ACT Research Report Series 99-7

The Relationship Between the Bock-Aitkin Procedure and the EM Algorithm for IRT Model Estimation

Yaowen Hsu

Terry A. Ackerman

Meichu Fan

ACT

November 1999

The Relationship Between the Bock-Aitkin Procedure and the EM Algorithm for IRT Model Estimation

Yaowen Hsu
Terry A. Ackerman
Meichu Fan

Abstract

It has been previously shown that the Bock-Aitkin procedure is an instance of the EM algorithm when trying to find the marginal maximum likelihood estimate for a *discrete* latent ability variable (latent trait). In this paper, it is shown that the Bock-Aitkin procedure is a numerical implementation of the EM algorithm for a *continuous* latent ability variable using numerical quadrature. Further, the relationship between the EM algorithm, marginal maximum likelihood estimation, and the Bock-Aitkin procedure is described for both the discrete and continuous cases. Some issues concerning the use of the Bock-Aitkin procedure and the EM algorithm are addressed.

Acknowledgements

Thanks to Bradley Hanson for helpful discussions and comments. The first author also thanks Subrata Kundu, Robert Tsutakawa, Louis DiBello, Daniel Eignor, Matthew Schulz, and colleagues in the STAR department for encouragement, comments, and/or reviews on earlier drafts.

The Relationship Between the Bock-Aitkin Procedure and the EM Algorithm for IRT Model Estimation

Bock and Aitkin (1981) described a procedure to solve marginal maximum likelihood (MML) estimation. It can be shown the Bock-Aitkin procedure is an EM algorithm, which was formulated by Dempster, Laird, and Rubin (1977). Along with the introduction of the software BILOG (Mislevy & Bock, 1993), the Bock-Aitkin procedure and the EM algorithm have become popular item parameter estimation approaches. For example, researchers have developed methods for using the Bock-Aitkin procedure under a variety of item response theory (IRT) models (Bock, Gibbons, & Muraki,1988; Mislevy & Bock, 1985; Muraki, 1992; Muraki & Carlson, 1995; Thissen, 1982). Other researchers have also developed procedures for using the EM algorithm for parameter estimation in IRT (Rigdon & Tsutakawa, 1983; Tsutakawa, 1984, 1985).

However, in their paper Bock and Aitkin (1981) did not show how their procedure is an instance of the EM algorithm, although they mentioned that the procedure "is closely related to the EM algorithm" (p.444) but "not the same as the general EM algorithm" (p.448; also quoted in Lewis, 1985, p.206). Lewis (1985) gave insightful comments and sketched a proof of the relationship between the EM algorithm and the Bock-Aitkin procedure with a general discrete distribution of θ-ability (conventionally, θ denotes the ability or latent trait. In this paper, the discussion includes multiple abilities. The θ distribution is referred to as unidimensional or multidimensional by context). Harwell, Baker, and Zwarts (1988) provided a mathematical background for MML estimation and the Bock-Aitkin procedure, and demonstrated the procedure under the assumption of a discrete θ distribution. Rigdon and Tsutakawa (1983) and Tsutakawa (1984,

1985) showed how to use the EM algorithm to find MML estimates of the IRT model parameters (The term EM/MMLE will be referred to as the EM algorithm for MML estimation). Woodruff and Hanson (1996) described the EM algorithm for maximum likelihood (ML) parameter estimation for finite mixtures (i.e., when θ is discrete). They then related the Bock-Aitkin procedure to theirs.

In spite of the various discussions relating to the Bock-Aitkin procedure and the EM algorithm, it still seems that the underlying mechanics of these two algorithms and their relationship are not well-understood. This understanding can clarify some issues. For example, is the underlying θ scale discrete or continuous when the Bock-Aitkin procedure is used? Bock and Aitkin (1981) heuristically described two approaches regarding the use of θ to implement their procedure. One approach uses Gauss-Hermite quadrature to compute the integral needed when θ has the normal distribution (here called BA1). Another uses a discrete distribution (here referred as discrete representation) to approximate a specified θ (continuous) distribution with a finite number of prespecified and equally spaced points (here called BA2). No integral is evaluated. The BA2 procedure is extended to the case where the θ distribution is unspecified and is estimated based on a discrete representation (Mislevy & Bock, 1985, here called BAM). In some literature (e.g., Mislevy & Stocking, 1989, p.61), the discrete points and probabilities of a discrete representation are labeled as the quadrature points and weights, although the approaches differ–for example, the quadrature points in Gauss-Hermite quadrature are not equally spaced. To date, the Bock-Aitkin procedure has only been shown to be an instance of the EM algorithm by using a discrete representation, or in fact, by assuming that θ is discrete with a finite number

