LSAC RESEARCH REPORT SERIES

Optimizing Information Using the Expectation-Maximization Algorithm in Item Response Theory

Alexander Weissman

Law School Admission Council
Research Report 11-01
March 2011

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation that provides unique, state-of-the-art admission products and services to ease the admission process for law schools and their applicants worldwide. More than 200 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC's services.

LSAC fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, Credential Assembly Service (CAS), and other matters may change without notice at any time. Up-to-date LSAC policies and procedures are available at LSAC.org, or you may contact our candidate service representatives.

Table of Contents

Executive Summary

The Law School Admission Test (LSAT) uses a mathematical model called item response theory (IRT) to ensure score and test-form comparability from administration to administration. In order to apply IRT to test data, a set of parameters for each administered question (item) is estimated from the data using a statistical method called marginal maximum likelihood (MML). The expectation-maximization (EM) algorithm underlies the efficient MML estimation of item parameters and is employed operationally for each LSAT administration.

This report first presents a mathematical development and overview of the EM algorithm as it is applied to the three-parameter logistic (3PL) IRT model, the same model used for estimating item parameters for the LSAT. It is then shown that estimating the 3PL IRT model is a special case of a more general, or unconstrained, estimation that applies to almost all IRT models that use dichotomously scored items. Next, the relationship between the EM algorithm and information theory, a mathematical theory of communication that sets limits on the amount of information that can be extracted from data, is examined. The equivalence of MML estimation via the EM algorithm to the minimization of the Kullback–Leibler divergence, a fundamental quantity in information theory, is then illustrated.

Using mathematical results from information theory, it is shown that the unconstrained estimation proposed here provides a fixed reference point against which many other models may be tested—in a sense, an overarching model to test models. A likelihood ratio test developed for testing models against this reference point is provided, and examples using both real and simulated data demonstrate the approach.

Introduction

A major challenge in educational and psychological measurement, broadly known by the term *testing*, is quantifying attributes of a person that cannot be observed directly. Examples of such attributes include skill, proficiency, and ability. Further complicating the process are the ever-present sources of error, due in part to the measurement instruments (tests) but also to the response behavior of subjects (test takers). Thus,

educational and psychological measurement is concerned with measuring unobservable constructs in the presence of non-negligible error. As a result, statistics and statistical models play a crucial role in measurement.

One of the earliest of these models may be traced as far back as Spearman (1907, 1913) and is usually referred to as the *classical true-score model* (Crocker & Algina, 1986). In this model, an observed score (e.g., a test score) is partitioned into two components: the *true score* and the *error score*. The true score is defined as the expectation of the observed score; that is, $\tau_i = E[X_i]$, where τ_i is the true score and X_i is the observed score for an individual i. The error score, Z_i, is then equal to $X_i - \tau_i$ (Crocker & Algina, 1986, pp. 106–107; Lord & Novick, 1968).

A few other assumptions of the classical true-score model give it some important statistical properties:

1. $E[Z_i] = 0$; that is, the expectation of the error score is zero.

2. For a set of individuals i and measurement occasions[1] r, the true scores $\tau_i = E[\tau_{ir}]$ and error scores Z_{ir} are uncorrelated; that is, $\text{cov}(\tau_i, Z_{ir}) = 0$.

3. Error scores across measurement occasions are uncorrelated; that is, for any two measurement occasions, say $r = 1$ and $r = 2$, $\text{cov}(Z_{i1}, Z_{i2}) = 0$.

In the classical true-score model, there are no explicit assumptions about the individual items comprising a test. The observed score is usually computed as a linear combination of item scores; commonly, the number correct or sum score suffices as the observed score. However, the items themselves do play an implicit role in the classical true-score model. For instance, while the classical definition of test-score reliability $\rho = \sigma_T^2 / \sigma_X^2$, where σ_T^2 is the variance of true scores and σ_X^2 is the variance of observed scores, can be stated without reference to items, the practical estimation of it relies on subsets of items, or on the individual items themselves. Furthermore, from the classical true-score model, test construction depends on item-level properties (e.g., item

[1]Strictly speaking, these are measurement occasions over what are referred to as "parallel forms."

difficulty) that are bound by the specific instrument (test) from which they came, as well as the specific sample (test takers) from which they were obtained.

Item response theory (IRT) was developed as a comprehensive latent trait theory for removing (or at least controlling for) the test dependence and sample dependence of item properties. The latent trait approach assumes that underlying a test taker's response behavior is an unobservable construct: the *latent trait*. Defining the nature of a latent trait has been, and continues to be, somewhat of a challenge. In their chapter introducing latent traits and item response functions, Lord & Novick (1968) write:

> Much of psychological theory is based on a trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behavior. (p. 385)

While the interpretation of a latent trait is open to debate, it is fortunate that the mathematical treatment of such is fairly well defined. We speak of an item response function as that which relates the latent trait level for a test taker i, symbolized by θ_i, to the probability of observing a specific response on item j, symbolized by X_{ij}. In the following, unless otherwise noted, we will assume that item responses are either correct $\left(X_{ij}=1\right)$ or incorrect $\left(X_{ij}=0\right)$, that $\theta_i \in \mathbb{R}$, and that the relationship between θ_i and the probability of a correct response can be modeled according to a three-parameter logistic (3PL) model (Lord & Novick, 1968). That is,

$$P\left(X_{ij}=1\middle|\theta_i\right)=c_j+\frac{\left(1-c_j\right)}{1+\exp\left[-Da_j\left(\theta_i-b_j\right)\right]}, \tag{1}$$

where

$X_{ij}=1$ is a realization of a correct response by test taker i to item j,

θ_i is the value of the latent trait for test taker i,

a_j is an item discrimination parameter ($a_j > 0$), quantifying the rate of change in $P\left(X_{ij} = 1 \middle| \theta\right)$ with respect to θ,

b_j is an item difficulty parameter ($b_j \in \mathbb{R}$), measured on the same scale as θ,

c_j is an item pseudo-guessing parameter ($0 \le c_j < 1$), which estimates the probability of correctly responding to item j by chance.

By setting $D = 1.702$, the logistic function in (1) can be scaled to match closely with the normal ogive model. The above restrictions on the parameters a_j and c_j ensure that $P\left(X_{ij} = 1 \middle| \theta\right)$ is monotonically increasing in θ. An example of an item response function is shown in Figure 1.

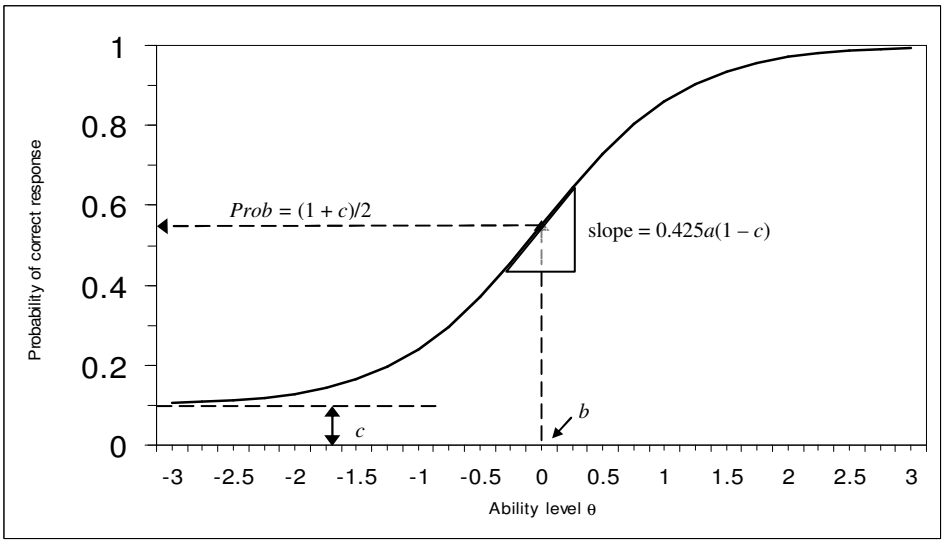

FIGURE 1. *An example of an item response function*

To simplify notation, we will refer to the 3PL IRT model item parameter vector for an item j as $\boldsymbol{\omega}_j = \left[a_j \quad b_j \quad c_j\right]^T$.

