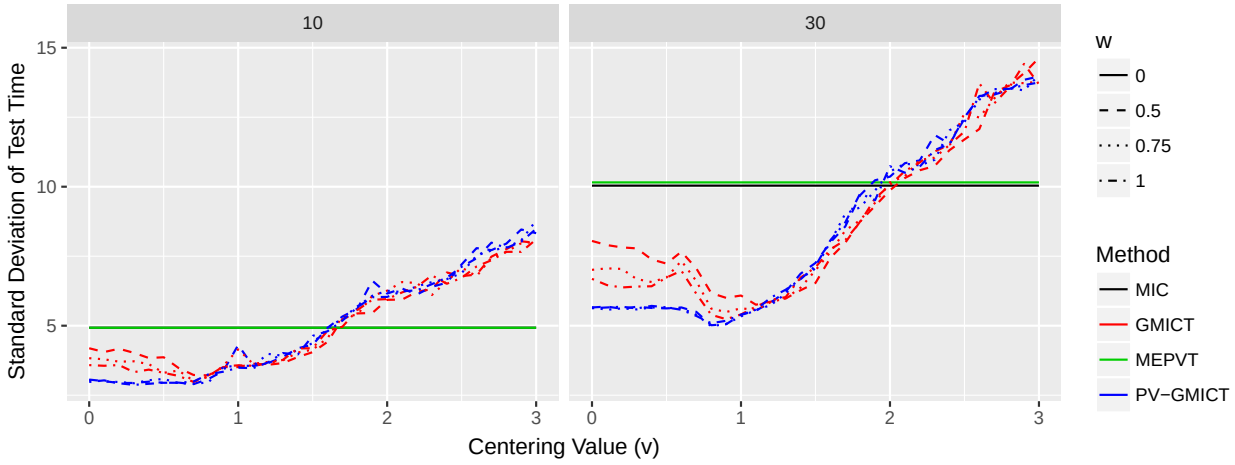


**Figure 4.6:** Plot of SD of test time for Simulation 2.

# Chapter 5

# Response-time Weighted Kullback-Leibler Information

## 5.1 Introduction

The most common technique for selecting items in an adaptive testing system is using the maximum information criterion (MIC). Essentially, in an adaptive test, the MIC chooses the next item to administer such that the item Fisher information

$$I_i(\theta_j) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_j^2}\ln(P_i(\theta_j))|\theta_j\right]$$

is maximized over all items in the item bank. It has been argued that the value of the item information function at $\theta$ is *local information*, and that the MIC as an item selection criterion works best when the estimate of ability $\hat{\theta}$ is close to the real ability level $\theta$ (Chang & Ying, 1996). When not close, the chosen item is less optimal than desired; this is especially true early on in an adaptive test.

An alternative is to use a *global information* technique, such as the Kullback-Leibler (K-L) information criterion proposed by Chang and Ying (1996). They argued that the log-likelihood ratio can best quantify global information, and, thus, proposed the use of the K-L information (or K-L distance) function, as it is a distance function—or, more correctly, a deviance function—of the log-likelihood. K-L information is defined as

$$K_i(\theta\|\theta_0) = \mathrm{E}\left[\log\left(\frac{L_i(\theta_0|X_i)}{L_i(\theta|X_i)}\right)\right] = P_i(\theta_0)\log\left(\frac{P_i(\theta_0)}{P_i(\theta)}\right) + Q_i(\theta_0)\log\left(\frac{Q_i(\theta_0)}{Q_i(\theta)}\right),$$

where $\theta_0$ is the true ability, $\theta$ is an estimate of $\theta_0$, and $L_i$ is the likelihood of the $i$th item.

Since we do not actually know the person's true ability, an index is necessary. One simple index is the area under the curve $K$ around an appropriate interval about our estimate of ability $\hat{\theta}_n$. This

is given as

$$K_i(\hat{\theta}_n) = \int_{\hat{\theta}-\delta_n}^{\hat{\theta}_n+\delta_n} K_i(\theta \| \hat{\theta}_n)\ d\theta, \tag{5.1}$$

where $\delta_n$ is a control to determine the interval size around our current estimate of ability $\hat{\theta}_n$. A natural choice for $\delta_n$ is

$$\delta_n = \frac{1}{\sqrt{n}}$$

so that as the test length increases, the interval around $\hat{\theta}_n$ will decrease accordingly, signifying less uncertainty around our ability estimate. With the K-L index as an item selection criterion, items are chosen such that the K-L index is maximized over all items in the item bank.

In keeping with the theme of this dissertation, it would be interesting to see how a modified version of the K-L index using response times performs in item selection. Here, it is proposed to replace the item Fisher information $I_i(\theta)$ in the numerator of (1.35) with the K-L index $K_i(\theta)$ in (5.1). Specifically, according to the K-L modified GMICT (KL-GMICT), the item with the largest value of

$$\mathrm{IT}_i^{KL} = \frac{K_i(\hat{\theta})}{|\mathrm{E}(T_i|\hat{\tau}) - v|^w} \tag{5.2}$$

will be chosen. This method has the advantage from using the K-L index is two-fold. First, a wider array of items will be considered at any given stage of the test, which could potentially even out the item exposure rate of the items in the item pool. As the GMICT already does this, the hope is that using the KL-GMICT the item exposure rate will even out above and beyond the GMICT. Second, the K-L index takes into account the inherent uncertainty of $\theta$-estimates, particularly at the beginning of the test, so more overall informative items should be chosen earlier. Thus, especially for shorter tests, we should get better estimates of examinee ability $\theta$. To understand the differences between these methods, MIC, GMICT, and KL-GMICT will be compared via two simulation studies, one using a simulated item bank and another using a real item bank.

## 5.2 Simulation 1: Simulated Item Bank and Examinee Populations

### 5.2.1 Method

Simulation studies are carried out to compare the MIC, GMICT, and KL-GMICT. Several factors are manipulated to determine the differences between the methods. First, the person parameters are simulated as

$$\begin{pmatrix} \theta \\ \tau \end{pmatrix} \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\theta\tau} & 1 \end{pmatrix} \right),$$

where $\rho_{\theta\tau}$ is either 0, .25, .50, or .75. In the CAT, ability and speededness are estimated using the joint MAP. Tests have fixed lengths of either 10, 20, or 30 items. Finally, items are selected using several combinations of $v$ and $w$ in GMICT selection procedure; here $v$ ranges from 0 to 3 in 0.1 increments and $w$ is either 0.50, 0.75, or 1.00. For each factor combination, ten replications of 2000 examinees are simulated.

A 500-item bank is simulated with the item parameters $a$, $b$, $c$, $\alpha$, and $\beta$ as defined earlier. Parameters are generated as follows:

- $(a^*, b, \beta) \sim N_3 \left( \begin{pmatrix} 0.3 \\ 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.10 & 0.16 & 0.00 \\ 0.16 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{pmatrix} \right)$
  where $a^* = \log a$;
- $c \sim beta(2, 10)$;
- $\alpha \sim unif(1, 4)$.

The choices of distributions here were chosen to simulate items that are commonly found in standardized testing. Because it is known that there is often a relationship between item difficulty and discrimination parameters, the covariance matrix has a non-zero relationship between $a^*$ and $b$. Furthermore, the covariance between the item difficulty and time intensity parameters is chosen because it is believed there is a moderate association between these parameters; often harder items

take longer to complete. Ability estimation is assessed using bias, RMSE, and correlation. To further assess the effectiveness of the test, mean test time (MTT), standard deviation of test time (STT), and the test overlap rate.

### 5.2.2 Results

The results of the simulation on estimation bias, RMSE, and correlation of ability are given in Figures 5.1, 5.2, and 5.3, respectively. Several conclusions can be made. First, it is clear that when choosing items using KL-GMICT with all combinations of $w$ and $v$, absolute bias increases, RMSE increases, and the correlation between estimated $\theta$ and true $\theta$ decreases compared to choosing items using MIC; that is, using KL-GMICT makes the accuracy of estimating ability worse. On the other hand, the test overlap, MTT, and STT all decrease for KL-GMICT compared to MIC for most values of $v$. Simulation results on test overlap rate, MTT, and STT are given in Figures 5.4, 5.5, and 5.6, respectively. The magnitude of the trend of using KL-GMICT over MIC is moderated by the choice of values of $w$. These broad observations of the effect that choosing items using KL-GMICT has in comparison to MIC are in the same as those using GMICT in comparison to MIC, but to a different extent. Surprisingly, KL-GMICT has a higher bias, higher RMSE, and lower correlation than GMICT. Conversely, the test overlap rate, MTT, and STT are all lower for KL-GMICT than GMICT.

For varying levels of $\rho_{\theta\tau}$ and varying test lengths, the direction of the effect of selection method is constant, though the magnitudes of the effects are moderated. For instance, as test length and $\rho_{\theta\tau}$ increase, the differences in ability estimation between the selection methods lessen. Also, as test length increases, MTT and STT increase, as expected, while an increase in $\rho_{\theta\tau}$ seems to very slightly decrease MTT and STT. Test overlap rate, on the other hand, is not affected by the strength of the relationship between ability and speededness, while it only very slightly increases test overlap. None of these effects are unexpected.

## 5.3 Simulation 2: Real Item Bank and Examinee Population

### 5.3.1 Method

Simulation studies with a real item bank are carried out to determine how the MIC, GMICT, and KL-GMICT compare with each other in item selection in an operational testing situation; several factors are manipulated. First, items are selected using either the MIC, GMICT, KL-GMICT. Second, for the GMICT and KL-GMICT methods, items are selected using several combinations of $v$ (0 to 3 in 0.1 increments) and $w$ (0.50, 0.75, or 1.00). Third, the tests have fixed lengths of either 10 or 30.

The item bank comes from a data set of a real high-stakes, large-scale standardized CAT. The data consists of raw responses and RTs from about 2000 examinees with an item pool containing about 500 multiple-choice items that were pre-calibrated according to 3PLM. The lognormal model item parameters $\alpha$ and $\beta$ were estimated using a modified version of van der Linden's (2007) MCMC routine that fixed the 3PLM item parameters to the pre-calibrated values, and the distribution of $\tau$ was set to have a mean of 0. A check of trace plots showed that all parameters appeared to converge using 10000 MCMC draws with a burn-in size of 5000.

Examinee ability and speededness parameters are simulated from a multivariate normal distribution with population parameters reflecting the population of test-takers from the real data set. The population parameters are as follows: $\mu_\theta = 0.5$, $\mu_\tau = 0$, $\sigma_\theta = 1$, $\sigma_\tau = 0.16$, and $\rho_{\theta\tau} = 0.76$. This simulation had ten replications each with 1000 examinees for every factor combination.

Bias, RMSE, and correlation are used to assess ability estimation. Furthermore, test performance is assessed using mean overall test time across individuals (MTT), the standard deviation of test times across individuals (STT), and the test overlap rate.

### 5.3.2 Results

The results in this simulation show a few differences from the Simulation 1. The first difference shows in the bias. In Simulation 1 KL-GMICT has a larger bias than GMICT, but in the current simulation, the bias looks to be about the same for the two methods. The other differences manifests in the mean test times and variability in test times; formerly the results showed that MTT and STT

were smaller for KL-GMICT, but the current simulation shows that MTT and STT are similar. All other results conform to the previous results. The results of the simulation on estimation bias, RMSE, and correlation of ability are given in Figures 5.7, 5.8, and 5.9, respectively. Simulation results for test overlap rate, MTT and STT are shown in Figures 5.10, 5.11, and 5.12.

## 5.4 Discussion

In this chapter, a new method combining the Kullback-Leibler index (Chang & Ying, 1996) with the GMICT was proposed, called the KL-GMICT. In much the same way as the GMICT, it seeks to balance out the competing goals of choosing items that best measure a given examinee's ability level, while minimizing the amount of time that said person needs to spend on that item. The competing interest of trying to maintain a balanced usage across the items of item bank is also taken into account. Taken together, KL-GMICT manages to succeed in balancing these interests against each other.

In a rather unfortunate manner, the KL-GMICT does not seem to balance the interests any better than GMICT does. In the current simulations, it seems that KL-GMICT has worse ability estimation accuracy than GMICT, while it better minimizes the mean and standard deviation of test times. Of course, just as in GMICT, changing the value of $w$ in KL-GMICT affects that balance. In particular, as $w$ increases, MTT and STT decrease, while bias and RMSE increase; this trend, of course, reverses with decreasing $w$. Thus, it seems that for any value of $w$ used in KL-GMICT that a corresponding value of $w$ used in GMICT should be able to be found that makes the outcomes of the methods close. Furthermore, KL-GMICT requires the computation of an intractable integral, which slows down the item selection process considerably. Taken together, this suggests that KL-GMICT as an item selection technique can be done just as easily with GMICT on its own.