of predetermined values (Harwell et al., 1988; Lewis, 1985; Woodruff & Hanson, 1996). For example, Mislevy and Stocking (1989) stated that the steps of the Bock-Aitkin procedure are "exactly the steps of the EM algorithm (Dempster, Laird, & Rubin, 1977), in the special case of missing multinomial indicators' because θs are limited to a finite number of values" (p.61). Then, because the EM algorithm can give MML estimates (Tsutakawa, 1985; also will be described later), we know that the estimates obtained from the Bock-Aitkin procedure are MML estimates (although it is not guaranteed in general). However, when the resulting item parameter estimates are applied to the estimation of θ, the θ scale is usually assumed to be continuous rather than discrete. Moreover, this can not justify how the procedure in which Bock (1995) used constant 243-point fractional quadrature for any number of multiple abilities can give MML estimates (the method of adaptive quadrature is also used by Bock & Schilling, 1997, cited in Bock, 1997), unless like using Gauss-Hermit quadrature the nodes and weights of fractional quadrature are treated as a discrete representation, but not elements for numerical integration (which means, however, a three-point discrete distribution–or a three-point histogram–may be used to approximate a normal distribution). Strictly speaking, we may say that the BA1 procedure in which the standard normal distribution of ability is assumed has not yet been shown as an implementation of MML estimation. As will be shown below, from a theoretical view of point the Bock-Aitkin procedure is a reformulation (or an instance) of EM/MMLE with a continuous or discrete θ distribution (note that the EM algorithm can be applied for a more general θ distribution). BA1, BA2, and BAM are the implementation versions of EM/MMLE for a continuous θ distribution. The use of fractional quadrature in the Bock-Aitkin procedure can be justified. If θ is discrete,

then BAM is the same as EM/MMLE.

Moreover, from an EM algorithm perspective, the Bock-Aitkin procedure can naturally include estimation of the θ-distribution parameters as well as item parameters, as BAM does. Note that this simultaneous estimation is not joint ML (JML) estimation in which, for example, individual examinee θ values (instead of θ-distribution parameters) are estimated. One feature of the Bock-Aitkin procedure is that it reduces some computation load of the EM algorithm. However, the Bock-Aitkin procedure, being an EM algorithm and hence inheriting the algorithm's statistical properties, may not converge to the MML estimate.

Therefore, different from earlier work restricted to the case of discrete θ, in this paper the relationship between both algorithms is described with a general θ distribution, although a continuous scale is emphasized. A parametric distribution of θ is assumed; that is, the (underlying) θ distribution is specified in which the distribution parameters are either known or unknown. BAM assumes a nonparametric/unspecified θ distribution, then the procedure is for semiparametric IRT models where the item response function (IRF) is specified but the θ distribution not (see Holland 1990; if both are not specified, then it is a nonparametric IRT model), and is beyond the scope of this paper. On the other hand, both BA2 and BAM use a discrete representation, so the θ distribution is considered being discrete, at least, for MML estimation. It is because in the course of MML estimation, a discrete distribution is actually used, although it is an approximation of a continuous distribution (B. Hanson, personal communication, December 8, 1997).

First, the MML estimation and the Bock-Aitkin procedure are described. Next, the EM algorithm is introduced and its application to IRT settings is given. Then, the relationship be-

tween the EM algorithm and the Bock-Aitkin procedure is described both when θ is discrete and continuous. Finally, some issues regarding the use of the EM algorithm and Bock-Aitkin procedure are discussed.