Problem statement. In practice, the item parameters and latent trait values in (1) are unknown, so they must be estimated from the observed pattern of item responses (i.e., a matrix $\mathbf{X} = \left[X_{ij} \right]$). Thus, a general formulation of the problem is given by the following:

$$\max_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \ln L\left(\mathbf{X} | \boldsymbol{\omega}\right), \tag{2}$$

where $\ln L\left(\mathbf{X} | \boldsymbol{\omega}\right)$ is the logarithm of the likelihood of observing $\mathbf{X}$ given a set of parameters $\boldsymbol{\omega}$, $\mathbf{X}$ is an $N \times J$ matrix of item responses for $i = 1,2,\ldots,N$ test takers responding to $j = 1,2,\ldots,J$ items, and $\boldsymbol{\omega}$ is a $P \times J$ matrix of P item parameters for J items. This problem, known in statistics as the *maximum likelihood problem* (Bain & Engelhardt, 1992), is to find a set of parameters $\boldsymbol{\omega}$, contained in the parameter space $\boldsymbol{\Omega}$, that maximizes the log likelihood function.

The problem in (2) may also be viewed as an optimization problem. As Wets (1999) notes, “Finding the ‘best’ estimate of a statistical parameter is, of course, an optimization problem of some type” (p. 79). Birnbaum’s joint maximum likelihood estimation, one of the first estimation methods proposed for solving (2) (Lord & Novick, 1968), yielded inconsistent parameter estimates for the 3PL model as sample size increased. The method of marginal maximum likelihood (MML), proposed by Bock and Lieberman (1970), yielded consistent parameter estimates, but the computational intensity of their procedure was such that it could only be applied to very short tests.

The breakthrough to solving (2) came when Dempster, Laird, and Rubin (1977) formalized the expectation-maximization (EM) algorithm. Bock & Aitkin (1981) later applied the EM algorithm to MML estimation of item parameters, yielding a procedure that was “both theoretically acceptable and computationally feasible” (Baker & Kim, 2004; Harwell, Baker, & Zwarts, 1988).

In the following we derive the estimating equations for a 3PL IRT model using the EM algorithm, but from a slightly different perspective. We begin with the idea of the EM algorithm as a lower-bound optimization (Beal, 2003; Dellaert, 2002; Harpaz & Haralick, 2006; Minka, 1998). We will examine this bound-optimization approach again below, in the section titled Optimizing Bounds on the Log Likelihood Function.

Our first step is to express the log likelihood function in (2) as a marginal likelihood over a (possibly multidimensional) latent variable $\boldsymbol{\theta}$:

$$\ln L(\mathbf{X}|\boldsymbol{\omega}) = \ln \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\omega}). \tag{3}$$

There are two things to note in (3). First, we have a logarithm of a sum, which is difficult to manipulate. Second, we have augmented our function with more unknowns, namely the latent variable $\boldsymbol{\theta}$. Surprisingly, what looks like a step in the wrong direction turns out to be a serendipitous maneuver, thanks to one of the most important inequalities in convex analysis: *Jensen's inequality* (e.g., Cover & Thomas, 2006, p. 27).

Theorem 1: *Jensen's inequality.* If f is a convex function and Y is a random variable,

$$Ef(Y) \geq f(EY), \tag{4}$$

where E is the expectation operator. Note that the logarithm function is concave, not convex, but we can still use Jensen's inequality because the negative of a concave function is a convex function. Thus, if f is a concave function and Y is a random variable, $Ef(Y) \leq f(EY)$.

Applying Jensen's inequality for concave functions to (3) allows us to move the logarithm inside the summation while simultaneously lower-bounding the log likelihood function. To do so, however, we must introduce a density $g(\boldsymbol{\theta})$ over which the expectation is taken:

$$\begin{aligned} \ln L(\mathbf{X}|\boldsymbol{\omega}) &= \ln \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\omega}) \\ &= \ln \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g(\boldsymbol{\theta}) \frac{L(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\omega})}{g(\boldsymbol{\theta})} \\ &\geq \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g(\boldsymbol{\theta}) \ln \frac{L(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\omega})}{g(\boldsymbol{\theta})}. \end{aligned} \tag{5}$$

The last line in (5) provides a lower bound on the log likelihood function by means of an auxiliary density $g(\boldsymbol{\theta})$. While a lower bound is good, a tight bound (i.e., equality) is optimal.

Fortunately, we can in fact provide a tight bound. By choosing the density $g(\boldsymbol{\theta})$ as a posterior density, we find that

$$\begin{aligned} \ln L(\mathbf{X}|\boldsymbol{\omega}) &= \ln \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\mathbf{X},\boldsymbol{\theta}|\boldsymbol{\omega}) \\ &= \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\omega}) \ln \frac{L(\mathbf{X},\boldsymbol{\theta}|\boldsymbol{\omega})}{g(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\omega})}, \end{aligned} \tag{6}$$

where, by Bayes' theorem, the posterior density is given by

$$g(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\omega}) = \frac{L(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\omega})\, p(\boldsymbol{\theta},\boldsymbol{\omega})}{\sum_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} L(\mathbf{X}|\boldsymbol{\theta}',\boldsymbol{\omega})\, p(\boldsymbol{\theta}',\boldsymbol{\omega})} \tag{7}$$

and $p(\boldsymbol{\theta},\boldsymbol{\omega})$ is a prior density. Substituting (7) into (5) will show the equality; Dellaert (2002) and Beal (2003) use Lagrange multipliers and functional derivatives to show that (7) is an optimal bound for (5).

We can rewrite (6) as a difference between two components and also introduce some additional notation:

$$\begin{aligned} \ln L(\mathbf{X}|\boldsymbol{\omega}) = &\underbrace{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\omega}^{(t)}) \ln L(\mathbf{X},\boldsymbol{\theta}|\boldsymbol{\omega})}_{Q(\boldsymbol{\omega}|\boldsymbol{\omega}^{(t)})} \\ &- \underbrace{\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\omega}^{(t)}) \ln g(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\omega})}_{H(\boldsymbol{\omega}|\boldsymbol{\omega}^{(t)})}. \end{aligned} \tag{8}$$

The EM algorithm is iterative, so to keep track of the iterations, we let t indicate the iteration number, $t = 0,1,2,\ldots$. On the right-hand side of (8), we have two terms: $Q(\boldsymbol{\omega}|\boldsymbol{\omega}^{(t)})$ and $H(\boldsymbol{\omega}|\boldsymbol{\omega}^{(t)})$. These correspond to the notation in Dempster et al. (1977).

The $Q\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)$ relates to the complete data log likelihood; $H\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)$ relates to the posterior density of the latent variable. An application of Jensen's inequality shows that for any $\boldsymbol{\omega}\in\boldsymbol{\Omega}$, $H\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)\le H\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)$ (Dempster et al., 1977, p. 6). The maximization step, or M-step, maximizes $Q\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)$ with each iteration $t\to t+1$, so $Q\left(\boldsymbol{\omega}^{(t+1)}\middle|\boldsymbol{\omega}^{(t)}\right)\ge Q\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)$. Thus, from (8), it can be seen that the marginal log likelihood $\ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}\right)$ is guaranteed not to decrease. That is,

$$\begin{aligned}\ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}^{(t+1)}\right)-\ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}^{(t)}\right) &= Q\left(\boldsymbol{\omega}^{(t+1)}\middle|\boldsymbol{\omega}^{(t)}\right)-H\left(\boldsymbol{\omega}^{(t+1)}\middle|\boldsymbol{\omega}^{(t)}\right)\\ &\quad-\left[Q\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)-H\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)\right]\\ &= \underbrace{\left[Q\left(\boldsymbol{\omega}^{(t+1)}\middle|\boldsymbol{\omega}^{(t)}\right)-Q\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)\right]}_{\ge 0}\\ &\quad+\underbrace{\left[H\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)-H\left(\boldsymbol{\omega}^{(t+1)}\middle|\boldsymbol{\omega}^{(t)}\right)\right]}_{\ge 0}\\ &\ge 0.\end{aligned} \tag{9}$$

The expectation step, or E-step, of the algorithm calculates the posterior density $g\left(\boldsymbol{\theta}\middle|\mathbf{X},\boldsymbol{\omega}^{(t)}\right)$ given the parameters $\boldsymbol{\omega}^{(t)}$ at iteration t. Since the parameters are exactly what we are trying to find, at initialization $(t=0)$ we can start the algorithm with reasonable estimates (guesses) $\boldsymbol{\omega}^{(0)}$. The M-step finds the value of $\boldsymbol{\omega}$ that maximizes $Q\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)$; that is, at iteration $t+1$, it finds $\boldsymbol{\omega}^{(t+1)}$ such that

$$\boldsymbol{\omega}^{(t+1)} = \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}}{\arg\max} \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g\left(\boldsymbol{\theta}\middle|\mathbf{X},\boldsymbol{\omega}^{(t)}\right) \ln L\left(\mathbf{X},\boldsymbol{\theta}\middle|\boldsymbol{\omega}\right). \tag{10}$$

In practice, the E-steps and M-steps alternate until a convergence criterion is met. This criterion can be a difference in parameter estimates $\left|\boldsymbol{\omega}^{(t+1)} - \boldsymbol{\omega}^{(t)}\right| < \varepsilon$ for some $\varepsilon > 0$, or a difference in the marginal log likelihood $\ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}^{(t+1)}\right) - \ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}^{(t)}\right)$, or perhaps a combination of both. Detailed discussions and proofs regarding the convergence of the EM algorithm may be found in Dempster et al. (1977) and Wu (1983).