On the other hand, this does not take into account the structure of the item bank itself. In Simulation 1, item parameters are simulated such that the population correlation matrix of the

item parameters is

$$\boldsymbol{\Sigma}_{\boldsymbol{\Psi}}^{(1)} = \begin{pmatrix} 1.00 & 0.51 & 0.00 & 0.00 & 0.00 \\ 0.51 & 1.00 & 0.00 & 0.00 & 0.50 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.50 & 0.00 & 0.00 & 1.00 \end{pmatrix}$$

,

where $\boldsymbol{\Psi} = (a^*, b, c, \alpha^*, \beta)$, $a^* \equiv \ln a$, and $\alpha^* \equiv \ln \alpha$, whereas in Simulation 2, the sample correlation matrix of the parameters in the real item bank is

$$\boldsymbol{\Sigma}_{\boldsymbol{\Psi}}^{(2)} = \begin{pmatrix} 1.00 & 0.32 & -0.03 & 0.04 & 0.09 \\ 0.32 & 1.00 & -0.08 & 0.20 & 0.45 \\ -0.03 & -0.08 & 1.00 & -0.03 & -0.06 \\ 0.04 & 0.20 & -0.03 & 1.00 & 0.53 \\ 0.09 & 0.45 & -0.06 & 0.53 & 1.00 \end{pmatrix}$$

.

This difference between the simulated bank and the real item bank suggests two things. First, a small relationship between $b$ and $\alpha^*$ may exist. Second, and perhaps more importantly, there is a rather large relationship between $\alpha^*$ and $\beta$. The strengths of these relationships may have an effect on the outcomes of the selection methods.