Marginal Maximum Likelihood Estimation

Consider dichotomously scored items (binary items; i.e., an answer to the item is classified into one of two score categories; the category of interest is labeled "correct", and "incorrect" otherwise). The arguments developed in this paper can be applied to polytomous items when each score category is treated as a "binary item" (see Muraki, 1992; Muraki & Carlson, 1995). Using conventional terminology, an item *response* refers to an item score. Let the random variable U denote the item score which takes on 1, if the answer is correct, and 0 otherwise. A test T of k given and fixed binary items administered to a population C of examinees is characterized by the random vector $\underline{U}$ which is the response pattern (or response vector), $\underline{U} = (U_1, U_2, \cdots, U_k)'$, where the prime ($'$) represents the matrix transpose. Note that as defined by Holland (1990), a test is "a specific set of questions with specific directions, given under standardized conditions of timing, item presentation, and so forth. If any of these elements change, the resulting test is different from T" (pp.577-578). The examinee is characterized (or assigned) by the m-component random vector $\underline{\Theta} = (\Theta_1, \Theta_2, \cdots, \Theta_m)'$, where each Θ_h denotes an ability (or latent trait) measured by the test T, and m is the number of dimensions of the complete latent space (of $\underline{\Theta}$), $1 \leq m \leq k$. The complete latent space refers to the minimum dimensionality necessary for local independence (for the test T) to hold–that is, $U_1, U_2, \cdots,$ and U_k are stochastically conditionally

independent, given that $\underline{\Theta}$ has the value $\underline{\theta}$. The assumption of local independence is sometimes referred to as the principle of local independence in the literature. Theoretically the θ scale can be any type, although a continuous one is of main interest here.

An IRF is used to define an IRT model (latent trait model). The IRF for an item with respect to T and $\underline{\Theta}$ (i.e., an item's IRF may vary for different test or examinee population) is specified as the conditional probability of a correct response to the item, given $\underline{\theta}$:

$$\Pr(U = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}) \equiv \Pr_{\underline{\phi}}(U = 1|\underline{\Theta} = \underline{\theta}), \tag{1}$$

on the m-dimensional complete latent space, where $\underline{\phi}$ is the vector of item parameters.

Given $\underline{\theta}$, the probability function of U can be expressed as

$$\Pr(U = u|\underline{\Theta} = \underline{\theta}, \underline{\phi}) = \Pr(U = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi})^u (1 - \Pr(U = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}))^{1-u}, \tag{2}$$

for u = 0 and 1; zero elsewhere. By local independence, the conditional probability of $\underline{U}$ of k items, given $\underline{\theta}$, is

$$\Pr(\underline{U} = \underline{u}|\underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^{k} \Pr(U_i = u_i|\underline{\Theta} = \underline{\theta}, \underline{\phi}_i), \tag{3}$$

where $\boldsymbol{\phi} = (\underline{\phi}_1, \underline{\phi}_2, \cdots, \underline{\phi}_k)$ is the item parameter matrix for the test T, where the components of the vector $\underline{\phi}_i$ for item i depend on the particular model.

Note that the IRF is called the item characteristic curve (ICC) or trace line, when m = 1. For a polytomous item, the IRF is the probability of a score category, given $\underline{\Theta} = \underline{\theta}$. Note that as Cressie and Holland (1983) pointed out, because the abilities are not directly observable, an alternative way that an IRT model can be characterized is by using manifest probabilities,

the proportions of examinees in a given population who obtain particular response patterns, $\Pr(\underline{U} = \underline{u})$ (also see Holland, 1981, 1990). This characterization also involves the specification of the θ distribution in addition to the IRF and explicitly expresses the assumption of local independence. Adams, Wilson, and Wu (1997) regarded this characterization as a structural measurement model, called a two-level nonlinear hierarchical item response model in which the specification of an IRF is still referred to as a conditional item response model and the specification of the θ distribution referred to as a population model. The manifest probabilities constitute the likelihood for MML estimation (see Equation 8). The first word "marginal" in the term MML can be dropped when this characterization is used (see Holland, 1990).