Parameter Estimation in Item Response Theory

We now start with the EM equation in (6). By applying the specifics of the 3PL IRT estimation problem, we end up with estimating equations. The following assumptions for our IRT estimation problem will simplify matters.

Assumption 1: *Test takers respond independently of one another.* Formally, if $L\left(\mathbf{X}_r\middle|\boldsymbol{\omega}\right)$ is the likelihood of response vector $\mathbf{X}_r$ for test taker r, and $L\left(\mathbf{X}_s\middle|\boldsymbol{\omega}\right)$ is the likelihood of response $\mathbf{X}_s$ for test taker s, then $L\left(\mathbf{X}_r,\mathbf{X}_s\middle|\boldsymbol{\omega}\right) = L\left(\mathbf{X}_r\middle|\boldsymbol{\omega}\right)L\left(\mathbf{X}_s\middle|\boldsymbol{\omega}\right)$ for all $r,s;\ r \neq s$. From this assumption, we can write

$$\ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}\right) = \ln \prod_{i=1}^{N} L\left(\mathbf{X}_i\middle|\boldsymbol{\omega}\right) = \sum_{i=1}^{N} \ln L\left(\mathbf{X}_i\middle|\boldsymbol{\omega}\right), \tag{11}$$

where $\mathbf{X}_i$ is a vector of item responses by test taker i to J items with parameters $\boldsymbol{\omega}$.

Substituting (11) into (6), we have

$$\begin{aligned} \ln L(\mathbf{X}|\boldsymbol{\omega}) &= \sum_{i=1}^{N} \ln L(\mathbf{X}_i|\boldsymbol{\omega}) \\ &= \sum_{i=1}^{N} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g(\boldsymbol{\theta}|\mathbf{X}_i, \boldsymbol{\omega}) \ln \frac{L(\mathbf{X}_i, \boldsymbol{\theta}|\boldsymbol{\omega})}{g(\boldsymbol{\theta}|\mathbf{X}_i, \boldsymbol{\omega})}. \end{aligned} \tag{12}$$

Assumption 2: *All item responses are conditionally independent given* $\boldsymbol{\theta}$. This assumption is also called *local independence* (Lord & Novick, 1968, p. 398). That is,

$$\begin{aligned} L(\mathbf{X}_i, \boldsymbol{\theta}|\boldsymbol{\omega}) &= L(\mathbf{X}_i|\boldsymbol{\theta}, \boldsymbol{\omega}) p(\boldsymbol{\theta}|\boldsymbol{\omega}) \\ &= \prod_{j=1}^{J} L(X_{ij}|\boldsymbol{\theta}, \boldsymbol{\omega}_j) p(\boldsymbol{\theta}|\boldsymbol{\omega}_j), \end{aligned} \tag{13}$$

where the conditional independence of the J item responses implies that the joint likelihood $L(\mathbf{X}_i, \boldsymbol{\theta}|\boldsymbol{\omega})$ may be factored into a product of likelihood functions.

Assumption 3: *There is independence between* $\boldsymbol{\theta}$ *and* $\boldsymbol{\omega}$. We assume that the latent trait $\boldsymbol{\theta}$ is independent of the item parameters $\boldsymbol{\omega}$; thus, $p(\boldsymbol{\theta}, \boldsymbol{\omega}) = p(\boldsymbol{\theta}) p(\boldsymbol{\omega})$. Hence, (13) simplifies to

$$L(\mathbf{X}_i, \boldsymbol{\theta}|\boldsymbol{\omega}) = \prod_{j=1}^{J} L(X_{ij}|\boldsymbol{\theta}, \boldsymbol{\omega}_j) p(\boldsymbol{\theta}). \tag{14}$$

Assumption 4: *The latent trait* $\boldsymbol{\theta}$ *is unidimensional*. The unidimensionality assumption reduces $\boldsymbol{\theta}$ to a vector, where we denote θ_k as the k^{th} element of this vector. This further simplifies the likelihood function, and it allows us to replace a summation over a set of elements with a summation over a single index k. These θ_k are called *quadrature points*. They are the support for the latent trait, which in the marginal maximum likelihood procedure is treated as a random effect.

Putting the above assumptions together, we obtain

$$\begin{aligned}
\ln L\left(\mathbf{X}\middle|\boldsymbol{\omega}\right) &= \sum_{i=1}^{N}\ln L\left(\mathbf{X}_i\middle|\boldsymbol{\omega}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right)\ln\frac{L\left(\mathbf{X}_i,\theta_k\middle|\boldsymbol{\omega}\right)}{g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right)} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right)\ln\frac{\prod_{j=1}^{J} L\left(X_{ij}\middle|\theta_k,\boldsymbol{\omega}_j\right)p(\theta_k)}{g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right)} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right)\left[\sum_{j=1}^{J}\ln L\left(X_{ij}\middle|\theta_k,\boldsymbol{\omega}_j\right)+\ln p(\theta_k)\right] \\
&\quad -\sum_{i=1}^{N}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right)\ln g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right).
\end{aligned} \tag{15}$$

Note that the last equality is analogous to the decomposition of $Q\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)$ and $H\left(\boldsymbol{\omega}\middle|\boldsymbol{\omega}^{(t)}\right)$ terms in (8). In addition, with these assumptions we obtain a simplified expression for the posterior density:

$$g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}\right) = \frac{\prod_{j=1}^{J} L\left(X_{ij}\middle|\theta_k,\boldsymbol{\omega}_j\right)p(\theta_k)}{\sum_{m=1}^{K}\prod_{j=1}^{J} L\left(X_{ij}\middle|\theta_m,\boldsymbol{\omega}_j\right)p(\theta_m)}. \tag{16}$$

The M-step, given in a more general form in (10), can be written here as

$$\begin{aligned}
\boldsymbol{\omega}^{(t+1)} &= \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}}{\arg\max}\sum_{i=1}^{N}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}^{(t)}\right)\ln L\left(\mathbf{X}_i,\theta_k\middle|\boldsymbol{\omega}\right) \\
&= \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}}{\arg\max}\sum_{i=1}^{N}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\omega}^{(t)}\right)\left[\sum_{j=1}^{J}\ln L\left(X_{ij}\middle|\theta_k,\boldsymbol{\omega}_j\right)+\ln p(\theta_k)\right].
\end{aligned} \tag{17}$$

We have just a few more steps before arriving at estimating equations that are useful for computation. First, we add one more assumption to our IRT parameter estimation problem.

Assumption 5: *Item responses follow a binomial sampling model*. For a dichotomously scored item response $X_{ij} = \{0,1\}$, where $X_{ij} = 1$ indicates that test taker i correctly responded to item j, the likelihood function is chosen according to the binomial sampling model. Thus, if we denote the success probability $P_{\boldsymbol{\omega}_j}(\theta_k) \equiv P\left(X_j = 1 \middle| \theta_k, \boldsymbol{\omega}_j\right)$, the log likelihood function is

$$\begin{aligned} \ln L\left(X_{ij} = x_{ij} \middle| \theta_k, \boldsymbol{\omega}_j\right) &= \ln\left[P_{\boldsymbol{\omega}_j}(\theta_k)^{x_{ij}}\left(1 - P_{\boldsymbol{\omega}_j}(\theta_k)\right)^{1-x_{ij}}\right] \\ &= x_{ij}\ln P_{\boldsymbol{\omega}_j}(\theta_k) + \left(1 - x_{ij}\right)\ln\left(1 - P_{\boldsymbol{\omega}_j}(\theta_k)\right). \end{aligned} \tag{18}$$

Substituting (18) into (17), and noting that the $\ln p(\theta_k)$ term in (17) can be omitted since it is not a function of $\boldsymbol{\omega}$, we obtain

$$\boldsymbol{\omega}^{t+1} = \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}}{\arg\max} \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right)\left[x_{ij}\ln P_{\boldsymbol{\omega}_j}(\theta_k) + \left(1 - x_{ij}\right)\ln\left(1 - P_{\boldsymbol{\omega}_j}(\theta_k)\right)\right]. \tag{19}$$

Equation (19) is sufficient for finding maximum likelihood estimates of $\boldsymbol{\omega}$.