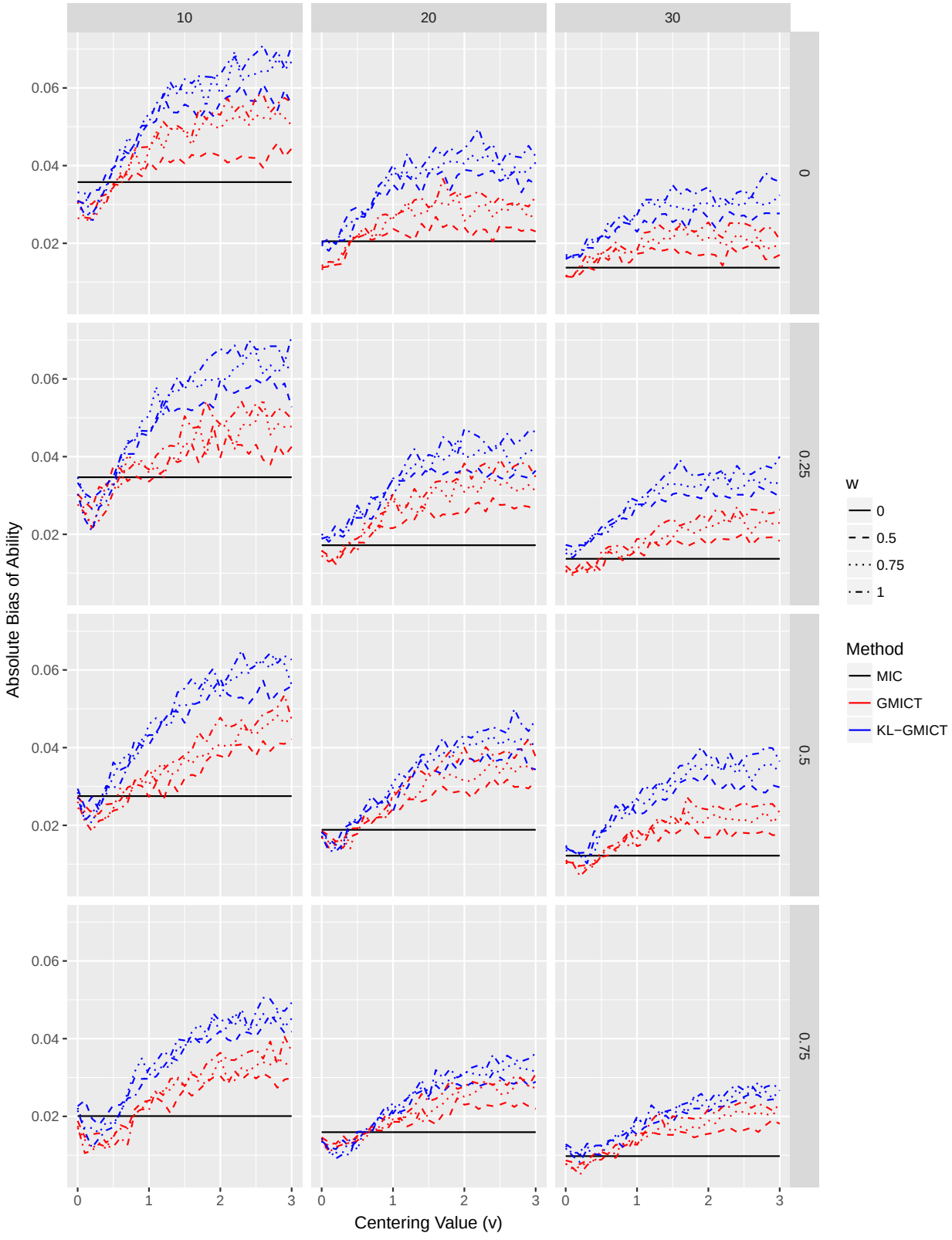


**Figure 5.1:** Plot of bias of ability estimates for Simulation 1.

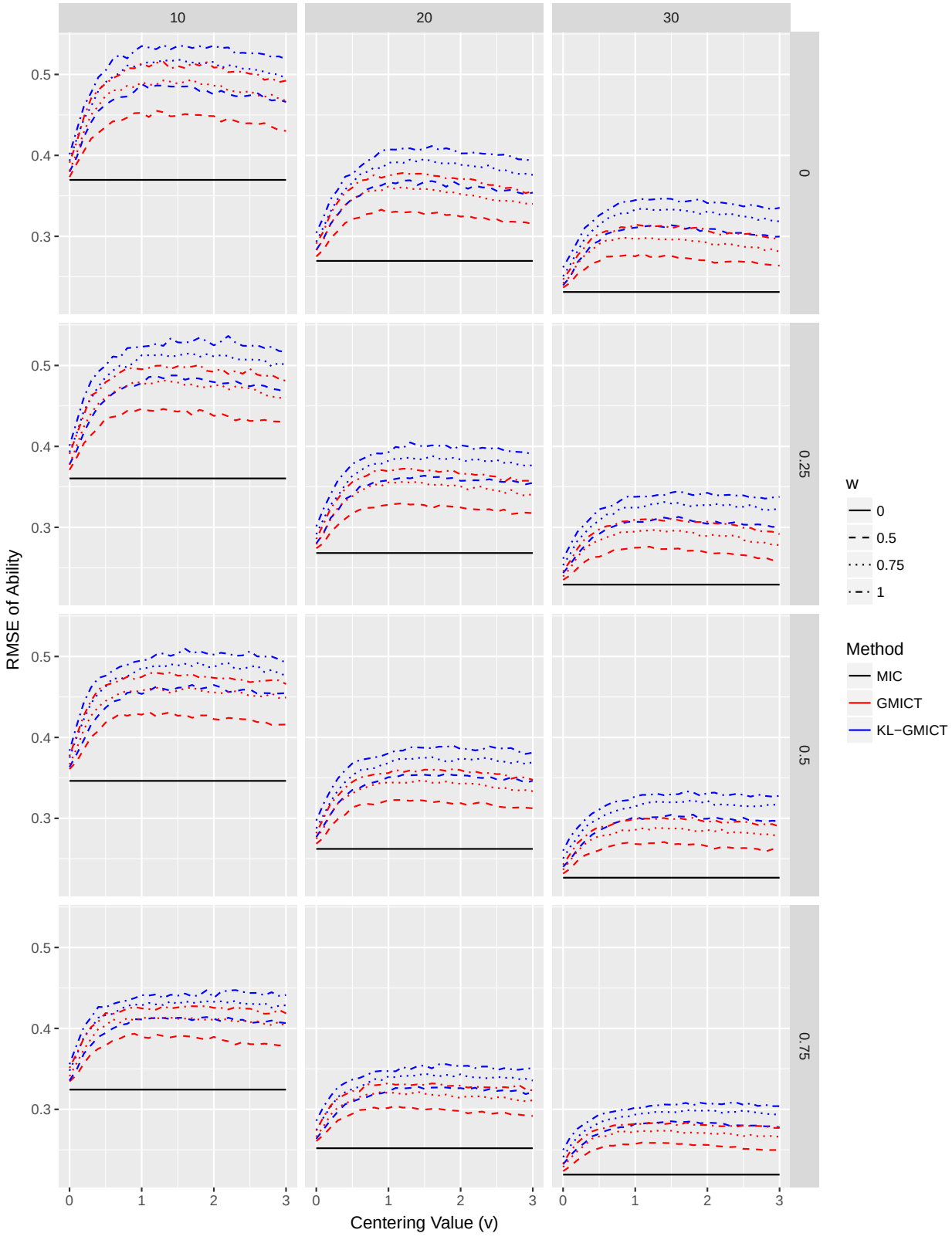


**Figure 5.2:** Plot of RMSE of ability estimates for Simulation 1.

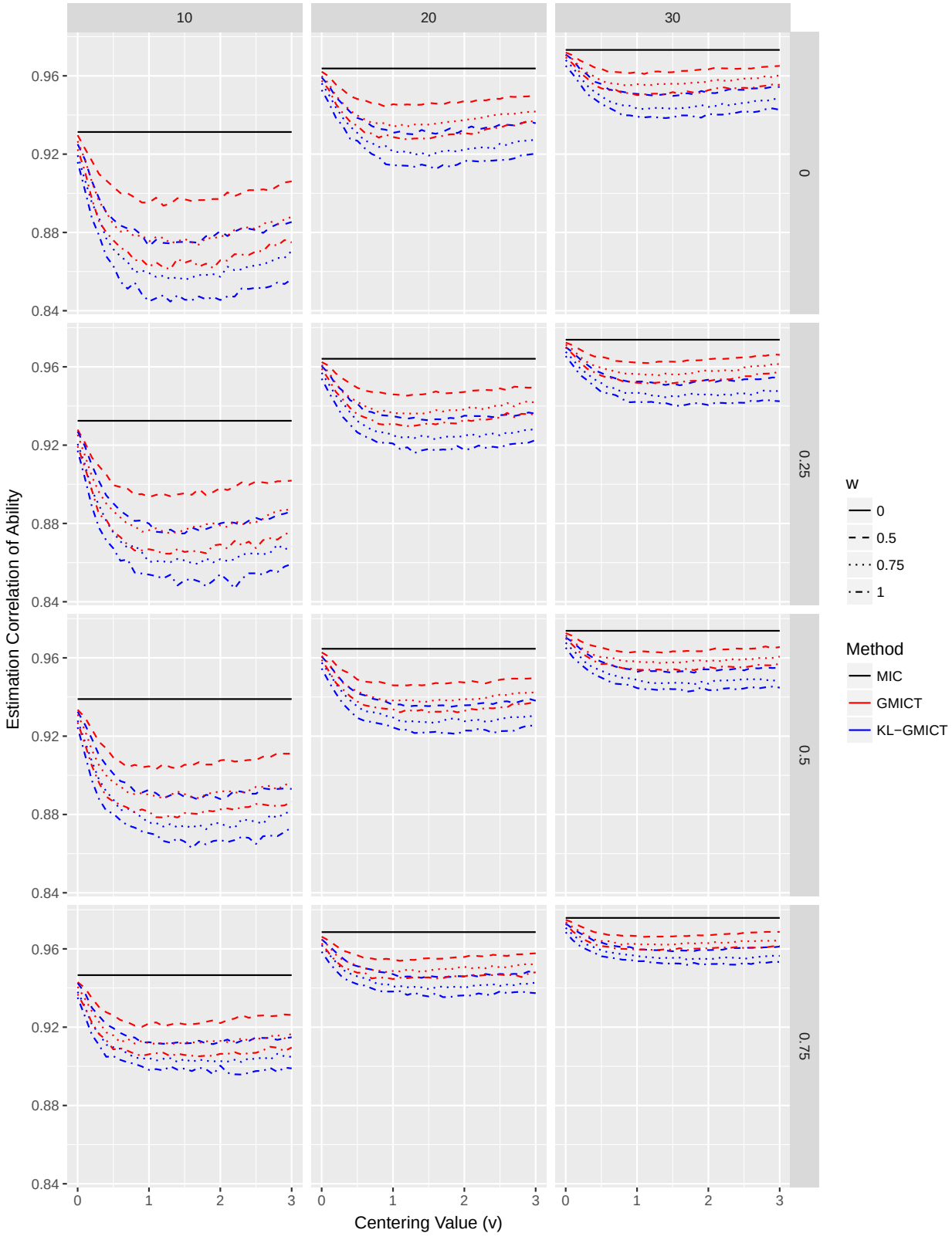


**Figure 5.3:** Plot of correlation of ability estimates with true values for Simulation 1.

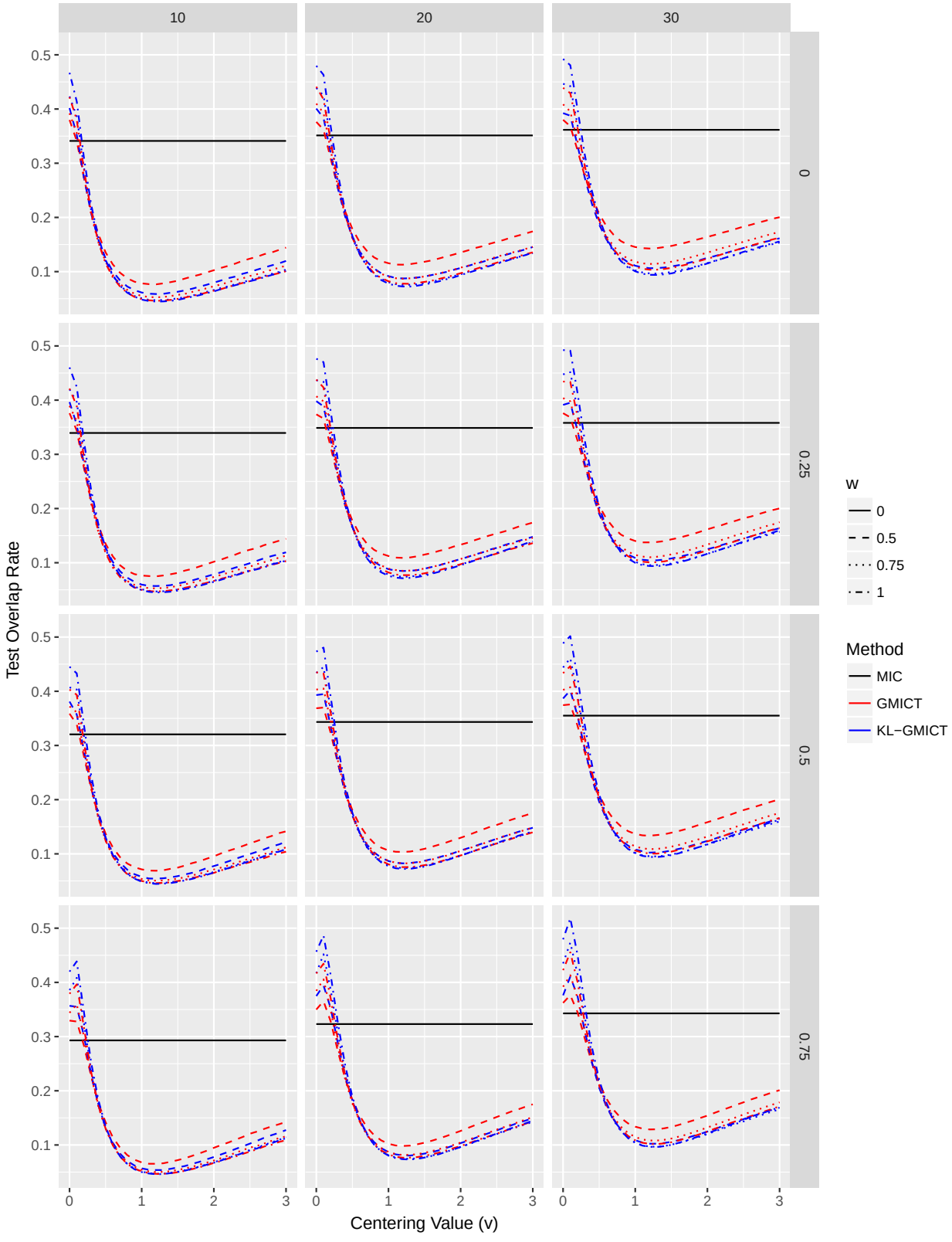


**Figure 5.4:** Plot of test overlap rate for Simulation 1.

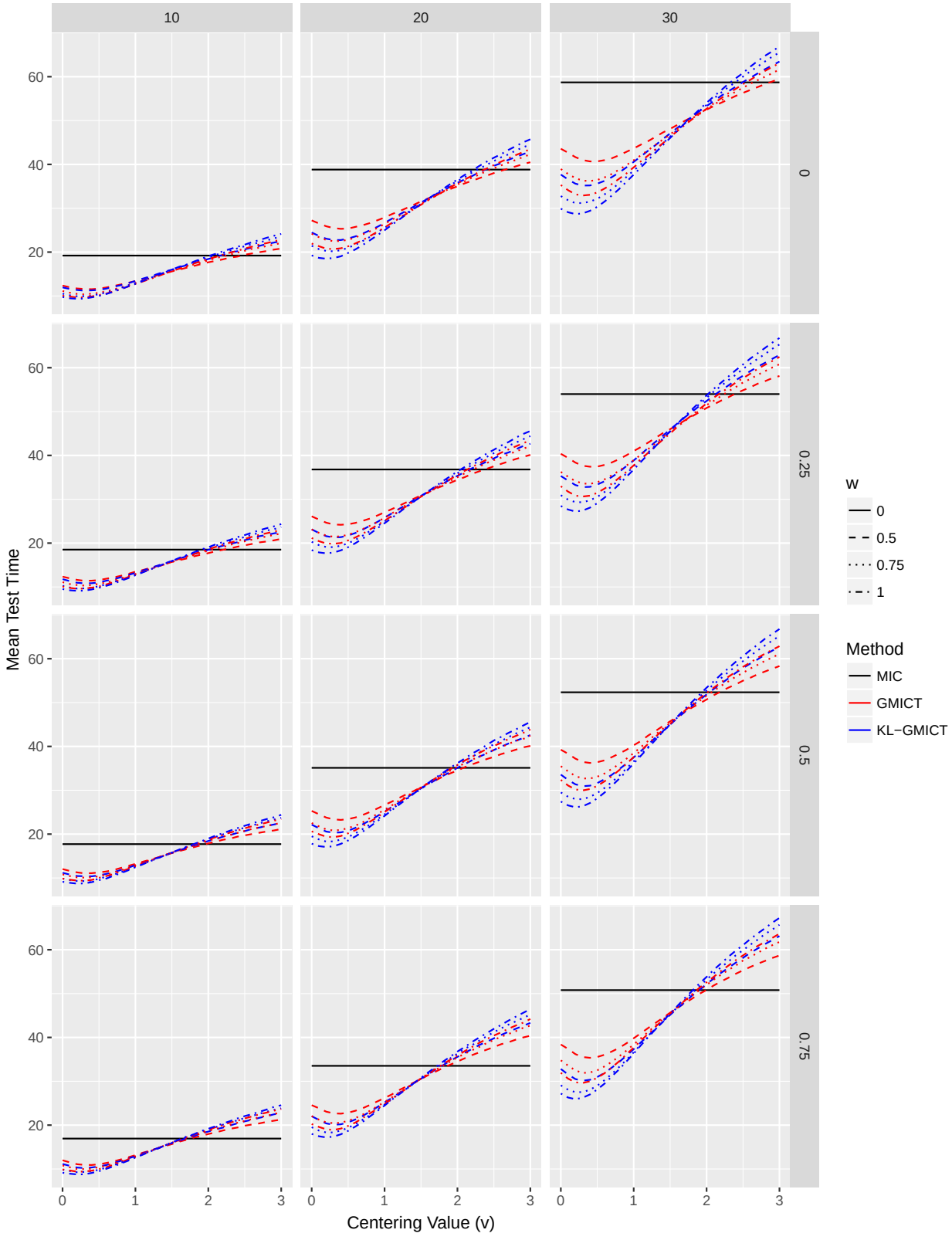


**Figure 5.5:** Plot of mean test time for Simulation 1.

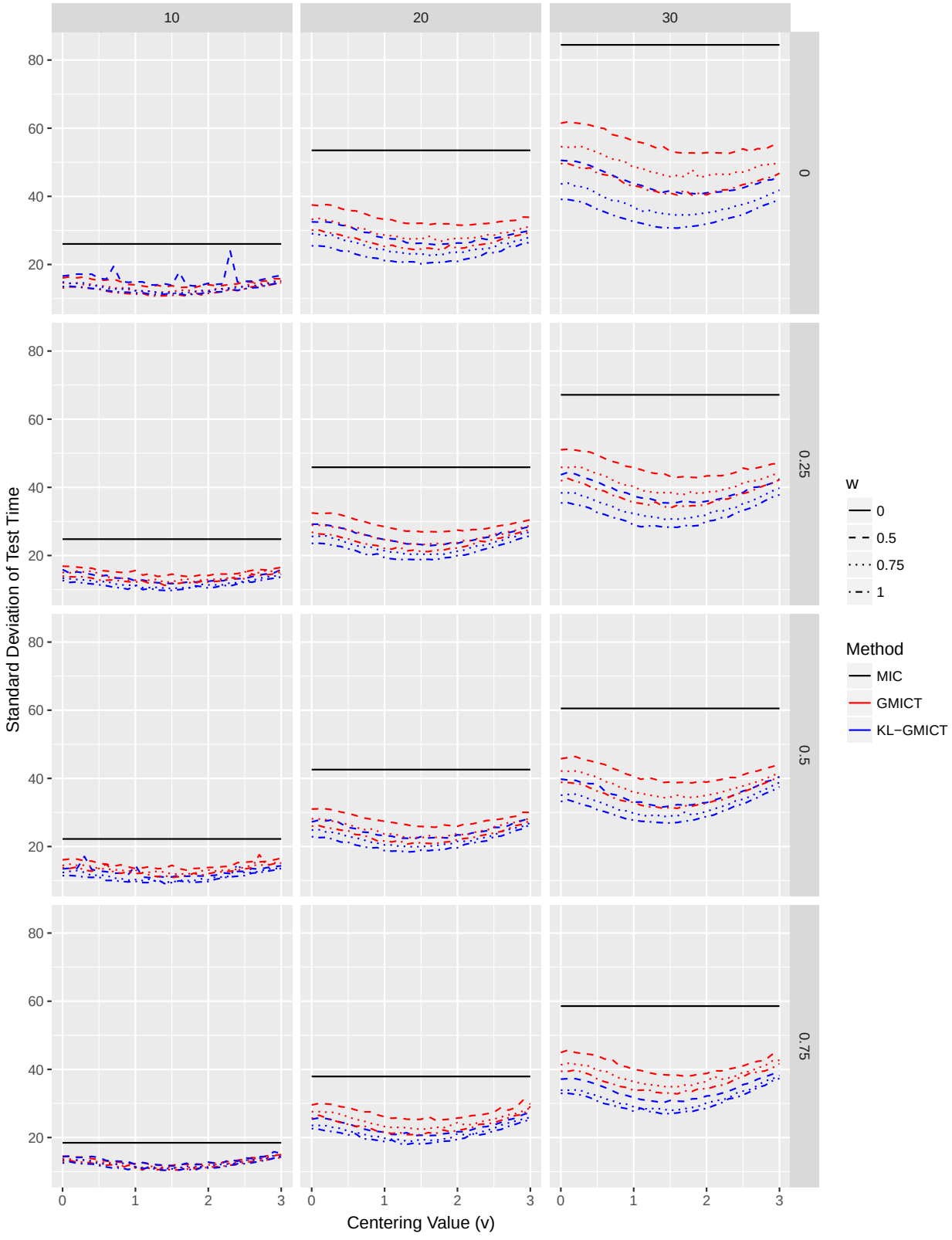


**Figure 5.6:** Plot of SD of test time for Simulation 1.

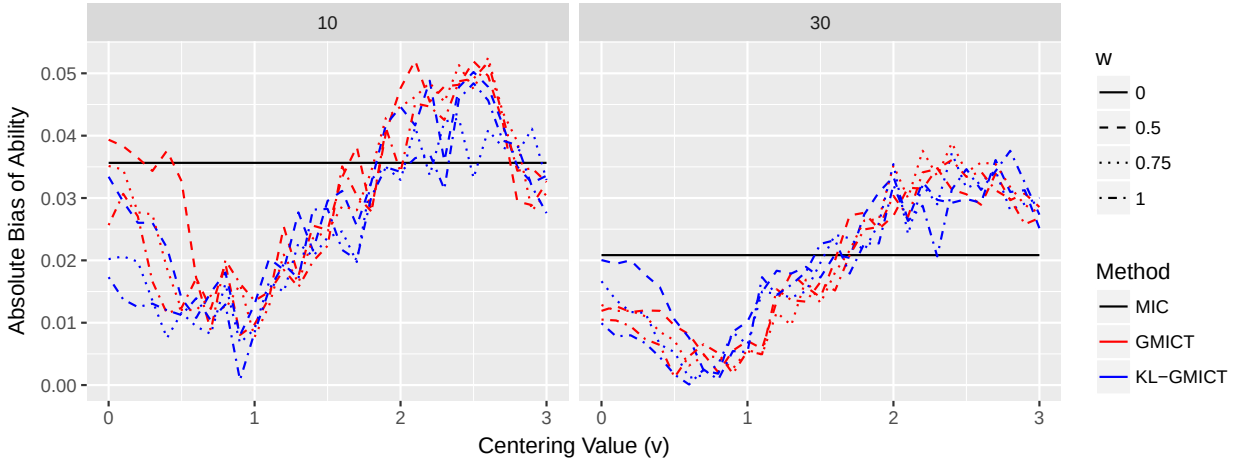


**Figure 5.7:** Plot of bias of ability estimates for Simulation 2.

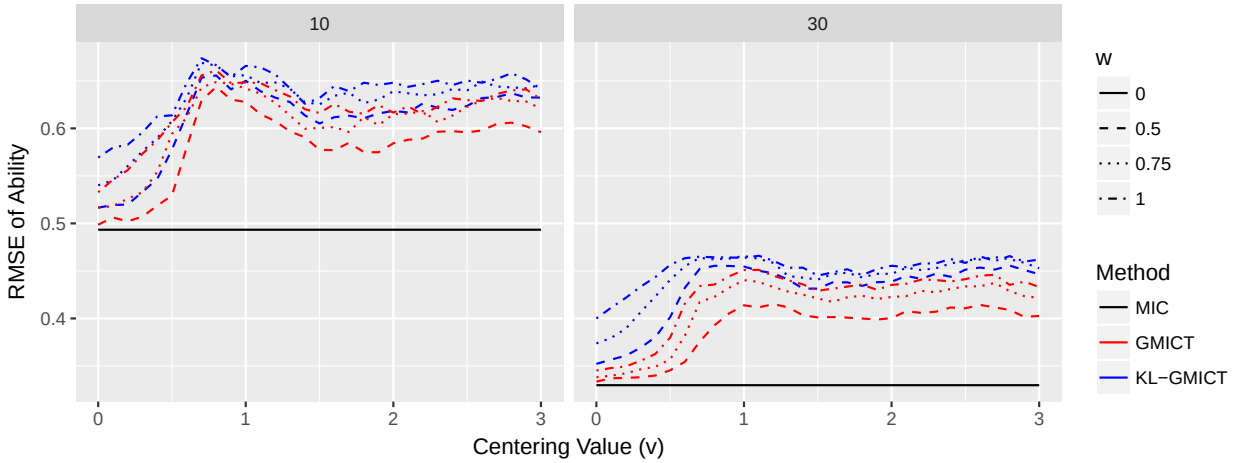


**Figure 5.8:** Plot of RMSE of ability estimates for Simulation 2.

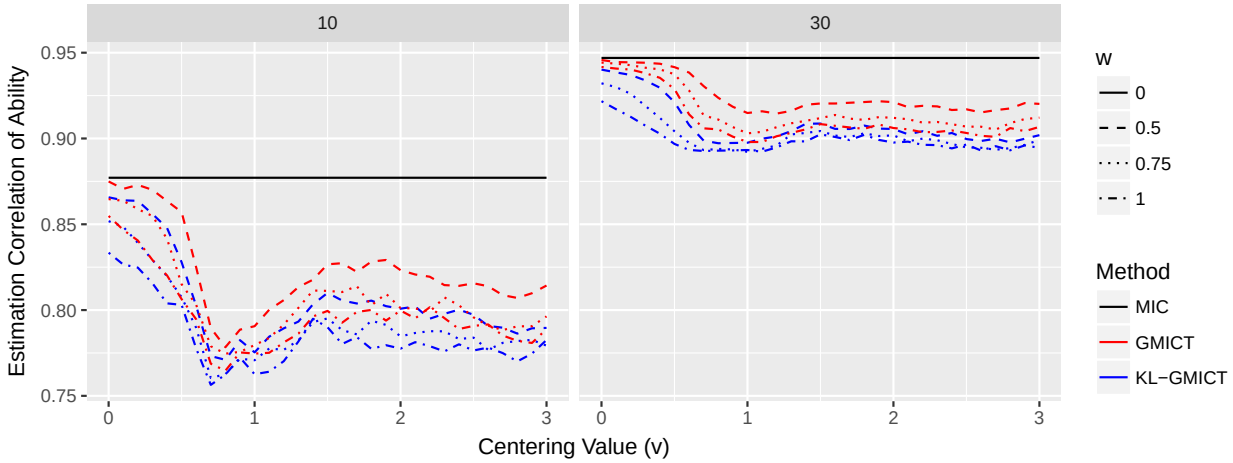


**Figure 5.9:** Plot of correlation of ability estimates with true values for Simulation 2.

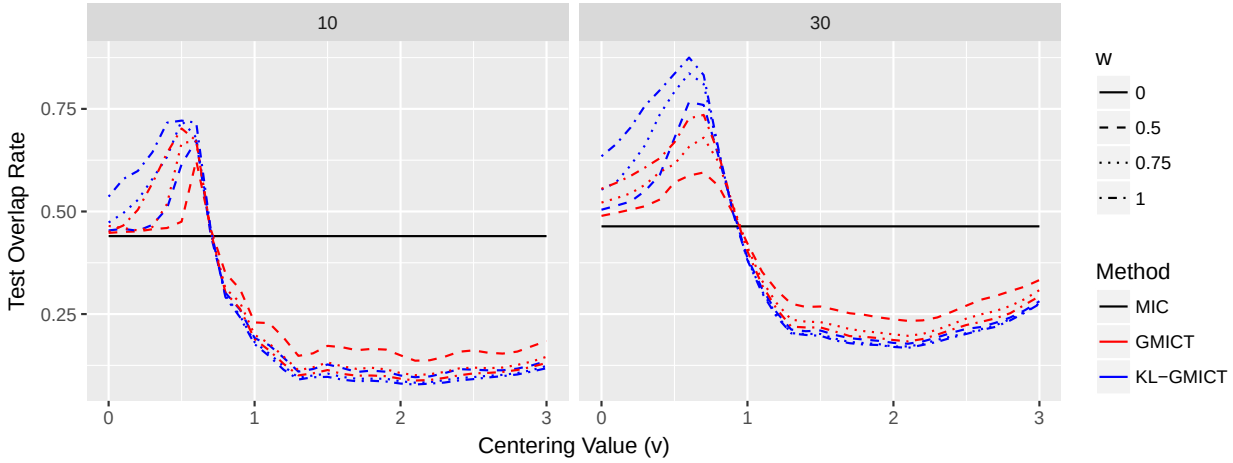


**Figure 5.10:** Plot of test overlap rate for Simulation 2.

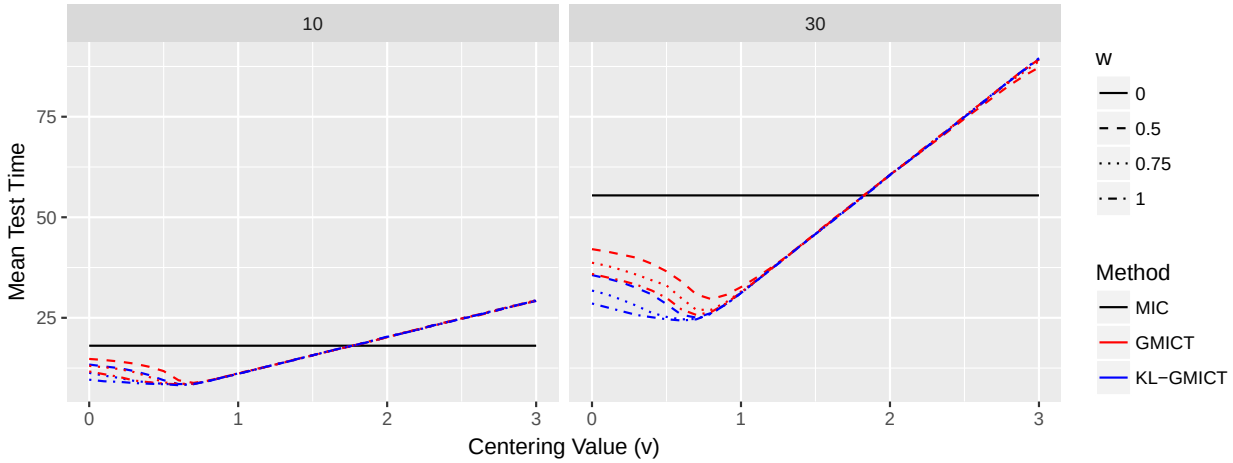


**Figure 5.11:** Plot of mean test time for Simulation 2.

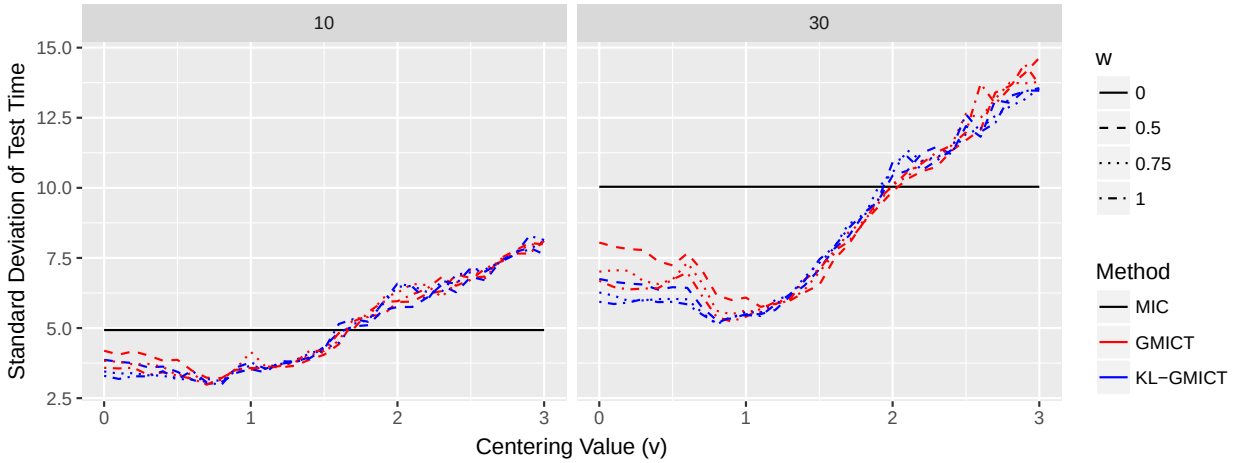


**Figure 5.12:** Plot of SD of test time for Simulation 2.

# Chapter 6

# Concluding Remarks

In this dissertation, I have examined several methods for incorporating response times into selection and estimation in adaptive testing. In Chapter 2, I introduced a joint maximum a posteriori estimator for the joint estimation of ability and speededness parameters. It improves estimation accuracy over the standard MLE, and if used in conjunction with the maximum information per time unit (MICT) approach, also makes test times decrease slightly over using MLE with MICT. In Chapter 3, I investigated the effect of using the MAP estimator in conjunction with the generalized MICT (GMICT; Choe & Kern, 2014) item selection method. This method successfully improves estimation accuracy compared to MLE with GMICT, while maintaining a low mean and standard deviation of test