Consider a random sample of n examinees from the population C. Let

$$\mathbf{U}_{n\times k} = (\underline{U}_1, \underline{U}_2, \cdots, \underline{U}_n)'$$

be the response data (item score) matrix, consisting of response patterns $\underline{U}_j = (U_{j1}, U_{j2}, \cdots, U_{jk})'$, where U_{ji} denotes the item score of examinee j on item i. The $n \times m$ ability matrix of n randomly selected examinees is

$$\mathbf{\Theta} = \begin{pmatrix} \underline{\Theta}_1' \\ \vdots \\ \underline{\Theta}_n' \end{pmatrix} = \begin{pmatrix} \Theta_{11} & \cdots & \Theta_{1m} \\ \vdots & \ddots & \vdots \\ \Theta_{n1} & \cdots & \Theta_{nm} \end{pmatrix}$$

For simplicity, let

$$p_i(\underline{\theta}_j) \equiv \Pr(U_{ji} = 1 | \underline{\Theta}_j = \underline{\theta}_j, \underline{\phi}_i). \tag{4}$$

The conditional probability that an $n \times k$ response data matrix $\boldsymbol{u}$ on a test will be produced by a

group of n examinees with $\underline{\theta}_j$ who respond independently is

$$\begin{aligned}\Pr(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{\phi}) &= \prod_{j=1}^{n} \Pr(\underline{U}_j = \underline{u}_j|\underline{\Theta}_j = \underline{\theta}_j, \boldsymbol{\phi}) \\ &= \prod_{j=1}^{n}\prod_{i=1}^{k} p_i(\underline{\theta}_j)^{u_{ji}}(1 - p_i(\underline{\theta}_j))^{1-u_{ji}}. \qquad (5)\end{aligned}$$

If the values of $\underline{\theta}_1, \underline{\theta}_2, \cdots,$ and $\underline{\theta}_n$ of the n particular examinees are of interest, that is, this group is the target population C, then $\boldsymbol{\theta}$ in the conditional probability of Equation 5 is regarded as a nonstochastic matrix of incidental (or person) parameters that can be estimated along with item parameters by observing the data matrix $\boldsymbol{u}$, since $\boldsymbol{\theta}$ is unknown but fixed.

Both item and person parameters can be estimated using ML estimation. The likelihood function is the conditionally joint probability of Equation 5, that is,

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{u}) = \Pr(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{\phi}).$$

A set of ML estimates (MLEs) for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ is obtained by simultaneously solving the likelihood equations which are

$$\frac{\partial \ln L(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{u})}{\partial \phi_z} = 0,$$

where ϕ_z is an element of $\boldsymbol{\phi}$, and

$$\frac{\partial \ln L(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{u})}{\partial \theta_{jl}} = 0 \quad \forall j, l,$$

where θ_{jl} is an element of $\boldsymbol{\theta}$ (there are $n \times m$ equations). Because the system of likelihood equations is nonlinear, an iterative procedure is needed, for example, the Newton-Raphson method.

However, there are some problems for this simultaneous estimation. For instance, it is impractical to compute the inverse of the Hessian matrix of second derivatives needed in the Newton-Raphson method, when n (the number of examinees) is large.

Bock and Lieberman (1970) proposed a procedure, called MML estimation (Bock & Aitkin, 1981), to estimate item parameters for the normal ogive model. Using the estimated item parameters, the ability level of each examinee can then be estimated/predicted by, for example, the conditional probability of Equation 3 or the conditional distribution of $\underline{\Theta}$ given an observed response pattern $\underline{u}$ (see Holland, 1990). In the terminology used in linear models, MML estimation gives mixed-effect solutions by treating $\boldsymbol{\phi}$ fixed but $\boldsymbol{\theta}$ random (Bock & Aitkin, 1981; Rigdon & Tsutakawa, 1983).