Following Harwell et al. (1988), we can give additional interpretation to the terms in (19) by focusing on the inner summation (over test takers i) for fixed j,k. Starting with the first term in (19), we note that $\ln P_{\boldsymbol{\omega}_j}(\theta_k)$ can be factored out of the summation, as it is constant for all i given θ_k. Thus,

$$\sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right) x_{ij} \ln P_{\boldsymbol{\omega}_j}(\theta_k) = \ln P_{\boldsymbol{\omega}_j}(\theta_k) \sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right) x_{ij}. \tag{20}$$

Note that only correct responses $x_{ij} = 1$ will contribute nonzero values to this summation. Thus, we can identify the expected number of correct responses to item j at quadrature point θ_k as

$$\hat{r}_{jk} = \sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right) x_{ij}. \tag{21}$$

Likewise, for the second term in (19), the expected number of incorrect responses to item j at θ_k is

$$\hat{w}_{jk} = \sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right)\left(1 - x_{ij}\right). \tag{22}$$

The sum of (21) and (22) must equal the expected number of test takers[2] responding to item j at θ_k; thus,

$$\hat{n}_{jk} = \sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right). \tag{23}$$

Since $\hat{w}_{jk} = \hat{n}_{jk} - \hat{r}_{jk}$, we can rewrite (22) as

$$\hat{n}_{jk} - \hat{r}_{jk} = \sum_{i=1}^{N} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\omega}^{(t)}\right)\left(1 - x_{ij}\right). \tag{24}$$

Using these identities and recalling that $P_{\boldsymbol{\omega}_j}\left(\theta_k\right) \equiv P\left(X_j = 1 \middle| \theta_k, \boldsymbol{\omega}_j\right)$, we can simplify (19) as

$$\boldsymbol{\omega}^{t+1} = \underset{\boldsymbol{\omega} \in \boldsymbol{\Omega}}{\arg\max} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{r}_{jk} \ln P_{\boldsymbol{\omega}_j}\left(\theta_k\right) + \left(\hat{n}_{jk} - \hat{r}_{jk}\right) \ln\left(1 - P_{\boldsymbol{\omega}_j}\left(\theta_k\right)\right). \tag{25}$$

Estimation of item parameters can be done one item at a time (as suggested by the outer summation over j in the above equation). Note that (25) shows us that the $\hat{r}_{jk}$ and $\hat{n}_{jk}$ are the sufficient statistics for estimating $\boldsymbol{\omega}_j$ (Harwell et al., 1988, pp. 257–258).

[2]The condition of missing responses (e.g., omitting an item) is not covered here.

Optimizing Bounds on the Log Likelihood Function

Earlier we saw that the E-step of the EM algorithm finds at iteration t the optimal lower bound for the marginal log likelihood function (see (6) and (7)). The M-step is also an optimization: It optimizes this lower bound by finding a set of parameters $\boldsymbol{\omega}^{(t+1)}$ such that $Q\left(\boldsymbol{\omega}^{(t+1)}\middle|\boldsymbol{\omega}^{(t)}\right) \geq Q\left(\boldsymbol{\omega}^{(t)}\middle|\boldsymbol{\omega}^{(t)}\right)$. for all valid $\boldsymbol{\omega} \in \Omega$. Since the summation in (25) is differentiable with respect to $\boldsymbol{\omega}$, the usual route to finding $\boldsymbol{\omega}^{(t+1)}$ is achieved by differentiating the function, setting this derivative to zero, and solving. As mentioned, estimating $\boldsymbol{\omega}^{(t+1)}$ can be done one item at a time; thus, for an item j, our system of first-order derivatives is

$$\frac{\partial}{\partial \boldsymbol{\omega}_j} \sum_{k=1}^{K} \hat{r}_{jk} \ln P_{\boldsymbol{\omega}_j}(\theta_k) + \left(\hat{n}_{jk} - \hat{r}_{jk}\right) \ln\left(1 - P_{\boldsymbol{\omega}_j}(\theta_k)\right) = 0, \tag{26}$$

which simplifies to

$$\sum_{k=1}^{K} \frac{\hat{r}_{jk} - \hat{n}_{jk} P_{\boldsymbol{\omega}_j}(\theta_k)}{P_{\boldsymbol{\omega}_j}(\theta_k)\left[1 - P_{\boldsymbol{\omega}_j}(\theta_k)\right]} \frac{\partial P_{\boldsymbol{\omega}_j}(\theta_k)}{\partial \boldsymbol{\omega}_j} = 0. \tag{27}$$

At this point, one may wonder whether a solution to (27) is the best we can do. More precisely, although we clearly are solving for an optimum, it should be realized that we are actually solving for a *constrained* optimum. The constraint is imposed by the item response function in (1), as parameterized by $\boldsymbol{\omega}_j$. Note that this constraint is enforced by the local behavior of the log likelihood function with respect to the parameters $\boldsymbol{\omega}$. Thus, while we are optimizing the bound on the log likelihood in the M-step, we must do so under the constraints imposed by the item response function $P_{\boldsymbol{\omega}_j}(\theta_k)$.

Is it possible, then, to free this constraint? If so, what would we be estimating? That is, since the item response function in (1) would no longer apply, there would be no parameters $\boldsymbol{\omega}$ to estimate. While it is true that the parametric constraints imposed by (1) would be removed, all IRT assumptions we made in Assumptions 1–5 would still hold. For instance, the success probability $P\left(X_j = 1 \middle| \theta_k\right)$ would remain conditional on the

latent trait θ_k, although we would no longer enforce a functional form for $P\left(X_j = 1 \middle| \theta_k\right)$. Further, the binomial sampling assumption (Assumption 5), which underlies almost all (dichotomous) IRT models, would remain. Below we will see that this assumption is particularly important (see section titled An Information-Theoretic Interpretation).

Pursuing an unconstrained optimization for the M-step closely follows the method used for lower-bounding the marginal log likelihood function in the E-step. In that case, the optimum density (see $g(\boldsymbol{\theta})$ in (5)–(7)) is found by considering not a function, but a functional, taking functional derivatives and employing Lagrange multipliers as needed. We extend the functional derivative approach taken by Dellaert (2002) and Beal (2003), this time focusing on the M-step (see (10) as well as the discussion of the $Q\left(\boldsymbol{\omega} \middle| \boldsymbol{\omega}^{(t)}\right)$ component).

Recalling the estimating equation we obtained in (25), that is,

$$\boldsymbol{\omega}^{(t+1)} = \underset{\boldsymbol{\omega} \in \boldsymbol{\Omega}}{\arg\max} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{r}_{jk} \ln P_{\boldsymbol{\omega}_j}(\theta_k) + \left(\hat{n}_{jk} - \hat{r}_{jk}\right) \ln\left(1 - P_{\boldsymbol{\omega}_j}(\theta_k)\right),$$

we note that this maximization is over the parameters $\boldsymbol{\omega}$. Now let us consider a more general problem. Instead of restricting our attention to the parameter space $\boldsymbol{\Omega}$, we will consider a functional

$$\Im_1\left[P_j(\theta)\right] = \sum_{k=1}^{K} \hat{r}_{jk} \ln P_j(\theta_k) + \left(\hat{n}_{jk} - \hat{r}_{jk}\right) \ln\left(1 - P_j(\theta_k)\right), \tag{28}$$

where the summation over j has been omitted because we can solve these equations separately for each item j. Further, the function $P_{\boldsymbol{\omega}_j}(\theta_k)$ has been replaced with the more general $P_j(\theta_k)$. Thus, we are no longer focusing on a parameter space $\boldsymbol{\Omega}$, but rather on a functional space. Taking the functional derivative of (28) with respect to $P_j(\theta_k)$, we get

$$\begin{aligned}\frac{\partial \Im_1\left[P_j(\theta)\right]}{\partial P_j(\theta_k)} &= \frac{\hat{r}_{jk}}{P_j(\theta_k)} - \frac{\left(\hat{n}_{jk} - \hat{r}_{jk}\right)}{1 - P_j(\theta_k)} \\ &= \frac{\hat{r}_{jk} - \hat{n}_{jk} P_j(\theta_k)}{P_j(\theta_k)\left[1 - P_j(\theta_k)\right]}.\end{aligned} \tag{29}$$

Notice that taking the functional derivative with respect to $P_j(\theta_k)$ made the summation "go away."