times. The only compromise is with respect to the test overlap rate, which, with an appropriate level of $v$ is still much lower than using MIC to select items. In Chapters 4 and 5, new selection methods—minimum expected posterior variance using RTs (MEPVT), time-weighted MEPVT (PV-GMICT), and time-weighted K-L index (KL-GMICT)—were developed in hopes of further improving the inherent trade-off between estimation accuracy, minimum test times, and test overlap. Unfortunately, the selection methods that were proposed did not seem to have much of an effect on outcomes over and above simply using the earlier proposed GMICT approach.

In examining the PV-GMICT and KL-GMICT, I have come to two conclusions. First, methods in this class of item selection procedures (which also includes GMICT) are sensitive to the choice of the weighting parameter $w$. This makes sense. A higher $w$ places more weight, ultimately, on choosing items that take less time to complete, whereas a lower $w$ places places more weight on choosing items that best determine the location of an examinee on the ability scale. The choice of $w$ can range from 0 to $\infty$. Interestingly, the specific choice of $w$ does not seem to have a uniform effect over the methods; a value of $w$ that is high for one method, is not necessarily high for another. This comes to a head when looking at PV-GMICT. In simulation results, it appears that the effect

of using PV-GMICT on each of the outcome measures is not sensitive to $w$. However, in truth, it must be that $w$ is simply near to point where the method is choosing items purely for shorter times, and that a choice of $w$ between 0 and 0.5 would put PV-GMICT closer to GMICT. This is also the case for KL-GMICT, but to a smaller extent. Which brings me to my second conclusion: it seems likely to be able to choose a value for $w$ such that GMICT, PV-GMICT, and KL-GMICT have highly similar results. In other words, any effect that a particular choice of $w$ may have while using PV-GMICT and KL-GMICT should be able to be closely replicated by simply using GMICT. Thus, GMICT as a selection method is highly flexible.