In Bock and Lieberman's paper the θ distribution is fixed to the standard normal distribution $N(0,1)$, hence no distribution parameter needs to be estimated. Bock (1997) stated that "Although this integral [i.e., Equation 6 below] includes the density function of the latent proficiency[/ability] distribution, the location and scale of that distribution may be absorbed in the item parameters and arbitrarily set to conventional values such as 0 and 1. Thus, if the latent distribution is indexed only by location and scale parameters, as is the normal distribution, it is fully specified and does not increase the number of parameters to be estimated" (p.27). However, Bock (1997) mentioned that one of the requirements of practical testing programs is to estimate the θ distribution jointly with the item parameters (p.27); or, "one of the strengths of IRT, compared to classical test theory, is that it provides estimation of the proficiency latent distribution that is largely invariant with respect to test length and item characteristics" (p.29). In fact, in

addition to item parameters, MML estimation as presented here naturally includes estimation of the θ-distribution parameters for those IRT models such as the one-parameter Rasch model and a model with a discrete θ scale. Bock (1997) pointed out the condition for this simultaneous estimation of both item parameters and θ-distribution parameters.

For a random sample of n examinees from a population C, let $\underline{\Theta}_1, \underline{\Theta}_2, \cdots$, and $\underline{\Theta}_n$, denote examinees' abilities that are independent and identically distributed with the density function, $g(\underline{\theta}|\underline{\beta})$, belonging to a family indexed by the vector of the distribution parameters $\underline{\beta}$. The distribution is usually assumed to be fixed as normal $N(0,1)$ for the unidimensional case, then $\underline{\beta}$ is known. One possible formulation of the (marginal) probability function of the response pattern $\underline{u}$ is

$$\begin{aligned} \Pr_{\boldsymbol{\phi},\,\underline{\beta}}(\underline{U}=\underline{u}) &= \Pr(\underline{U}=\underline{u}|\boldsymbol{\phi},\underline{\beta}) = \int f(\underline{u},\underline{\theta}|\boldsymbol{\phi},\underline{\beta})\,d\underline{\theta} \\ &= \int \Pr(\underline{U}=\underline{u}|\underline{\Theta}=\underline{\theta},\boldsymbol{\phi})g(\underline{\theta}|\underline{\beta})\,d\underline{\theta}. \end{aligned} \tag{6}$$

As mentioned above, $\Pr(\underline{U}=\underline{u}|\boldsymbol{\phi},\underline{\beta})$ is the manifest probability that Cressie and Holland (1983) and Holland (1981, 1990) used to characterize an IRT model. Moreover, the MML estimation procedure described here can be extended for other possible formulations of $\Pr(\underline{U}=\underline{u})$ which may have different or additional types of model parameters, as long as these parameters can be attributed as item parameters or θ-distribution parameters; for example, Adams, Wilson, and Wu (1997) introduced examinee demographic variables into the θ distribution.

For k items, there are 2^k possible response patterns. The n randomly selected examinees

will give n (not necessarily distinct) observed response patterns. Let $\underline{u}_w$ denote the distinct response patterns that appear in a test, w = 1, 2, ... , s and $s \leq \min\{n, 2^k\}$. Let r_w denote the number of examinees whose response pattern is $\underline{u}_w$ such that $\sum_{w=1}^{s} r_w = n$.

Note that if $\underline{\Theta}$ is discrete with q possible values $\underline{x}_l$, $l = 1, \ldots, q$, then the θ-distribution parameters are q probabilities β_l associated with the values of $\underline{x}_l$, where

$$\beta_l = g(\underline{x}_l|\beta_l) = \Pr(\underline{\Theta} = \underline{x}_l|\beta_l),$$

and $\sum_l \beta_l = 1$ and Equation 6 becomes

$$\begin{aligned}
\Pr(\underline{U} = \underline{u}|\boldsymbol{\phi}, \underline{\beta}) &= \sum_{l=1}^{q} f(\underline{u}, \underline{x}_l|\boldsymbol{\phi}, \underline{\beta}) \\
&= \sum_{l=1}^{q} \Pr(\underline{U} = \underline{u}|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}) \Pr(\underline{\Theta} = \underline{x}_l|\beta_l) \\
&= \sum_{l=1}^{q} \Pr(\underline{U} = \underline{u}|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi})\beta_l, \qquad (7)
\end{aligned}$$

where $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$.