Now, if we set (29) equal to zero, we obtain

$$P_j(\theta_k) = \frac{\hat{r}_{jk}}{\hat{n}_{jk}}, \tag{30}$$

provided that $P_j(\theta_k)$ is in the open interval $(0,1)$. The solution in (30) appears elsewhere in the statistical estimation literature, albeit by different names. Prescher (2004) refers to this solution as a *relative frequency estimate*, or an "instance of the unrestricted probability model" (p. 4). This interpretation is consistent with the unconstrained optimization we have been considering. Wets (1999) refers to it as the *empirical probability mass function*, and goes on to say, "processing (all) the statistical information available, i.e. the information provided by the sample, yields the empirical distribution as best estimate" (p. 85). In our case, we have shown that (30) is the solution to the unconstrained optimization of the functional $\Im_1$.

To show the consistency of this approach with the constrained optimization we obtained in (27), we now suppose that the function $P_j(\theta_k)$ depends on a set of parameters $\boldsymbol{\omega}_j$. That is, $P_j(\theta_k) = P_{\boldsymbol{\omega}_j}(\theta_k)$ for every k. To enforce these constraints, we use a Lagrange multiplier λ_k for each of the k quadrature points:

$$\sum_{k=1}^{K} \lambda_k \left[P_j(\theta_k) - P_{\boldsymbol{\omega}_j}(\theta_k) \right] = 0. \tag{31}$$

Then we have a new functional, incorporating (31) into (28),

$$\begin{aligned}\Im_2\left[P_j(\theta)\right] = &\sum_{k=1}^{K} \hat{r}_{jk} \ln P_j(\theta_k) + \left(\hat{n}_{jk} - \hat{r}_{jk}\right)\ln\left(1 - P_j(\theta_k)\right) \\ &+ \sum_{k=1}^{K} \lambda_k \left[P_j(\theta_k) - P_{\boldsymbol{\omega}_j}(\theta_k)\right].\end{aligned} \tag{32}$$

As before, we take the functional derivative with respect to $P_j(\theta_k)$, and not surprisingly, our result looks quite similar to what we found in (29), with one important difference—the Lagrange multiplier λ_k :

$$\frac{\partial \Im_2\left[P_j(\theta)\right]}{\partial P_j(\theta_k)} = \frac{\hat{r}_{jk} - \hat{n}_{jk} P_j(\theta_k)}{P_j(\theta_k)\left[1 - P_j(\theta_k)\right]} + \lambda_k. \tag{33}$$

Setting this functional derivative equal to zero and solving gives us what looks like an uninteresting result:

$$\lambda_k = -\frac{\hat{r}_{jk} - \hat{n}_{jk} P_j(\theta_k)}{P_j(\theta_k)\left[1 - P_j(\theta_k)\right]}. \tag{34}$$

However, (34) will soon play an important role. We take the usual (not functional) derivative of (31) with respect to the parameters $\boldsymbol{\omega}_j$, set this derivative to zero, and find that

$$\frac{\partial}{\partial \boldsymbol{\omega}_j} \sum_{k=1}^{K} \lambda_k \left[P_j(\theta_k) - P_{\boldsymbol{\omega}_j}(\theta_k)\right] = \sum_{k=1}^{K} \lambda_k \left[-\frac{\partial}{\partial \boldsymbol{\omega}_j} P_{\boldsymbol{\omega}_j}(\theta_k)\right] = 0. \tag{35}$$

Combining the result in (34) with (35), we obtain

$$\sum_{k=1}^{K}\left[-\frac{\hat{r}_{jk}-\hat{n}_{jk}P_j(\theta_k)}{P_j(\theta_k)\left[1-P_j(\theta_k)\right]}\right]\left[-\frac{\partial}{\partial\boldsymbol{\omega}_j}P_{\boldsymbol{\omega}_j}(\theta_k)\right]=0$$

$$\Rightarrow\sum_{k=1}^{K}\frac{\hat{r}_{jk}-\hat{n}_{jk}P_j(\theta_k)}{P_j(\theta_k)\left[1-P_j(\theta_k)\right]}\frac{\partial P_{\boldsymbol{\omega}_j}(\theta_k)}{\partial\boldsymbol{\omega}_j}=0. \tag{36}$$

Finally, substituting our constraint $P_j(\theta_k) = P_{\boldsymbol{\omega}_j}(\theta_k)$ into (36) for each k, we arrive at the same estimating equation we obtained in (27).

An Information-Theoretic Interpretation

An obvious benefit of the unconstrained optimization of (28) is that its solution, given by (30), will result in the "best" M-step possible, given Assumptions 1–5. That is, the $Q(\cdot|\cdot)$ function for the unconstrained optimization will always be greater than or equal to the $Q(\cdot|\cdot)$ function for the constrained optimization. Another benefit, perhaps less obvious, is by that freeing the parametric dependence of the $P_j(\theta_k)$ while maintaining the binomial sampling assumption (see Assumption 5), we are squarely in the regular exponential family of functions (e.g., Dempster et al., 1977, pp. 3–4). Of particular importance for this family of functions is the convexity of its parameter space. In fact, convexity undergirds a fundamental convergence theorem of Csiszar and Tusnady (1984), which itself is intricately connected to the EM algorithm. The link between them is the information-theoretic quantity Kullback–Leibler (KL) divergence, or relative entropy (Cover & Thomas, 2006; Kullback & Leibler, 1951).

First we review the definition of KL divergence. For two probability density functions $f_1(x)$ and $f_2(x)$ with support $x\in X$, the KL divergence between $f_1(x)$ and $f_2(x)$ is

$$D(f_1\|f_2)=\sum_{x\in X}f_1(x)\ln\frac{f_1(x)}{f_2(x)}. \tag{37}$$

With an application of Jensen's inequality, it can be shown that $D(f_1 \| f_2) \geq 0$. Further, $D(f_1 \| f_2) = 0$ if and only if $f_1(x) = f_2(x), \forall x \in X$ (Cover & Thomas, 2006, p. 28).

Let us now recall the EM equation given in (6). As before, we will apply the IRT Assumptions 1–5. However, we will consider the unconstrained optimization for the M-step, and so we must modify our notation somewhat. Since the dependence on parameters $\boldsymbol{\omega}$ has been removed, we use a subscript on the relevant functions to denote an unconstrained optimization:

$$\begin{aligned} \ln L_0(\mathbf{X}) &= \sum_{i=1}^{N} \ln L_0(\mathbf{X}_i) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} g_0(\theta_k | \mathbf{X}_i) \ln \frac{L_0(\mathbf{X}_i, \theta_k)}{g_0(\theta_k | \mathbf{X}_i)}. \end{aligned} \tag{38}$$

If we multiply both sides of (38) by –1, we obtain

$$-\ln L_0(\mathbf{X}) = \sum_{i=1}^{N} \sum_{k=1}^{K} g_0(\theta_k | \mathbf{X}_i) \ln \frac{g_0(\theta_k | \mathbf{X}_i)}{L_0(\mathbf{X}_i, \theta_k)}. \tag{39}$$

Here we recognize the KL divergence between two density functions: the posterior density $g_0(\theta_k | \mathbf{X}_i)$, which is the conditional distribution of the latent variable given the observables; and the joint likelihood $L_0(\mathbf{X}_i, \theta_k)$ of the latent and observable variables. Following the definition of KL divergence in (37), we can write (39) as

$$\begin{aligned} -\ln L_0(\mathbf{X}) &= -\sum_{i=1}^{N} \ln L_0(\mathbf{X}_i) \\ &= \sum_{i=1}^{N} D\big(g_0(\theta | \mathbf{X}_i) \| L_0(\mathbf{X}_i, \theta)\big). \end{aligned} \tag{40}$$

The EM algorithm can then be interpreted as a method of alternating minimization of the KL divergence (Amari, 1994; Csiszar & Tusnady, 1984; Ip & Lalwani, 2000). That is, maximizing the marginal log likelihood $\ln L_0(\mathbf{X})$ is equivalent to minimizing the sum of KL divergences $D\left(g_0(\theta|\mathbf{X}_i)\|L_0(\mathbf{X}_i,\theta)\right)$, $i=1,2,\ldots,N$.