Now, one possibility that has not been fully explored is the effect of the item bank on the outcome values given the different selection methods. In particular, the real item bank has two relationships not modeled in the simulated item bank: $r_{b,\alpha^*} = 0.20$ and $r_{\alpha^*,\beta)=0.53}$. In particular, the relationship between the difficulty parameters $b$ and the time-discrimination parameters $\alpha$ in interesting to consider. The less related the measurement model parameters and the RT model parameters are, the more pronounced the effects of the item selection method should be. This needs to be explored further in the future.

In the end, in reflecting on the results and what they mean, I have come to the conclusion that this dissertation is primarily concerned with two things that psychometricians have been concerned themselves with for well over a century: reliability and validity.

Reliability as a central concern of this dissertation should be fairly obvious; estimation accuracy, by and large, is a direct measure of reliability. However, there is more to the issue of reliability than that. As noted in Section 1.4.1, in the Classical Test Theory model, Gulliksen (1950) showed that the total error scores $E$ can be decomposed into two parts: the number of items $W$ which the examinee gives an incorrect answer, and the number of items $U$ that the examinee does not reach. In this viewpoint, $N - X = W + U$, so a pure speed test is one where $P(W = 0) = 1$, and a pure power test is a test where $P(U = 0) = 1$. Thus, for all practical tests, the total error score is a function of both $U$ and $W$. A standard result of this viewpoint is that the reliability of the test is artificially increased by the presence of speededness (Crocker & Algina, 1986). Unsurprisingly, it seems that in real multiple-choice tests with time-limits, people simply start guessing at the end of a test. This has the implication that the reliability *decreases* for multiple-choice tests in the

presence of time-limits (Attali, 2004). Unfortunately, time-limits are a fact of life in the testing world, especially in multiple-choice testing regimes, so the best we can do in creating a testing system is to either directly model the change in test-taking strategies that occur due to time-limits, or to try to limit the effect that time-limits have on scores. What better way to limit the effect of time-limits than to simply choose items in such a way that an examinee never reaches the time-limit? The methods given here seek to minimize the amount of time that examinees spend on tests. While the time-limit effect and the extent to which these methods minimize it are not examined directly, indirect evidence that these methods would help is that not only do mean test times across examinees lessen, but also the variability in those test times lessen. Thus, the distribution of test times is more focused, with a smaller proportion of people exceeding a given time-limit. Future studies should be done to actually study this time-limit effect.

While reliability is a given, validity as a concern in this dissertation may be less obvious, and so may take some words to elucidate; those that follow are my attempt to do so. It has been known for a long time that not only is the accuracy of an examinee related to their ability, but the speededness of an examinee is as well. Van der Linden's (2006) hierarchical model can, in another light, be viewed as two-factor model with one factor $\theta$ for ability, and one factor $\tau$ for speededness. If this model is correct, then there an important implication. Simply, response times do have a relationship with ability, but that they are explained via a separable factor, and treating them as simply related to an ability (as in a one-factor model with item accuracies and RTs loading onto it) will bias the item parameters, which, in turn, will bias the estimates of ability. Of course, simply discarding response times may remedy this, and arguably it would, but only if the item parameters are truly independent. Unfortunately, it would seem that this should not be the case, since the relationship between ability and speededness in the real item bank was high ($\rho_{\theta\tau} = 0.76$). Rather it should be that estimating item parameters in the presence of response times should affect the values compared to estimating them without response times; this, in turn, would propagate the error forward to the estimated abilities. This was not investigated in this dissertation, but it seems worthy of some study going forward as the literature on misspecified models in IRT begins to widen (Bolt, Beng, & Lee, 2014; Lautenschlager & Park, 1988; Sahin, Walker, & Gelbal, 2014; Sun, 2015; Zhao & Hambleton, 2017). As such, including speededness in a model of test structure can be seen

as an important and, arguably, essential component of the overall validity of the score. For what it is worth, I believe that it is highly likely that two-factor model is not the best model for describing the relationship between response times and accuracies in a test, because as George Box famously said: "All models are wrong; some models are useful." Simply put, the hierarchical model is useful, if for no other reason than it has allowed for a fruitful conversation in the measurement world.

# Appendix A

# Posterior Predictive Distribution

It can be shown that the posterior predictive distribution of the response and response time to the $i$th item given $\tau$ and the responses/response times to the previous $i-1$ items is

$$p_i(X_i = x_i, T_i = t_i|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) = p_i(X_i = x_i, T_i = t_i|\mathbf{x}_{(-i)}, \tau)$$
$$= \frac{f(t_i|\tau)\int_{-\infty}^{\infty} f(x_i|\theta)\left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k}Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}{\int_{-\infty}^{\infty}\left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k}Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}.$$

*Proof.* By definition, we know that

$$f(x_i, \mathbf{x}_{(-i)}, t_i, \mathbf{t}_{(-i)}, \tau) = \int f(x_i, \mathbf{x}_{(-i)}, t_i, \mathbf{t}_{(-i)}, \theta, \tau)\ d\theta$$
$$= \int f(x_i, \mathbf{x}_{(-i)}, t_i, \mathbf{t}_{(-i)}|\theta, \tau) f(\theta, \tau)\ d\theta.$$

From the local independence assumption of the hierarchical model (van der Linden, 2007),

$$\int f(x_i, \mathbf{x}_{(-i)}, t_i, \mathbf{t}_{(-i)}|\theta, \tau) f(\theta, \tau)\ d\theta = \int f(x_i|\theta) f(\mathbf{x}_{(-i)}|\theta) f(t_i|\tau) f(\mathbf{t}_{(-i)}|\tau) f(\theta, \tau)\ d\theta$$
$$= \int f(x_i|\theta) f(t_i|\tau) f(\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \theta, \tau)\ d\theta$$
$$= \int f(x_i|\theta) f(t_i|\tau) f(\theta|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) f(\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)\ d\theta$$
$$= f(\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) f(t_i|\tau) \int f(x_i|\theta) f(\theta|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)\ d\theta.$$

Therefore,

$$f(x_i, \mathbf{x}_{(-i)}, t_i, \mathbf{t}_{(-i)}, \tau) = f(\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) f(t_i|\tau) \int f(x_i|\theta) f(\theta|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)\ d\theta$$

implies that

$$f(x_i, t_i|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) \equiv \frac{f(x_i, \mathbf{x}_{(-i)}, t_i, \mathbf{t}_{(-i)}, \tau)}{f(\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)} = f(t_i|\tau) \int f(x_i|\theta) f(\theta|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau)\ d\theta. \quad \text{(A.1)}$$

Now,

$$f(\theta|\mathbf{x}, \mathbf{t}, \tau) \propto f(\mathbf{x}|\theta, \mathbf{t}, \tau) f(\theta|\mathbf{t}, \tau)$$

and

$$f(\mathbf{x}|\theta, \mathbf{t}, \tau) = \frac{f(\mathbf{x}, \theta, \mathbf{t}, \tau)}{f(\theta, \mathbf{t}, \tau)}.$$

By local independence,

$$f(\mathbf{x}, \theta, \mathbf{t}, \tau) = f(\mathbf{x}, \mathbf{t}|\theta, \tau) f(\theta, \tau) = f(\mathbf{x}|\theta) f(\mathbf{t}|\tau) f(\theta|\tau) f(\tau),$$

and

$$f(\theta, \mathbf{t}, \tau) = f(\theta|\mathbf{t}, \tau) f(\mathbf{t}, \tau) = f(\theta|\mathbf{t}, \tau) f(\mathbf{t}|\tau) f(\tau).$$

Putting these together, we find that,

$$f(\mathbf{x}|\theta, \mathbf{t}, \tau) = \frac{f(\mathbf{x}|\theta) f(\mathbf{t}|\tau) f(\theta|\tau) f(\tau)}{f(\theta|\mathbf{t}, \tau) f(\mathbf{t}|\tau) f(\tau)} = \frac{f(\mathbf{x}|\theta) f(\theta|\tau)}{f(\theta|\mathbf{t}, \tau)}.$$

Therefore,

$$\begin{aligned} f(\theta|\mathbf{x}, \mathbf{t}, \tau) &\propto f(\mathbf{x}|\theta, \mathbf{t}, \tau) f(\theta|\mathbf{t}, \tau) \\ &= \frac{f(\mathbf{x}|\theta) f(\theta|\tau)}{f(\theta|\mathbf{t}, \tau)} f(\theta|\mathbf{t}, \tau) \\ &= f(\mathbf{x}|\theta) f(\theta|\tau), \end{aligned}$$

and thus,

$$\begin{aligned} f(\theta|\mathbf{x}_{(-i)}, \mathbf{t}_{(-i)}, \tau) &= \frac{f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)}{\int f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta} \\ &= \frac{\left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta}{\int \left[\prod_{k=1}^{i-1} P_k(\theta)^{x_k} Q_k(\theta)^{1-x_k}\right] f(\theta|\tau)\ d\theta} \end{aligned} \quad \text{(A.2)}$$

Substituting (A.2) into (A.1) leads to the desired result. □

# Appendix B

# Posterior Variance

It can be shown that the posterior variance of $\theta$ given $\tau$ and the responses/response times to the $i$ items is

$$\begin{aligned}\mathrm{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau) &= \mathrm{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \tau) \\ &= \mathrm{E}(\theta^2|\mathbf{x}_{(-i)}, X_i = x_i, \tau) - \mathrm{E}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \tau)^2 \\ &= \frac{\int \theta^2 f(x_i|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta}{\int f(x_i|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta} - \left[\frac{\int \theta f(x_i|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta}{\int f(x_i|\theta) f(\mathbf{x}_{(-i)}|\theta) f(\theta|\tau)\ d\theta}\right]^2 \quad \text{(B.1)}\end{aligned}$$

*Proof.* By definition,

$$\begin{aligned}&\mathrm{Var}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau) \\ &\quad = \mathrm{E}(\theta^2|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau) - \mathrm{E}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau)^2. \quad \text{(B.2)}\end{aligned}$$

Additionally, by definition,

$$\mathrm{E}(\theta^2|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau) = \int \theta^2 f(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau)\ d\theta \quad \text{(B.3)}$$

$$\mathrm{E}(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau) = \int \theta f(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau)\ d\theta. \quad \text{(B.4)}$$

From (A.2), we know that

$$f(\theta|\mathbf{x}_{(-i)}, X_i = x_i, \mathbf{t}_{(-i)}, T_i = t_i, \tau) = \frac{f(\mathbf{x}_{(-i)}|\theta) f(x_i|\theta) f(\theta|\tau)}{\int f(\mathbf{x}_{(-i)}|\theta) f(x_i|\theta) f(\theta|\tau)\ d\theta} \quad \text{(B.5)}$$