For (marginal) ML estimation of parameters $\boldsymbol{\phi}$ and $\underline{\beta}$, the (marginal) likelihood function is the probability of observing a response data matrix $\boldsymbol{u}$ which is, for $\{r_1, r_2, \ldots, r_s\}$ being multinomially distributed,

$$L \equiv \Pr(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\phi}, \underline{\boldsymbol{\beta}}) = \frac{n!}{\prod_{w=1}^{s} r_w!} \prod_{w=1}^{s} [\Pr(\underline{U} = \underline{u}_w|\boldsymbol{\phi}, \underline{\beta})]^{r_w}. \qquad (8)$$

The log-likelihood is, up to a constant factor independent of $\boldsymbol{\phi}$ and $\underline{\beta}$,

$$\ln L = \sum_{w=1}^{s} r_w \ln \Pr(\underline{U} = \underline{u}_w|\boldsymbol{\phi}, \underline{\beta}). \qquad (9)$$

The MML estimates (MMLEs) are obtained by solving the likelihood equations,

$$\frac{\partial \ln L}{\partial \phi_{it}} = 0, \tag{10}$$

and

$$\frac{\partial \ln L}{\partial \beta_{\nu}} = 0, \tag{11}$$

where ϕ_{it} denotes an element of the item parameter matrix $\boldsymbol{\phi}$, and β_{ν} an element of the θ-distribution parameter vector $\underline{\beta}$.

The first derivatives of $\ln L$ with respect to an item parameter and a θ-distribution parameter are, respectively,

$$\begin{aligned}
\frac{\partial \ln L}{\partial \phi_{it}} &= \sum_{w=1}^{s} \frac{r_w}{\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})} \frac{\partial \Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})}{\partial \phi_{it}} \\
&= \sum_{w=1}^{s} \int \frac{r_w \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) g(\underline{\theta} | \underline{\beta})}{\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})} \left[\frac{u_{wi} - p_i(\underline{\theta})}{p_i(\underline{\theta})(1 - p_i(\underline{\theta}))} \right] \frac{\partial p_i(\underline{\theta})}{\partial \phi_{it}} \, d\underline{\theta},
\end{aligned} \tag{12}$$

and

$$\begin{aligned}
\frac{\partial \ln L}{\partial \beta_{w}} &= \sum_{w=1}^{s} \frac{r_w}{\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})} \frac{\partial \Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})}{\partial \beta_{\nu}} \\
&= \sum_{w=1}^{s} \int \frac{r_w \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) g(\underline{\theta} | \underline{\beta})}{\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})} \frac{\partial \ln g(\underline{\theta} | \underline{\beta})}{\partial \beta_{\nu}} \, d\underline{\theta},
\end{aligned} \tag{13}$$

where $p_i(\underline{\theta}) = \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i)$, and in Equation 12,

$$\begin{aligned}
&\frac{\partial \Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})}{\partial \phi_{it}} \\
&\quad = \frac{\partial \int \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) g(\underline{\theta} | \underline{\beta}) \, d\underline{\theta}}{\partial \phi_{it}}
\end{aligned}$$

$$= \int g(\underline{\theta}|\underline{\beta}) \frac{\partial \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi})}{\partial \phi_{it}} \, d\underline{\theta}$$

$$= \int g(\underline{\theta}|\underline{\beta}) \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) \frac{\partial \ln \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi})}{\partial \phi_{it}} \, d\underline{\theta}, \tag{14}$$

also,

$$\begin{aligned} \frac{\partial \ln \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi})}{\partial \phi_{it}} &= \frac{\partial \sum_{h=1}^{k} [u_{wh} \ln p_h(\underline{\theta}) + (1 - u_{wh}) \ln(1 - p_h(\underline{\theta}))]}{\partial \phi_{it}} \\ &= \frac{u_{wi}}{p_i(\underline{\theta})} \frac{\partial p_i(\underline{\theta})}{\partial \phi_{it}} + \frac{1 - u_{wi}}{1 - p_i(\underline{\theta})} \frac{\partial (1 - p_i(\underline{\theta}))}{\partial \phi_{it}} \\ &= \frac{u_{wi} - p_i(\underline{\theta})}{p_i(\underline{\theta})(1 - p_i(\underline{\theta}))} \frac{\partial p_i(\underline{\theta})}{\partial \phi_{it}}, \end{aligned} \tag{15}$$