Theorem 2. We now state the alternating minimization theorem of Csiszar and Tusnady (1984) and show how it applies to the EM algorithm. Let P and Q be convex sets of finite measures, let $p_0 \in P$ be arbitrary, and let for each $n \geq 0$ $q_n \in Q$ minimize the KL informational divergence $D(p_n\|q)$ for $q \in Q$ while $p_{n+1} \in P$ minimizes $D(p\|q_n)$ for $p \in P$. Then $D(p_n\|q_n)$ converges to the infimum of $D(p\|q)$ on $P_O \times Q$, where P_O is the set of all $p \in P$ such that $D(p\|q_n) < +\infty$ for some n.

To apply this theorem to the unconstrained optimization, we must identify the convex sets P and Q. As a normalized probability density, the posterior density $g_0(\theta|\mathbf{X}_i)$ is an element of a convex set, so we note that $g_0(\theta|\mathbf{X}_i) \in P$. Now, Theorem 2 would apply directly to the marginal log likelihood in (40) if we could show that $L_0(\mathbf{X}_i,\theta)$ is an element of a convex set. However, $L_0(\mathbf{X}_i,\theta)$ is not a normalized probability density function. Ideally, if we could find a set Q, where for each $q \in Q$, q is also a probability density function, then Q would be convex. With such a $q \in Q$—if we can show that minimizing $D(p\|q)$ also minimizes (40)—we have assurance from Theorem 2 that the EM algorithm for the unconstrained optimization will converge to a global optimum of the marginal log likelihood.

The solution is to consider first the likelihood function for a single item and individual test taker. That is, for a test taker i responding to item j, the likelihood function at the k^{th} quadrature point is

$$L\left(X_{ij}=x_{ij}\middle|\theta_k,p_{jk}\right)=p_{jk}{}^{x_{ij}}\left(1-p_{jk}\right)^{1-x_{ij}}, \tag{41}$$

where $p_{jk} = P_j(\theta_k)$, as denoted earlier. Thus, the likelihood of a response is itself a normalized probability density function. If we let

$$\pi_{ijk} = \pi_{jk}(x_{ij}) = \begin{cases} 1 - p_{jk}, & x_{ij} = 0 \\ p_{jk,} & x_{ij} = 1, \end{cases} \tag{42}$$

then we see that π_{ijk} is a normalized probability and is thus an element of a convex set. Then Theorem 2 applies for test taker i and item j as

$$D\left(g\left(\theta|\mathbf{X}_i, \boldsymbol{\pi}_i\right)\middle\|\pi_{ij}\right) = \sum_{k=1}^{K} g\left(\theta|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln \frac{g\left(\theta|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right)}{\pi_{ijk}}, \tag{43}$$

where the parameter vector $\boldsymbol{\pi}_{ik} = [\pi_{i1k} \quad \pi_{i2k} \quad \cdots \quad \pi_{iJk}]$. For J items, we have

$$\begin{aligned} \sum_{j=1}^{J} D\left(g\left(\theta|\mathbf{X}_i, \boldsymbol{\pi}_i\right)\middle\|\pi_{ij}\right) &= \sum_{j=1}^{J}\sum_{k=1}^{K} g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln \frac{g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right)}{\pi_{ijk}} \\ &= J\sum_{k=1}^{K} g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \\ &\quad -\sum_{j=1}^{J}\sum_{k=1}^{K} g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln \pi_{ijk}. \end{aligned} \tag{44}$$

Now, in applying our notation from (42) to (40), we find that

$$\begin{aligned} -\ln L_0(\mathbf{X}_i) &= D\left(g_0\left(\theta|\mathbf{X}_i\right)\middle\|L_0(\mathbf{X}_i, \theta)\right) \\ &= \sum_{k=1}^{K} g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln \frac{g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right)}{L\left(\mathbf{X}_i, \theta_k|\mathbf{p}_k\right)} \\ &= \sum_{k=1}^{K} g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \\ &\quad -\sum_{k=1}^{K} g\left(\theta_k|\mathbf{X}_i, \boldsymbol{\pi}_{ik}\right) \ln \prod_{j=1}^{J} L\left(X_{ij}|\theta_k, p_{jk}\right) p(\theta_k), \end{aligned} \tag{45}$$

recalling that the joint likelihood $L(\mathbf{X}_i, \theta_k | \mathbf{p}_k) = \prod_{j=1}^{J} L(X_{ij} | \theta_k, p_{jk}) p(\theta_k)$. Since $\pi_{ijk} = L(X_{ij} | \theta_k, p_{jk})$, (45) becomes

$$\begin{aligned} D\left(g_0(\theta|\mathbf{X}_i) \| L_0(\mathbf{X}_i, \theta)\right) = & \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \\ & - \sum_{j=1}^{J} \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln \pi_{ijk} \\ & - \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln p(\theta_k). \end{aligned} \tag{46}$$

Returning now to (44), we see that by rearranging terms,

$$\begin{aligned} \sum_{j=1}^{J} \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln \pi_{ijk} = & J \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \\ & - \sum_{j=1}^{J} D\left(g(\theta|\mathbf{X}_i, \boldsymbol{\pi}_i) \| \pi_{ij}\right). \end{aligned} \tag{47}$$

Substituting (47) into the second term on the right-hand side of (46), we find that

$$\begin{aligned} D\left(g_0(\theta|\mathbf{X}_i) \| L_0(\mathbf{X}_i, \theta)\right) = & \sum_{j=1}^{J} D\left(g(\theta|\mathbf{X}_i, \boldsymbol{\pi}_i) \| \pi_{ij}\right) \\ & - (J-1) \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \\ & - \sum_{k=1}^{K} g(\theta_k | \mathbf{X}_i, \boldsymbol{\pi}_{ik}) \ln p(\theta_k). \end{aligned} \tag{48}$$

We now show that the alternating minimization of $\sum_{j=1}^{J} D\left(g(\theta|\mathbf{X}_i, \boldsymbol{\pi}_i) \| \pi_{ij}\right)$ is equivalent to the alternating minimization of $D\left(g_0(\theta|\mathbf{X}_i) \| L_0(\mathbf{X}_i, \theta)\right)$. First consider minimizing $\sum_{j=1}^{J} D\left(g(\theta|\mathbf{X}_i, \boldsymbol{\pi}_i) \| \pi_{ij}\right)$ with respect to π_{ij}, holding the posterior density $g\left(\theta | \mathbf{X}_i, \boldsymbol{\pi}_i^0\right)$ fixed. This minimization is analogous to the unconstrained M-step of the EM algorithm (see (28) and (29)). Since the prior distribution $p(\theta_k)$ is also fixed, the second and third

terms on the right-hand side of (48) are constants with respect to π_{ij}. Thus, minimizing $\sum_{j=1}^{J} D\left(g\left(\theta\middle|\mathbf{X}_i,\boldsymbol{\pi}_i\right)\middle\|\pi_{ij}\right)$ is equivalent to minimizing $D\left(g_0\left(\theta\middle|\mathbf{X}_i\right)\middle\|L_0\left(\mathbf{X}_i,\theta\right)\right)$, with $g_0\left(\theta\middle|\mathbf{X}_i\right)$ fixed.