(B.1) directly follows from (B.2), (B.3), (B.4), and (B.5). □

# Appendix C

# C++ Code

```
#include <RcppArmadillo.h>
#include <string>
#include <algorithm>
// [[Rcpp::depends(RcppArmadillo)]]

using namespace Rcpp;

// This computes P_i(theta_j) for the 3PL model.
// [[Rcpp::export]]
double threepl(double a, double b, double c, double theta){
  double p = c + (1 - c) / (1 + exp(-a * (theta - b)));
  return p;
}

// This generates a response matrix assuming a 3PL model.
// [[Rcpp::export]]
arma::mat genres(arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &theta){
  unsigned int n_items = a.n_elem;
  unsigned int n_persons = theta.n_elem;
  arma::mat response(n_persons, n_items);
  for(unsigned int i = 0; i < n_items; i++){
    for(unsigned int j = 0; j < n_persons; j++){
      response(j, i) = 1 * (R::runif(0, 1) <= threepl(a(i), b(i), c(i),
```

```
                                                    theta(j)));
    }
  }
  return(response);
}

// This generates a matrix log-times assuming the log-normal model of RTs.
// [[Rcpp::export]]
arma::mat genlogtime(arma::vec &alpha, arma::vec &beta, arma::vec &tau){
  unsigned int n_items = alpha.n_elem;
  unsigned int n_persons = tau.n_elem;
  arma::mat logtime(n_persons, n_items);
  for(unsigned int i = 0; i < n_items; i++){
    for(unsigned int j = 0; j < n_persons; j++){
      logtime(j, i) = R::rnorm(beta(i) - tau(j), 1 / alpha(i));
    }
  }
  return(logtime);
}

// This computes Fisher information for a single item.
// [[Rcpp::export]]
double Fisher_info(double a, double b, double c, double theta){
  double p = threepl(a, b, c, theta);
  double FI = pow(a, 2.0) * ((1 - p)/p) * pow((p - c)/(1 - c), 2.0);
  return FI;
}

// MIC: Maximum Information Criterion
```

```
// This returns the index for the item that maximizes Fisher information.
// [[Rcpp::export]]
unsigned int MIC(arma::vec &a, arma::vec &b, arma::vec &c, double theta){
  unsigned int n_items = a.n_elem;
  arma::vec FI(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    FI(i) = Fisher_info(a(i), b(i), c(i), theta);
  }
  return arma::index_max(FI);
}

// This computes the KL_index for a single item.
// [[Rcpp::export]]
double KL_index(double a, double b, double c, double theta, double delta = 1,
                unsigned int n_quad = 30){
  arma::vec p(n_quad);
  arma::vec quad_pts = arma::linspace<arma::vec>(theta - delta, theta + delta,
                                                 n_quad);
  double width = 2 * delta / (n_quad - 1);

  for(unsigned int i = 0; i < n_quad; i++){
    p(i) = threepl(a, b, c, quad_pts(i));
  }
  double p_theta = threepl(a, b, c, theta);

  double integral1 = width * arma::sum(log(p));
  double integral2 = width * arma::sum(log(1 - p));

  return 2 * delta * (p_theta * log(p_theta) + (1 - p_theta) *
```

```
    log(1 - p_theta)) - (p_theta * integral1 + (1 - p_theta) * integral2);
}

// KLI: Kullback-Leibler Index
// This returns the index for the item that maximizes the K-L index.
// [[Rcpp::export]]
unsigned int KLI(arma::vec &a, arma::vec &b, arma::vec &c, double theta,
                 double delta = 1, unsigned int n_quad = 30){
  unsigned int n_items = a.n_elem;
  arma::vec KI(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    KI(i) = KL_index(a(i), b(i), c(i), theta, delta, n_quad);
  }
  return arma::index_max(KI);
}

// GMICT: Generalized Time-weighted Maximum Information Criterion,
//        MICT (v = 0, w = 1)
// This returns the index for the item that maximizes the time-weighted
//   Fisher information.
// [[Rcpp::export]]
unsigned int GMICT(arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &alpha,
                   arma::vec &beta, double theta, double tau, double v = 0,
                   double w = 1){
  unsigned int n_items = a.n_elem;
  arma::vec GMI(n_items);
  arma::vec denom(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    double expected_RT = exp(beta(i) - tau + 1/(2 * pow(alpha(i), 2)));
```

```
    denom(i) = pow(std::abs(expected_RT - v), w);
    GMI(i) = Fisher_info(a(i), b(i), c(i), theta);
  }
  GMI = GMI / denom;
  return arma::index_max(GMI);
}

// KLIT: KL-GMICT
// This returns the index for item that maximizes the time-weighted K-L index.
// [[Rcpp::export]]
unsigned int KL_GMICT(arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &alpha,
                      arma::vec &beta, double theta, double tau, double delta,
                      double v = 0, double w = 1, unsigned int n_quad = 30){
  unsigned int n_items = a.n_elem;
  arma::vec KI(n_items);
  arma::vec denom(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    double expected_RT = exp(beta(i) - tau + 1/(2 * pow(alpha(i), 2)));
    denom(i) = pow(std::abs(expected_RT - v), w);
    KI(i) = KL_index(a(i), b(i), c(i), theta, delta, n_quad);
  }
  KI = KI / denom;
  return arma::index_max(KI);
}

// This is a light-weight helper function that performs the task of computing
// the likelihood across a set of quadrature points for a single item.
// [[Rcpp::export]]
Rcpp::List posterior_var_index_list(double tau, double a, double b, double c,
```

```
              arma::vec &fx, arma::vec &fxi, double mu_th = 0, double mu_tau = 0,
              double var_th = 1, double var_tau = 1, double covar = 0.5,
              double lower_th = -4, double upper_th = 4,
              unsigned int n_quad = 11){
  arma::vec quad_pts = arma::linspace<arma::vec>(lower_th, upper_th, n_quad);
  arma::vec quad_pts2 = quad_pts % quad_pts;
  double mu_th_given_tau = mu_th + (covar / var_tau) * (tau - mu_tau);
  double var_th_given_tau = var_th - covar * sqrt(var_th / var_tau);
  arma::vec f_th_given_tau(n_quad);
  arma::vec p_new(n_quad);
  for(unsigned int i = 0; i < n_quad; i++){
    f_th_given_tau(i) = R::dnorm(quad_pts(i), mu_th_given_tau,
                                 sqrt(var_th_given_tau), 0);
    p_new(i) = threepl(a, b, c, quad_pts(i));
  }
  arma::vec fx_new = fx % fxi;
  return Rcpp::List::create(Rcpp::Named("fx", fx_new),
                            Rcpp::Named("fxi0", 1 - p_new),
                            Rcpp::Named("fxi1", p_new));
}

// This function fully computes the posterior variance for a single item.
// [[Rcpp::export]]
double posterior_var_index(double tau, double a, double b, double c,
          arma::vec &fx, arma::vec &fxi, double mu_th = 0, double mu_tau = 0,
          double var_th = 1, double var_tau = 1, double covar = 0.5,
          double lower_th = -4, double upper_th = 4, unsigned int n_quad = 11){
  double integ1_0, integ1_1, integ2_0, integ2_1, integ3_0, integ3_1;
  arma::vec quad_pts = arma::linspace<arma::vec>(lower_th, upper_th, n_quad);
```

```
  arma::vec quad_pts2 = quad_pts % quad_pts;
  double delta_th = (upper_th - lower_th) / (n_quad - 1);
  double mu_th_given_tau = mu_th + (covar / var_tau) * (tau - mu_tau);
  double var_th_given_tau = var_th - covar * sqrt(var_th / var_tau);
  arma::vec f_th_given_tau(n_quad);
  arma::vec p_new(n_quad);
  for(unsigned int i = 0; i < n_quad; i++){
    f_th_given_tau(i) = R::dnorm(quad_pts(i), mu_th_given_tau,
                                 sqrt(var_th_given_tau), 0);
    p_new(i) = threepl(a, b, c, quad_pts(i));
  }
  arma::vec fx_new = fx % fxi;
  arma::vec part0 = fx_new % f_th_given_tau;
  arma::vec part1 = p_new % part0;
  part0 = (1 - p_new) % part0;
  integ1_0 = delta_th * arma::sum(quad_pts2 % part0);
  integ1_1 = delta_th * arma::sum(quad_pts2 % part1);
  integ2_0 = delta_th * arma::sum(quad_pts % part0);
  integ2_1 = delta_th * arma::sum(quad_pts % part1);
  integ3_0 = delta_th * arma::sum(part0);
  integ3_1 = delta_th * arma::sum(part1);
  double out = (integ1_0 - pow(integ2_0, 2.0) / integ3_0) +
              (integ1_1 - pow(integ2_1, 2.0) / integ3_1);
  return out;
}

// MEPV: Minimum Time-weighted Expected Posterior Variance Criterion
// This function returns the index of the item that minimizes the posterior
// variance.
```

```
// [[Rcpp::export]]
Rcpp::List MEPV(double tau, arma::vec &a, arma::vec &b, arma::vec &c,
                arma::vec &fx, arma::vec &fxi, double mu_th = 0,
                double mu_tau = 0, double var_th = 1, double var_tau = 1,
                double covar = 0.5, double lower_th = -4, double upper_th = 4,
                unsigned int n_quad = 11){
  unsigned int N = a.n_elem;
  arma::vec pvars(N);
  for(unsigned int i = 0; i < N; i++){
    pvars(i) = posterior_var_index(tau, a(i), b(i), c(i), fx, fxi, mu_th, mu_tau,
          var_th, var_tau, covar, lower_th, upper_th, n_quad);
  }
  unsigned int chosen = arma::index_min(pvars);
  Rcpp::List out = posterior_var_index_list(tau, a(chosen), b(chosen), c(chosen),
                            fx, fxi, mu_th, mu_tau, var_th, var_tau, covar,
                            lower_th, upper_th, n_quad);
  return Rcpp::List::create(Rcpp::Named("choice", chosen),
                            Rcpp::Named("fx", out(0)),
                            Rcpp::Named("fxi0", out(1)),
                            Rcpp::Named("fxi1", out(2)));
}

// MEPVT: PV-GMICT
// This function returns the index of the item that minimizes the time-weighted
// posterior variance.
// [[Rcpp::export]]
Rcpp::List PV_GMICT(double tau, arma::vec &a, arma::vec &b, arma::vec &c,
              arma::vec &alpha, arma::vec &beta, arma::vec &fx, arma::vec &fxi,
              double v = 0, double w = 1, double mu_th = 0, double mu_tau = 0,
```

```
                double var_th = 1, double var_tau = 1, double covar = 0.5,
                double lower_th = -4, double upper_th = 4,
                unsigned int n_quad = 11){
  unsigned int N = a.n_elem;
  arma::vec pvars(N);
  arma::vec denom(N);
  for(unsigned int i = 0; i < N; i++){
    double expected_RT = exp(beta(i) - tau + 1/(2 * pow(alpha(i), 2)));
    denom(i) = pow(std::abs(expected_RT - v), w);
    pvars(i) = posterior_var_index(tau, a(i), b(i), c(i), fx, fxi, mu_th, mu_tau,
          var_th, var_tau, covar, lower_th, upper_th, n_quad);
  }
  unsigned int chosen = arma::index_min(pvars % denom);
  Rcpp::List out = posterior_var_index_list(tau, a(chosen), b(chosen), c(chosen),
                   fx, fxi, mu_th, mu_tau, var_th, var_tau, covar, lower_th,
                     upper_th, n_quad);
  return Rcpp::List::create(Rcpp::Named("choice", chosen),
                            Rcpp::Named("fx", out(0)),
                            Rcpp::Named("fxi0", out(1)),
                            Rcpp::Named("fxi1", out(2)));
}

// This function compute the log-likelihood of the 3PL model at a single theta
// value.
// [[Rcpp::export]]
double threepl_likelihood(arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &x,
                          double theta){
  unsigned int n_items = x.n_elem;
  double log_likelihood = 0;
```

```
  for(unsigned int i = 0; i < n_items; i++){
    if(x(i) == 0){
      log_likelihood += log(1 - threepl(a(i), b(i), c(i), theta));
    }else{
      log_likelihood += log(threepl(a(i), b(i), c(i), theta));
    }
  }
  return exp(log_likelihood);
}

// This function computes the first derivative of the log-likelihood
// of the 3PL model at a single theta value.
// [[Rcpp::export]]
double dlogfth(arma::vec &x, double theta, arma::vec &a, arma::vec &b,
               arma::vec &c){
  unsigned int n_items = x.n_elem;
  arma::vec p(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    p(i) = threepl(a(i), b(i), c(i), theta);
  }
  arma::vec part = (a % (p - c)) / ((1 - c) % p);
  double dlog = arma::conv_to< double >::from(part.t() * (x - p));
  return dlog;
}

// This function computes the first and second derivatives of the
// log-likelihood of the 3PL model at a single theta value.
// [[Rcpp::export]]
arma::vec dlogfth_vec(arma::vec &x, double theta, arma::vec &a, arma::vec &b,
```

```
                      arma::vec &c){
  unsigned int n_items = x.n_elem;
  arma::vec p(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    p(i) = threepl(a(i), b(i), c(i), theta);
  }
  arma::vec part1 = (1 - c) % p;
  arma::vec part2 = a % (p - c) / part1;
  arma::vec dlog(2);
  dlog(0) = arma::sum(part2 % (x - p));
  dlog(1) = arma::sum(part2 % a % (1 - p) % (x % c - p % p) / part1);
  return dlog;
}

// This function gives the EAP estimate of theta.
// [[Rcpp::export]]
double EAPcpp(arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &x,
              double th_mu = 0., double th_var = 1., double lower_th = -4.,
              double upper_th = 4., unsigned int n_points = 21){
  arma::vec quad_pts = arma::linspace(lower_th, upper_th, n_points);
  double numerator = 0;
  double denominator = 0;
  for(unsigned int i = 0; i < n_points; i++){
    double value = R::dnorm(quad_pts(i), th_mu, sqrt(th_var), 0) *
                   threepl_likelihood(a, b, c, x, quad_pts(i));
    denominator += value;
    numerator += value * quad_pts(i);
  }
  return numerator / denominator;
```

```
}

// This function gives the ML estimate of tau.
// [[Rcpp::export]]
double MLE_tau(arma::vec &alpha, arma::vec &beta, arma::vec &lnt){
  arma::vec alp2 = alpha % alpha;
  double out = arma::sum(alp2 % (beta - lnt)) / arma::sum(alp2);
  return out;
}

// This function gives the ML estimate of theta if possible. If the
// MLE is not finite, then it automatically gives the EAP estimate instead.
// [[Rcpp::export]]
double MLE_EAP(arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &x,
          double th_mu = 0., double th_var = 1., double lower_lim = -4.,
          double upper_lim = 4., unsigned int n_points = 21,
          double tol = .00000001){
  unsigned int n_items = x.n_elem;
  if((sum(x) == 0)|(sum(x) == n_items)){
    return EAPcpp(a, b, c, x, th_mu, th_var, lower_lim, upper_lim, n_points);
  }else{
    double low = -1;
    double high = 1;
    double f_low = dlogfth(x, low, a, b, c);
    double f_high = dlogfth(x, high, a, b, c);
    while(((f_low > 0) - (f_low < 0)) == ((f_high > 0) - (f_high < 0))){
      low += -1;
      high += 1;
      f_low = dlogfth(x, low, a, b, c);
```

```
  f_high = dlogfth(x, high, a, b, c);
  if((((f_low > 0) - (f_low < 0)) == 0) |
     (((f_high > 0) - (f_high < 0)) == 0) | (low == lower_lim)){
    return EAPcpp(a, b, c, x, th_mu, th_var, lower_lim, upper_lim, n_points);
  }
}
lower_lim = low;
upper_lim = high;
double th_new = R::runif(lower_lim, upper_lim);
double change = 1000;
unsigned int iter = 0;
while(change > tol){
  ++iter;
  double th_old = th_new;
  arma::vec d = dlogfth_vec(x, th_old, a, b, c);
  double z = th_old - d(0) / d(1);
  if((z > lower_lim) & (z < upper_lim) & (iter <= 30)){
    th_new = z;
  }else{
    iter = 0;
    double m = lower_lim + (upper_lim - lower_lim)/2;
    f_low = dlogfth(x, lower_lim, a, b, c);
    double f_m = dlogfth(x, m, a, b, c);
    if(((f_low > 0) - (f_low < 0)) == ((f_m > 0) - (f_m < 0))){
      lower_lim = m;
    }else{
      upper_lim = m;
    }
    th_new = R::runif(lower_lim, upper_lim);
```

```
      }
      change = std::abs(th_new - th_old);
    }
    return th_new;
  }
}

// This function computes the first and second derivatives of posterior of
// theta and tau under van der Linden's model.
// [[Rcpp::export]]
arma::vec dlogfth_tau(arma::vec &x, arma::vec &lnt, double theta, double tau,
             arma::vec &a, arma::vec &b, arma::vec &c, arma::vec &alpha,
             arma::vec &beta, double mu_th = 0, double mu_tau = 0,
             double var_th = 1, double var_tau = 1, double covar = .5){
  unsigned int n_items = x.n_elem;
  arma::vec p(n_items);
  for(unsigned int i = 0; i < n_items; i++){
    p(i) = threepl(a(i), b(i), c(i), theta);
  }
  arma::vec part1 = (1 - c) % p;
  arma::vec part2 = a % (p - c) / part1;
  double det_sig = var_th * var_tau - pow(covar, 2.0);
  arma::vec alp2 = alpha % alpha;

  arma::vec dlogf(5);
  // dlogf_dth
  dlogf(0) = arma::sum(part2 % (x - p)) -
             (var_tau * (theta - mu_th) - covar * (tau - mu_tau)) / det_sig;
  // dlogf_dtau
```

```
  dlogf(1) = -arma::sum(alp2 % (lnt - (beta - tau))) -
             (var_th * (tau - mu_tau) - covar * (theta - mu_th)) / det_sig;
  // d2logf_dth2
  dlogf(2) = arma::sum(part2 % a % (1 - p) % (x % c - p % p) / part1) -
             (var_tau / det_sig);
  // d2logf_dtau2
  dlogf(3) = -arma::sum(alp2) - var_th / det_sig;
  // d2logf_dthdtau
  dlogf(4) = covar / det_sig;
  return dlogf;
}

// This function computes the MAP estimates of theta and tau under
// van der Linden's model.
// [[Rcpp::export]]
arma::vec MAP_th_tau(arma::vec &x, arma::vec &lnt, arma::vec &a, arma::vec &b,
             arma::vec &c, arma::vec &alpha, arma::vec &beta, double mu_th = 0,
             double mu_tau = 0, double var_th = 1, double var_tau = 1,
             double covar = 0.5, double tol = 0.00001, double start_th = NA_REAL,
             double start_tau = NA_REAL){
  arma::vec alp2 = alpha % alpha;
  double th_start;
  double tau_start;
  if(NumericVector::is_na(start_th)){
    th_start = R::rnorm(0, 1);
  }else{
    th_start = start_th;
  }
  if(NumericVector::is_na(start_tau)){
```

```
  tau_start = sum(alp2 % (lnt - beta)) / sum(alp2);
}else{
  tau_start = start_tau;
}
double change = 1000;
unsigned int i = 0;
double th_new = th_start;
double tau_new = tau_start;
while( change > tol ){
  ++i;
  double th_old = th_new;
  double tau_old = tau_new;
  arma::vec d = dlogfth_tau(x, lnt, th_old, tau_old, a, b, c, alpha, beta,
                            mu_th, mu_tau, var_th, var_tau, covar);
  double detA = d(2) * d(3) - pow(d(4), 2.0);
  double change_th = (d(0) * d(3) - d(1) * d(4))/detA;
  double change_tau = (d(1) * d(2) - d(0) * d(4))/detA;
  th_new = th_old - change_th;
  tau_new = tau_old - change_tau;
  change = pow(change_th, 2.0) + pow(change_tau, 2.0);
  if((i == 40) | (NumericVector::is_na(change)) |
     any(is_infinite(NumericVector::create(change)))){
    i = 0;
    th_new = R::rnorm(0, 1);
    tau_new = tau_start;
    change = 1000;
  }
}
arma::vec out(2);
```

```
  out(0) = th_new;
  out(1) = tau_new;
  return out;
}
```