and in Equation 13,

$$\begin{aligned} \frac{\partial \Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}, \underline{\beta})}{\partial \beta_\nu} &= \frac{\partial \int \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) g(\underline{\theta}|\underline{\beta}) \, d\underline{\theta}}{\partial \beta_\nu} \\ &= \int \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) \frac{\partial g(\underline{\theta}|\underline{\beta})}{\partial \beta_\nu} \, d\underline{\theta} \\ &= \int \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) g(\underline{\theta}|\underline{\beta}) \frac{\partial \ln g(\underline{\theta}|\underline{\beta})}{\partial \beta_\nu} \, d\underline{\theta}. \end{aligned} \tag{16}$$

Tsutakawa (1984, 1985) compared Equations 12 (MML estimation) with 40 (EM algorithm) using the two-parameter logistic model.

For the (unidimensional) two-parameter normal ogive model and the standard normal distribution of Θ, Bock and Lieberman (1970) suggested using the Newton-Raphson method and Gauss-Hermite quadrature to solve the likelihood equations given in Equation 10. That is, Equation 12 (in which $\underline{\Theta}$ is replaced by Θ and the known $\underline{\beta}$ is dropped), is approximated as

$$\frac{\partial \ln L}{\partial \phi_{it}} \approx \sum_{l=1}^{q} \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w | \Theta = x_l, \boldsymbol{\phi}) A(x_l)}{\tilde{P}(\underline{U} = \underline{u}_w | \boldsymbol{\phi})} \left[\frac{u_{wi} - p_i(x_l)}{p_i(x_l)(1 - p_i(x_l))} \right] \frac{\partial p_i(x_l)}{\partial \phi_{it}}, \tag{17}$$

where $\{x_l\colon l = 1, \ldots, q\}$ are tabled nodes on the θ scale and $\{A(x_l)\}$ are associated weights in the Gauss-Hermite quadrature with appropriate transformation (see Bock & Lieberman, 1970), and

$$\tilde{P}(\underline{U} = \underline{u}_w | \phi) = \sum_{l=1}^{q} \Pr(\underline{U} = \underline{u}_w | \Theta = x_l, \phi) A(x_l).$$

They obtained stable estimates of item parameters for five items. However, "computational difficulties limit the approach to not more than 10 or 12 items (Bock & Lieberman, 1970, p.180)." The computational demand arises through the need, in each iteration, for the inverse of a $2k \times 2k$ information matrix. Moreover, each element of the information matrix is the sum of 2^k terms, regardless of the value of s. And, each of 2^k terms involves evaluation of the integration.

Because of these problems, Bock and Aitkin (1981) reformulated Equation 17 and presented a procedure (called the Bock-Aitkin procedure in this paper) of the EM-algorithm type. Harwell et al. (1988) stated that this reformulation "produces consistent item parameter estimates" (p.254). (Note that Mislevy and Stocking (1989, p.59) stated that the MML approach "yields consistent estimates of item parameters", conditional on the veracity of the IRT model.) Moreover, Bock and Aitkin (1981, p.444) mentioned that by this reformulation, "a computationally feasible solution is possible for both small and large numbers of items," and freed "from arbitrary assumptions about the distribution of ability in the population effectively sampled."

Bock-Aitkin Procedure

In the Bock-Aitkin procedure (Bock & Aitkin, 1981), the θ distribution is fixed at each estimation iteration. (Mislevy and Bock (1985) extended the procedure so that the distribution

may vary for each iteration.) Assume that $\underline{\Theta}$ has the standard multivariate normal density $g(\underline{\theta})$, then the distribution parameters are known, and Equation 7 becomes

$$\ln L = \sum_{w=1}^{s} r_w \ln \Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}), \tag{18}$$

where $\Pr(\underline{U} = \underline{u}_w|\boldsymbol{\phi}) = \int \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}) g(\underline{\theta})\, d\underline{\theta}$. The approximation in Equation 17 of the first derivative of $\ln L$ with respect to an item parameter is reformulated, in terms of the $m \times 1$ vector $\underline{\Theta}$, as