Now consider minimizing the right-hand side of (48) with respect to $g\left(\theta\middle|\mathbf{X}_i,\boldsymbol{\pi}_i\right)$. This minimization is analogous to the E-step. If $\pi_{ij}^{(t)}$ denotes the optimal set of parameters obtained in the previous minimization, then from earlier arguments on finding an optimal bound on the marginal log likelihood function (see (6) and (7)), we suppose that the posterior distribution $g\left(\theta\middle|\mathbf{X}_i,\boldsymbol{\pi}_i^{(t)}\right)$ minimizes (48). To show that $g\left(\theta\middle|\mathbf{X}_i,\boldsymbol{\pi}_i^{(t)}\right)$ does indeed minimize (48), start with (44):

$$\begin{aligned}\sum_{j=1}^{J} D\left(g\left(\theta\middle|\mathbf{X}_i,\boldsymbol{\pi}_i^{(t)}\right)\middle\|\pi_{ij}^{(t)}\right) = J\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\ln g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\\ -\sum_{j=1}^{J}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\ln \pi_{ijk}^{(t)}.\end{aligned} \tag{49}$$

Then (48) becomes

$$\begin{aligned}D\left(g_0\left(\theta\middle|\mathbf{X}_i\right)\middle\|L_0\left(\mathbf{X}_i,\theta\right)\right) = J\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\ln g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\\ -\sum_{j=1}^{J}\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\ln \pi_{ijk}^{(t)}\\ -(J-1)\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\ln g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\\ -\sum_{k=1}^{K} g\left(\theta_k\middle|\mathbf{X}_i,\boldsymbol{\pi}_{ik}^{(t)}\right)\ln p(\theta_k).\end{aligned} \tag{50}$$

Since $g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\pi}_{ik}^{(t)}\right) = \prod_{j=1}^{J} \pi_{ijk}^{(t)} p(\theta) \Big/ L_0^{(t)}(\mathbf{X}_i)$, (50) simplifies to

$$
\begin{aligned}
D\left(g_0\left(\theta \middle| \mathbf{X}_i\right) \middle\| L_0\left(\mathbf{X}_i, \theta\right)\right) = & -L_0^{(t)}(\mathbf{X}_i) + \sum_{k=1}^{K} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\pi}_{ik}^{(t)}\right)\left[\sum_{j=1}^{J} \ln \pi_{ijk}^{(t)} + \ln p(\theta_k)\right] \\
& - \sum_{j=1}^{J}\sum_{k=1}^{K} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\pi}_{ik}^{(t)}\right) \ln \pi_{ijk}^{(t)} \\
& - \sum_{k=1}^{K} g\left(\theta_k \middle| \mathbf{X}_i, \boldsymbol{\pi}_{ik}^{(t)}\right) \ln p(\theta_k) \\
= & -\ln L_0^{(t)}(\mathbf{X}_i).
\end{aligned}
\tag{51}
$$

Thus, the posterior density $g\left(\theta \middle| \mathbf{X}_i, \boldsymbol{\pi}_i^{(t)}\right)$ does minimize the right-hand side of (48), attaining the optimal bound (i.e., equality) on the marginal log likelihood function.

Testing Models Against an Information Bound

The result in (48), combined with Theorem 2, yields a bound on the information provided by a sample under weaker assumptions than the parameterized response model in (1). The generality of this result is such that *any* latent trait model conforming to Assumptions 1–5 can be tested against this information bound. Not only does that include other IRT models such as the 2PL and 1PL models (e.g., Hambleton & Swaminathan, 1985), but it extends to a host of other IRT models such as the noisy input deterministic ‘and’ gate (NIDA); the deterministic input noisy ‘and’ gate (DINA); nonparametric IRT; and cognitive diagnosis models (de la Torre, 2009; Junker & Sijtsma, 2001).

Typically, latent trait models can only be compared against one another. With this bound, we can determine whether a specific parameterized model differs from the unconstrained optimization in any significant way. That is, we have a fixed point of reference.

Consider a likelihood ratio test between two optimized marginal log likelihoods: $\ln L_0(\mathbf{X})$ in the unconstrained case and $\ln L(\mathbf{X}|\boldsymbol{\omega})$ in the constrained case. The asymptotic likelihood ratio test (Spanos, 1999, p. 715) is

$$\lambda = -2\left[\ln \frac{L(\mathbf{X}|\boldsymbol{\omega})}{L_0(\mathbf{X})}\right] = -2\left[\ln L(\mathbf{X}|\boldsymbol{\omega}) - \ln L_0(\mathbf{X})\right] \sim \chi^2_r, \tag{52}$$

where under the null hypothesis λ is asymptotically distributed as a χ^2 distribution with r degrees of freedom. The quantity r indicates the difference in the degrees of freedom, or the difference in the number of restrictions, associated with $L_0(\mathbf{X})$ and $L(\mathbf{X}|\boldsymbol{\omega})$.

Before using the likelihood ratio test in (52), we should note that sample estimates of the marginal log likelihood functions are biased. Akaike (1973), in developing an information criterion for model comparisons, corrects for this bias by adjusting the log likelihood function by the number of free parameters in the model. The Akaike Information Criterion (AIC) for a model with q free parameters (Akaike, 1973; Bozdogan, 1987) is given by

$$AIC(q) = -2\log L(\mathbf{X}) + 2q. \tag{53}$$

From (53), we identify a bias-corrected marginal log likelihood as

$$l_c(\mathbf{X}) = -\frac{1}{2} AIC(q) = \log L(\mathbf{X}) - q. \tag{54}$$

With this bias correction, we can rewrite the likelihood ratio test in (52) as

$$\lambda_c = 2\left|l_c\left(\mathbf{X}|\boldsymbol{\omega}\right) - l_c\left(\mathbf{X}\right)\right| \sim \chi_r^2, \tag{55}$$

where the absolute value is required because, after correcting for the bias, $l_c\left(\mathbf{X}|\boldsymbol{\omega}\right) - l_c\left(\mathbf{X}\right)$ might be greater than, less than, or equal to zero. That is, before applying the bias correction, $-2\left[\ln L\left(\mathbf{X}|\boldsymbol{\omega}\right) - \ln L_0\left(\mathbf{X}\right)\right] \geq 0$; however,

$$\begin{aligned} -2\left[l_c\left(\mathbf{X}|\boldsymbol{\omega}\right) - l_c\left(\mathbf{X}\right)\right] &= -2\left\{\left[\ln L\left(\mathbf{X}|\boldsymbol{\omega}\right) - q_{\boldsymbol{\omega}}\right] - \left[\ln L_0\left(\mathbf{X}\right) - q_0\right]\right\} \\ &= -2\left[\ln L\left(\mathbf{X}|\boldsymbol{\omega}\right) - \ln L_0\left(\mathbf{X}\right)\right] - 2\left(q_0 - q_{\boldsymbol{\omega}}\right), \end{aligned} \tag{56}$$

where q_0 and $q_{\boldsymbol{\omega}}$ are the number of free parameters from the unconstrained and constrained optimized marginal log likelihoods, respectively.

We now examine the number of degrees of freedom, or free parameters, associated with the two optimized marginal log likelihoods in (52) and (56). For the unconstrained optimization in (28), there are jk values of $P_j\left(\theta_k\right)$ to estimate, and these are obtained by (30). However, for a given item j, the $\hat{r}_{jk}$ appearing in (30) must sum to the total number correct across k quadrature points; thus, only $k-1$ values of $P_j\left(\theta_k\right)$ are free to vary per item. For j conditionally independent items, the total number of degrees of freedom for the unconstrained optimization is then $j\left(k-1\right)$. In the case of the constrained optimization, the degrees of freedom will vary depending on the dimensionality of $\boldsymbol{\omega}$. For example, the 3PL model in (1) estimates three parameters for each item; thus, there are $3j$ degrees of freedom for j items. Hence, a likelihood ratio test of the 3PL model against the unconstrained optimization would have a difference in degrees of freedom of $r = j\left(k-1\right) - 3j = j\left(k-4\right)$.

Empirical Examples

We now explore applications of this bound and the associated likelihood ratio test to real-world data problems. For all of the following examples, code written by the author in the SAS-IML language (SAS Institute, 2008) was used.

Example 1. *Constrained and unconstrained optimizations for a dataset of observed item responses*. In this example, data collected from a large-scale standardized assessment were used. The data consisted of 10,000 test-taker responses to 60 dichotomously scored items. The constrained optimization employed the EM algorithm under the 3PL IRT model as defined in (1). The unconstrained optimization also employed the EM algorithm, but under the less restrictive conditions described above (see section titled Optimizing Bounds on the Log Likelihood Function). For both optimizations, 31 quadrature points were used. At convergence, $\ln L\left(\mathbf{X}|\boldsymbol{\omega}\right) = -349942$ for the constrained optimization, and $\ln L_0\left(\mathbf{X}\right) = -343447$ for the unconstrained optimization. With $3\times 60 = 180$ free parameters for the constrained optimization and with $30\times 60 = 1{,}800$ free parameters for the unconstrained optimization, the bias-corrected marginal log likelihoods were -350122 and -345247, respectively. The difference in degrees of freedom between the two optimizations was $r = 60\left(31-4\right) = 1{,}620$. Using the likelihood ratio test in (55), $\lambda_c = 9{,}750$; for a χ^2 distribution with 1,620 degrees of freedom, the upper-tail probability $P\left(\chi^2_{1620} \geq 9{,}750\right) \doteq 0$. Thus, we can conclude that the constrained marginal log likelihood is significantly different from the bound obtained from the unconstrained optimization.