# Appendix D

# R Code

```
#-----#
# CAT #
#-----#

# This is a full CAT simulation. It incorporates all of the selection methods
# described in this dissertation. Several options are described below, as well
# as the input values.
#
# examinees: examinee parameters (in order of theta,tau) specified as n x 2
#   matrix
# items: item parameters (in order of a,b,c,A,B) specified as m x 5 matrix
# test.length: test length (50 by default)
# n.start.items: number of random items to start (1 by default)
# method: item selection criterion (MIC, MICT, GMICT, KLI, KLIT, MEPV, MEPVT)
# w: exponent values for GMICT/KLIT/MEPVT (scalar or vector)
# v: centering values for GMICT/KLIT/MEPVT (scalar or vector)
# seed: set seed value for direct replications (NULL by default)
# width: a constant governing the range of integration for KLI and KLIT (3 by
#   default)
# weighted: either bias, MSE, and correlation are computed by weighting by
#   f(theta) or not (F by default)
# conditional: either conditional bias and conditional MSE are computed or not
#   (T by default)
```

```
# MAP: either use joint MAP or not. If not, use MLE instead. (F by default)
# n.quad1: number of quadrature points for KLI and KLIT (250 by default)
# n.quad2: number of quadrature points for MEPV and MEPVT (21 by default)

CAT.cond <- function(examinees, items, test.length=50, n.start.items=1, th.mu=0,
                     th.var=1, tau.mu = 0, tau.var = 1, covar = .5, method='MIC',
                     v = 0, w = 1, width = 3, seed=NA, weighted = T,
                     conditional = T, MAP = F, n.quad1 = 30, n.quad2 = 25){
  if(!is.na(seed)){
    set.seed(seed)
  }
  n.persons <- nrow(examinees)
  n.items <- nrow(items)
  theta <- examinees[,1]
  tau <- examinees[,2]
  a <- items[,1]
  b <- items[,2]
  c <- items[,3]
  A <- items[,4]
  B <- items[,5]
  responses <- genres(a, b, c, theta)
  log.times <- genlogtime(A, B, tau)
  if(any(method == c('MIC', 'MICT', 'MEPV', 'KLI'))){
    w <- 1; v <- 0;
    results <- data.frame(matrix(0, 1, 9))
    names(results) <- c('MSE.the', 'MSE.tau', 'bias.the', 'bias.tau', 'cor.the',
                        'cor.tau', 'MTT', 'STT', 'Chi2')
  }
  if(any(method == c('GMICT', 'MEPVT', 'KLIT'))){
```

```
  results <- data.frame(matrix(0, length(w)*length(v), 11))
  names(results) <- c('w', 'v', 'MSE.the', 'MSE.tau', 'bias.the', 'bias.tau',
                      'cor.the', 'cor.tau', 'MTT', 'STT', 'Chi2')
}
if(conditional){
  results2 <- list()
}
q <- 0
for(w.ind in w){
  for(v.ind in v){
    if(!is.na(seed)){
      seed <- seed + 1
      set.seed(seed)
    }
    if(any(method == c('GMICT', 'MEPVT', 'KLIT'))){
      cat(paste('w = ', w.ind, ', v = ', v.ind, sep = ''),'\n')
      flush.console()
    }
    q <- q + 1
    ec.MIC <- rep(0, n.items)
    theta.MIC <- rep(0, n.persons)
    tau.MIC <- rep(0, n.persons)
    test.time <- rep(0, n.persons)
    for(i in 1:n.persons){
      it <- c(rep(0, test.length), n.items + 1)
      u <- rep(2, test.length)
      lnt <- rep(0, test.length)
      if(i %% 1000 == 0){
        print(paste('On', 'Person', i))
```

```
  flush.console()
}
if(n.start.items > 0){
  it[1:n.start.items] <- sample(n.items, n.start.items)
  ec.MIC[it[1:n.start.items]] <- ec.MIC[it[1:n.start.items]] + 1
  u[1:n.start.items] <- responses[i, it[1:n.start.items]]
  lnt[1:n.start.items] <- log.times[i, it[1:n.start.items]]
  if(MAP == T){
    temp <- MAP_th_tau(u[1:n.start.items], lnt[1:n.start.items],
                       a[it[1:n.start.items]], b[it[1:n.start.items]],
                       c[it[1:n.start.items]], A[it[1:n.start.items]],
                       B[it[1:n.start.items]], th.mu, tau.mu,
                       th.var, tau.var, covar)
    th.est <- temp[1]
    tau.est <- temp[2]
  }else{
    th.est <- MLE_EAP(a[it[1:n.start.items]], b[it[1:n.start.items]],
                      c[it[1:n.start.items]], u[1:n.start.items],
                      th.mu, th.var)
    tau.est <- MLE_tau(A[it[1:n.start.items]], B[it[1:n.start.items]],
                       lnt[1:n.start.items])
  }
}
if(n.start.items == 0){
  th.est <- tau.est <- 0
}
if(method == 'MEPV' | method == 'MEPVT'){
  xx <- seq(-4, 4, len = n.quad2)
  fx <- numeric(n.quad2)
```

```
  for(quad.point in 1:n.quad2){
    fx[quad.point] <-
      threepl_likelihood(a[it[1:n.start.items]], b[it[1:n.start.items]],
                         c[it[1:n.start.items]], u[1:n.start.items],
                         xx[quad.point])
  }
  fxi <- numeric(n.quad2) + 1
}
for(j in (n.start.items + 1):test.length) {
  if(method=='MIC'){
    it[j] <- (1:n.items)[-it][MIC(a[-it], b[-it], c[-it], th.est) + 1]
  }
  if(any(method==c('MICT','GMICT'))){
    it[j] <- (1:n.items)[-it][GMICT(a[-it], b[-it], c[-it], A[-it],
                                    B[-it], th.est, tau.est, v.ind,
                                    w.ind) + 1]
  }
  if(method == 'MEPV'){
    temp1 <- MEPV(tau.est, a[-it], b[-it], c[-it], fx, fxi, th.mu,
                  tau.mu, th.var, tau.var, covar, -4, 4, n.quad2)
    fx <- temp1$fx
    it[j] <- (1:n.items)[-it][temp1$choice + 1]
  }
  if(method == 'MEPVT'){
    temp1 <- PV_GMICT(tau.est, a[-it], b[-it], c[-it], A[-it], B[-it],
                      fx, fxi, v.ind, w.ind, th.mu, tau.mu, th.var,
                      tau.var, covar, -4, 4, n.quad2)
    fx <- temp1$fx
    it[j] <- (1:n.items)[-it][temp1$choice + 1]
```

```
}
if(method == 'KLI'){
  #quads <- ceiling(n.quad1 * sqrt(2 / j))
  delta <- width/sqrt(j)
  it[j] <- (1:n.items)[-it][KLI(a[-it], b[-it], c[-it], th.est, delta,
                                n.quad1) + 1]
}
if(method == 'KLIT'){
  #quads <- ceiling(n.quad1 * sqrt(2 / j))
  delta <- width/sqrt(j)
  it[j] <- (1:n.items)[-it][KL_GMICT(a[-it], b[-it], c[-it], A[-it],
                                     B[-it], th.est, tau.est, delta,
                                     v.ind, w.ind, n.quad1) + 1]
}
ec.MIC[it[j]] <- ec.MIC[it[j]] + 1
u[j] <- responses[i, it[j]]
if(method == 'MEPV' | method == 'MEPVT'){
  if(u[j] == 0){
    fxi <- temp1$fxi0
  }else{
    fxi <- temp1$fxi1
  }
}
lnt[j] <- log.times[i, it[j]]
if(MAP == T){
  temp <- MAP_th_tau(u[1:j], lnt[1:j], a[it[1:j]], b[it[1:j]],
                     c[it[1:j]], A[it[1:j]], B[it[1:j]], th.mu,
                     tau.mu, th.var, tau.var, covar)
  th.est <- temp[1]
```

```
      tau.est <- temp[2]
    }else{
      th.est <- MLE_EAP(a[it[1:j]], b[it[1:j]], c[it[1:j]], u[1:j],
                        th.mu, th.var)
      tau.est <- MLE_tau(A[it[1:j]], B[it[1:j]], lnt[1:j])
    }
  }
  theta.MIC[i] <- th.est
  tau.MIC[i] <- tau.est
  test.time[i] <- sum(exp(lnt))
}
if(weighted){
  wvarthe1 <- sum(dnorm(theta, th.mu, sqrt(th.var)) *
                  (theta.MIC - mean(theta.MIC))^2) /
              sum(dnorm(theta, th.mu, sqrt(th.var)))
  wvarthe2 <- sum(dnorm(theta, th.mu, sqrt(th.var)) *
                  (theta - mean(theta))^2) /
              sum(dnorm(theta, th.mu, sqrt(th.var)))
  wcovthe <- sum(dnorm(theta, th.mu, sqrt(th.var)) *
                 (theta.MIC - mean(theta.MIC)) * (theta - mean(theta))) /
             sum(dnorm(theta, th.mu, sqrt(th.var)))
  wvartau1 <- sum(dnorm(tau, tau.mu, sqrt(tau.var)) *
                  (tau.MIC - mean(tau.MIC))^2) /
              sum(dnorm(tau, tau.mu, sqrt(tau.var)))
  wvartau2 <- sum(dnorm(tau, tau.mu, sqrt(tau.var)) *
                  (tau - mean(tau))^2) /
              sum(dnorm(tau, tau.mu, sqrt(tau.var)))
  wcovtau <- sum(dnorm(tau, tau.mu, sqrt(tau.var)) *
                 (tau.MIC - mean(tau.MIC)) * (tau - mean(tau))) /
```

```
            sum(dnorm(tau, tau.mu, sqrt(tau.var)))

  MSE.the <- sum(dnorm(theta, th.mu, sqrt(th.var)) *
                (theta.MIC - theta)^2) /
            sum(dnorm(theta, th.mu, sqrt(th.var)))
  MSE.tau <- sum(dnorm(tau, tau.mu, sqrt(tau.var)) *
                (tau.MIC - tau)^2) /
            sum(dnorm(tau, tau.mu, sqrt(tau.var)))
  bias.the <- sum(dnorm(theta, th.mu, sqrt(th.var)) *
                 (theta.MIC - theta)) /
             sum(dnorm(theta, th.mu, sqrt(th.var)))
  bias.tau <- sum(dnorm(tau, tau.mu, sqrt(tau.var)) *
                 (tau.MIC - tau)) /
             sum(dnorm(tau, tau.mu, sqrt(tau.var)))
  cor.the <- wcovthe / sqrt(wvarthe1 * wvarthe2)
  cor.tau <- wcovtau / sqrt(wvartau1 * wvartau2)
}else{
  MSE.the <- sum((theta.MIC - theta)^2) / n.persons
  MSE.tau <- sum((tau.MIC - tau)^2) / n.persons
  bias.the <- sum((theta.MIC - theta)) / n.persons
  bias.tau <- sum((tau.MIC - tau)) / n.persons
  cor.the <- cor(theta.MIC, theta)
  cor.tau <- cor(tau.MIC, tau)
}
if(conditional){
  uni.the <- unique(theta)
  cond.bias.the <- numeric(length(uni.the))
  cond.mse.the <- numeric(length(uni.the))
  for(i in 1:length(uni.the)){
```

```
    temp.the <- theta[theta == uni.the[i]]
    temp.est <- theta.MIC[theta == uni.the[i]]
    temp.n <- length(temp.the)
    cond.bias.the[i] <- sum(temp.est - temp.the) / temp.n
    cond.mse.the[i] <- sum((temp.est - temp.the)^2) / temp.n
  }
  if(any(method == c('MIC', 'MICT', 'MEPV', 'KLI'))){
    results2[[q]] <- data.frame(theta = uni.the, cond.bias = cond.bias.the,
                                cond.mse = cond.mse.the)
  }
  if(any(method == c('GMICT', 'MEPVT', 'KLIT'))){
    results2[[q]] <- data.frame(w = w.ind, v = v.ind, theta = uni.the,
                                cond.bias = cond.bias.the,
                                cond.mse = cond.mse.the)
  }
}

MTT <- mean(test.time)
STT <- sd(test.time)
Chi2 <- sum((ec.MIC/n.persons - test.length/n.items)^2) /
        (test.length/n.items)
if(any(method == c('MIC', 'MICT', 'MEPV', 'KLI'))){
  results[q,] <- c(MSE.the, MSE.tau, bias.the, bias.tau, cor.the, cor.tau,
                   MTT, STT, Chi2)
}
if(any(method == c('GMICT', 'MEPVT', 'KLIT'))){
  results[q,] <- c(w.ind, v.ind, MSE.the, MSE.tau, bias.the, bias.tau,
                   cor.the, cor.tau, MTT, STT, Chi2)
}
```

```
    }
  }

  if(conditional){
    output = list(results = results, results2 = results2)
  }else{
    output = list(results = results)
  }

  return(output)
}
```