In light of what is known about the fitting of psychometric models to data obtained from educational assessments, this result is not unexpected. Rather, it is consistent with what is almost always observed in the field: Psychometric models capture some, not all, of the information contained in test-taker responses. Another gauge on the information loss is the value of the respective functionals, summed over all items, at convergence (see (28) and (32)). For the unconstrained optimization, $\sum_j \Im_1\left[P_j\left(\theta\right)\right] = -323399;$ for

the constrained optimization, $\sum_j \Im_2\left[P_j(\theta)\right] = -339676$. As expected, the functional for the unconstrained optimization is greater than that for the constrained optimization.

Example 2. *Constrained and unconstrained optimizations for a dataset of simulated item responses*. In this example, the 3PL item parameter estimates for the 60 items obtained from the constrained optimization in Example 1 were used to simulate the responses of 10,000 test takers to those 60 items. Test-taker latent trait θ was drawn from a $N(0,1)$ distribution. Constrained and unconstrained optimizations were conducted in the same manner as described in Example 1. At convergence, $\ln L(\mathbf{X}|\boldsymbol{\omega}) = -349007$ for the constrained optimization and $\ln L_0(\mathbf{X}) = -347802$ for the unconstrained optimization. Once again, 31 quadrature points were used; the bias-corrected marginal log likelihoods were -349187 for the constrained optimization and -349602 for the unconstrained optimization. The likelihood ratio test statistic in this case was $\lambda_c = 830$; for a χ^2 distribution with 1,620 degrees of freedom, the upper-tail probability $P\left(\chi^2_{1620} \geq 830\right) \doteq 1$. Thus, we conclude that the constrained and unconstrained marginal log likelihoods are not significantly different.

This result is what should be obtained when the generating model (i.e., the model used to simulate the response data) and the estimating model (i.e., the model fitted to the data) are identical. In contrast to the observed response data, the 3PL IRT model should capture all of the systematic information contained in the simulated item response data. For comparison, the values of the functionals were $\sum_j \Im_1\left[P_j(\theta)\right] = -333481$ and $\sum_j \Im_2\left[P_j(\theta)\right] = -338613$ for the unconstrained and constrained optimizations, respectively.

Discussion

As mentioned in the introduction, educational and psychological measurement is concerned with measuring unobservable constructs in the presence of non-negligible error; thus, statistical models play a crucial role in measurement. IRT models are a class of such statistical models, rooted in the broader latent trait theory. However, more

complex IRT models (e.g., the 3PL model) did not become practical until the advent of a computationally feasible method of parameter estimation. The EM algorithm provided the necessary methodology for efficiently estimating IRT models. It should be noted that over time, and with increased availability of more powerful computational resources, other methods such as Markov chain Monte Carlo (MCMC) have gained in popularity (e.g., Patz & Junker, 1999). In fact, the two approaches can be combined, leading to algorithms such as Monte Carlo EM (see McLachlan & Krishnan, 2008 for an overview).

A persistent challenge for the EM algorithm is that while it ensures convergence to a local maximum, there is no guarantee that it will converge to the global maximum. One exception is when convexity of the parameter spaces can be established, as was explored in the present report for the unconstrained M-step. For more general cases, flexible optimization schemes with global convergence properties are in order, thus setting the stage nicely for collaboration between the disciplines of statistics and operations research.

Such collaboration is the focus of Jank's (2006) chapter on stochastic implementations of EM, where he reviews nonstochastic ("deterministic") EM, provides examples, and points out some of its limitations, including the possibility of the algorithm's convergence to a local maximum. With the goal of global optimization in mind, he proposes a stochastic implementation of EM via a genetic algorithm. Further, he suggests "…[bridging] a gap between the field of statistics, which is home to extensive research on the EM algorithm, and the field of operations research, in which work on global optimization thrives." Wets (1999), although not focusing specifically on the EM algorithm, provides an elucidating commentary on the connections between statistics and optimization, further exploring the role of constraints in statistical estimation problems. The cross-entropy method described in de Boer, Kroese, Mannor, and Rubinstein (2005) shares many features of the EM algorithm. In fact, although motivated by different problems, the solution to the unconstrained optimization of the binomial likelihood function discussed here appears in their work as well (see pp. 25–26, 35–36).

In this report, the sometimes overlooked relationship between the EM algorithm and information theory is brought to the fore. The reward for doing so is of both theoretical

and practical interest. On the theoretical side, the EM algorithm as a minimization of KL divergence admits the convergence results from Csiszar and Tusnady (1984), a consequence of the convexity of the parameter space of the binomial likelihood common to latent trait models with dichotomous response variables. These results apply for an unconstrained optimization in the M-step, derived by functional differentiation and later compared with a constrained optimization. On the practical side, a likelihood ratio test between the marginal log likelihood functions from the constrained and unconstrained optimizations provides a method for comparing models against a fixed reference. An apparent benefit of the unconstrained optimization is that it provides an overarching model to test models.

In many respects the information-theoretic approach to model comparisons suggested in this report is close in spirit to information criteria such as the AIC (e.g., Bozdogan, 1987). Also at the heart of the AIC are the minimization of KL divergence and the likelihood ratio tests. Bozdogan (1987) writes:

> [The AIC] has enormous practical importance and is one more demonstration of the importance of the likelihood ratio criterion in statistical inference.… [The] major development of AIC lies in the direct extension of an entropic or information theoretic interpretation of the method of maximum likelihood. Its introduction is based on the entropy maximization principle, or minimizing its negative; it is based on the minimization of the K-L information quantity" (p. 347).

While these approaches clearly share a common heritage, how they should be applied in practice differs. Given a set of models fitted to a dataset, the model resulting in the minimum AIC suggests an optimal balance between goodness-of-fit and parsimony. Thus, comparing among AIC values is particularly useful when a researcher has a number of competing models and must choose the "best" one. In that case, all comparisons among the potential models are relative. The approach suggested in this report is to find an information bound. Models can then be compared not only to one another (relative comparison) but also to this bound (absolute comparison).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceeding of the Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Amari, S.-i. (1994). *Information geometry of the EM and em algorithms for neural networks*. Tokyo: Department of Mathematical Engineering, University of Tokyo.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics*. Pacific Grove, CA: Duxbury.

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

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference.* London: University of London.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*(2), 179–197.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory, 2nd ed.* Hoboken, NJ: John Wiley & Sons, Inc.

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

Csiszar, I., & Tusnady, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions: Supplement Issue 1*, 205–237.

de Boer, P.-T., Kroese, D., Mannor, S., & Rubinstein, R. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research, 134*(1), 19–67.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

Dellaert, F. (2002). *The expectation maximization algorithm* (No. GIT-GVU-02-20). Atlanta, GA: Georgia Institute of Technology.

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

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

Harpaz, R., & Haralick, R. (2006). *The EM algorithm as a lower bound optimization technique*: New York, NY: Graduate Center, City University of New York.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*(3), 243–271.

Ip, E. H., & Lalwani, N. (2000). Notes and comments: A note on the geometric interpretation of the EM algorithm in estimating item characteristics and student abilities. *Psychometrika, 65*(4), 533.

Jank, W. (2006). The EM algorithm, its randomized implementation and global optimization: Some challenges and opportunities for operations research. In F. B. Alt, M. C. Fu, & B. L. Golden (Eds.), *Perspectives in Operations Research* (Vol. 36, pp. 367–392). New York, NY: Springer.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

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

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions*. Hoboken, NJ: Wiley-Interscience.

Minka, T. P. (1998). Expectation-maximization as lower bound maximization. Retrieved from http://research.microsoft.com/en-us/um/people/minka/papers/em.html

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178.

Prescher, D. (2004). A tutorial on the expectation-maximization algorithm including maximum-likelihood estimation and EM training of probabilistic context-free grammars. Retrieved from http://arxiv.org/abs/cs/0412015.

SAS Institute. (2008). SAS-IML: Interactive Matrix Language (Version 9.2). Cary, NC: Author.

Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161–169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417–426.

Wets, R. J. B. (1999). Statistical estimation from an optimization viewpoint. *Annals of Operations Research, 85*(1), 79.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics, 11*(1), 95–103.