# References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement*, *12*, 33–51.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, *49*, 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*, 253–276.

Attali, Y. (2004). *Reliability of speeded number-right multiple-choice tests* (Research Report No. RR-04-15). Princeton, NJ: Educational Testing Service.

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

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems* (Series Report No. No. 58-16). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability* (Tech. Rep. No. No. 17). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1958b). *On the estimation of mental ability* (Series Report No. No. 15). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21–33.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika*, *46*, 443–456.

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

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

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

Chang, H.-H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika*, *61*, 445–463.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333-353.

Chang, H.-H., Qian, J., & Ying, Z. (2001). $a$-Stratified multistage computerized adaptive testing with $b$ blocking. *Applied Psychological Measurement*, *25*, 333–341.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

Chang, H.-H., & Ying, Z. (1999). $a$-Stratified multistage computerized adaptive testing. *Applied Psycholo-*

*gical Measurement*, *23*, 211–222.
Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*, 1466–1488.
Chang, H.-H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, *67*, 387–398.
Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369–383.
Choe, E. M., & Kern, J. L. (2014, April). *Controlling item exposure for response time-informed item selection in CAT.* (Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA)
Coombs, C. H. (1964). *A theory of data.* New York, NY: Wiley.
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart, and Winston.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.
Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A programmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale, NJ: Lawrence Erlbaum Associates.
Douglas, J., Kosorok, M., & Chewning, B. (1999). A latent variable model for multivariate psychometric response times. *Psychometrika*, *64*, 69–82.
Drasgow, F., Chernyshenko, O. O., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*, 465–476.
Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 577–636). Palo Alto, CA: Consulting Psychologists Press.
Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., & Hulin, C. L. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army selection and classification decisions* (Tech. Rep. No. 1311). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380–396.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.
Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. A. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655–670.
Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.
Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York, NY: Springer.
Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of response and response times with the package cirt. *Journal of Statistical Software*, *20*, 1–14.
Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), *The handbook of abnormal psychology.* London, UK: Pitman Medical.
Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, *56*, 1–104.
Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.
Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351–373.
Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage Publications.

Hoijtinck, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika*, *55*, 641–656.
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
Jansen, M. G. H. (1986). A Bayesian version of Rasch's multiplicative Poisson model for the number of errors on achievement tests. *Journal of Educational Statistics*, *11*, 147–160.
Jansen, M. G. H. (1997a). Rasch model for speed tests and some extensions with applications to incomplete designs. *Journal of Educational and Behavioral Stastistics*, *22*, 125–140.
Jansen, M. G. H. (1997b). Rasch's model for reading speed with manifest exploratory variables. *Psychometrika*, *62*, 393–409.
Jansen, M. G. H., & van Duijn, M. A. J. (1992). Extensions of Rasch's multiplicative Poisson model. *Psychometrika*, *57*, 405–414.
Kennedy, M. (1930). Speed as a personality trait. *Journal of Social Psychology*, *1*, 286–298.
Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.
Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: Modeling approach using responses and response times. *Psychological Methods*, *14*, 54–75.
Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457–477.
Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement*, *12*, 365–376.
Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.
Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517–548.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233–245.
Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*, 347–364.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York, NY: Oxford University Press.
Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement*, *46*, 3228–3237.
Maris, E. (1993). Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.
Maydeu-Olivares, A., Hernández, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research*, *41*, 445–471.
Meyer, J. P. (2008). *A mixture Rasch model with item response-time components.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
Mislevy, R. J., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika*, *65*, 149–156.
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*, 197–219.
Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296.
Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 77–101). New York, NY: Springer.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report No. No. RR-69-92). Princeton, NJ: Educational Testing Service.

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

Pieters, J. P. M., & van der Ven, A. H. G. S. (1982). Precision, speed, and distraction in time-limit tests. *Applied Psychological Measurement*, *6*, 93–109.

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, *78*, 538–544.

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

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

Roberts, J. S. (1995). *Item response theory approaches to attitude measurement* (Unpublished doctoral dissertation). University of South Carolina, Columbia, SC.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1998). *The generalized graded unfolding model: A general parametric item response model for unfolding graded responses* (Research Report No. No. RR-98-32). Princeton, NJ: Educational Testing Service.

Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement*, *25*, 177–192.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam: North-Holland.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). Springer-Verlag.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Roussos, L., & Stout, W. (1996). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215–230.

Sahin, S. G., Walker, C. M., & Gelbal, S. (2014). The impact of model misspecification with multidimensional test data. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research: The 79th annual meeting of the psychometric society, madison, wisconsin, 2014* (pp. 145–172). New York, NY: Springer.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*, 18–38.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issures of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Lawrence Erlbaum Associates.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Sorensen, D., & Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in quantitative genetics.* New York, NY: Springer.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* London, UK: Macmillan.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The generalized graded unfolding model for multi-unidimensional paired comparison responses* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Sternberg, R. J. (1999). *The nature of cognition.* Cambridge, MA: MIT Press.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57–75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure control of items in CAT. In

W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston, MA: Kluwer-Nijhof Publishing.

Sun, Y. (2015). *Constructing a misspecified item response model that yields a specified estimate and a specified model misfit value* (Unpublished doctoral dissertation). Ohio State University, Columbus, OH.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement*, *8*, 353–374.

Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incoporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference* (pp. 236–256). University of Minnesota, Department of Psychology, Psychometric Methods Program.

ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*, 575–579.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic Press.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Traub, R. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*, 8–14.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201–216.

van der Linden, W. J. (1999). Multidimensional adative testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 387–308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J. (2009a). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*, 378–394.

van der Linden, W. J. (2009b). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272.

van der Linden, W. J. (2009c). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25–41.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

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

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

Veldkamp, B. P. (2016). On the issue of item selection in computerized adaptive testing with response times.

*Journal of Educational Measurement*, *53*, 212–228.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. (1997). A logistic model fot time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York, NY: Springer-Verlag.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*, 428–449.

Wang, C., Chang, H.-H., & Boughton, K. (2011). Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*, *76*, 13–39.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144–168.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*, 381–417.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wang, W.-C., Liu, C.-W., & Wu, S.-L. (2013). The random-threshold generalized unfolding model and its application of computerized adaptive testing. *Applied Psychological Measurement*, *37*, 179–200.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *4*, 17–27.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41–58.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Woodworth, R. S. (1899). Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements*, *3*, 1–114.

Yi, Q., Zhang, J., & Chang, H.-H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement*, *32*, 543–558.

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*, 1–11.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 331–354). New York, NY: Springer.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–251.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics*, *25*, 225–247.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computer-adaptive tests. *Applied Psychological Measurement*, *18*, 121–140.