$$\frac{\partial \ln L}{\partial \phi_{it}} \approx \sum_{l=1}^{q} \frac{\overline{r}_{li} - \overline{n}_l p_i(\underline{x}_l)}{p_i(\underline{x}_l)(1 - p_i(\underline{x}_l))} \frac{\partial p_i(\underline{x}_l)}{\partial \phi_{it}}, \tag{19}$$

where m-dimensional Gauss-Hermite quadrature with appropriate transformation (see Bock & Aitkin, 1981; Bock & Lieberman, 1970) is employed to evaluate the integral (note that by using the standard multivariate normal distribution, the abilities are assumed independent), $\{\underline{x}_l = (x_{l1}, x_{l2}, \ldots, x_{lm})' : l = 1, \ldots, q\}$, are quadrature nodes on the m-dimensional space, and $\{A(\underline{x}_l) = A(x_{l1})A(x_{l2}) \cdots A(x_{lm})\}$ are the associated weights, and

$$\overline{n}_l = \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}) A(\underline{x}_l)}{\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi})}, \tag{20}$$

$$\overline{r}_{li} = \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}) A(\underline{x}_l) u_{wi}}{\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi})}, \tag{21}$$

$$\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi}) = \sum_{l=1}^{q} \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}) A(\underline{x}_l). \tag{22}$$

A Bock-Aitkin procedure (the BA1 procedure) can then be briefly described as follows:

1. Using provisional values of item parameters and pre-specified nodes, compute $\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi})$ for response pattern w, $w = 1, 2, \ldots, s$.

2. (E-step) Compute $\overline{n}_l$ and $\overline{r}_{li}$ for each item i and each node l.

3. (M-step) Obtain the item parameter estimates by solving the system of likelihood equations for all item parameters via Equation 19, where $\overline{n}_l$ and $\overline{r}_{li}$ are treated as constants.

4. Go back to step 1 unless the convergence criterion is reached.

A comparison between the second line of Equation 13 and the second line of Equation 12 leads to an extension of the Bock-Aitkin procedure where the θ-distribution parameters can be estimated along with item parameter estimation. Note that Bock and Aitkin (1981) gave heuristic meanings of $\overline{n}_l$ and $\overline{r}_{li}$, by treating the quadrature points and weights $\{\underline{x}_l, A(\underline{x}_l)\}$ as a discrete distribution approximation to the standard normal distribution (because $\sum A(\underline{x}_l) = 1$, $A(\underline{x}_l)$ can be interpreted as the probability at point $\underline{x}_l$). However, the meanings will be obvious from an EM algorithm perspective discussed later. Bock and Aitkin also proposed to use a discrete distribution (or discrete representation) on a finite number of equally spaced points to approximate a continuous distribution of ability which has finite mean and variance (also see Mislevy & Bock, 1985, 1993). Bock and Aitkin (1981, pp.449-450) described a process using this discrete approximation which can give a nonparametric estimate of the continuous (normal or any other) distribution of ability. It should be noted that this approximation is not for numerical quadrature to compute the integral, although the discrete points and probabilities are labeled as quadrature points and weights in some literature. In this paper, this discrete representation is treated as if the underlying θ distribution is discrete in item parameter estimation.

EM Algorithm

The name EM refers to the two steps of the algorithm: the expectation (E) step and the maximization (M) step. The EM algorithm (Dempster et al., 1977) is a general iterative process of ML estimation for incomplete data. The complete data contains two parts: one is incomplete data, and another can be referred to as missing data that can be missing values, unobservable data, or parameters.

Suppose there is a model for the complete data $\boldsymbol{Y}$, with an associated probability or density function $f(\boldsymbol{y}|\varphi)$ where φ is the set of unknown parameters to be estimated. Let $\boldsymbol{Y}_{\text{obs}}$ represent the observed part of $\boldsymbol{Y}$, and $\boldsymbol{Y}_{\text{mis}}$ denote the missing part. The complete data is given as $\boldsymbol{Y}= (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$. The EM algorithm finds the MLE, $\hat{\varphi}_{\text{MLE}}$, of φ that maximizes the likelihood function based on the observed data $\boldsymbol{Y}_{\text{obs}}$ which is

$$L(\varphi|\boldsymbol{y}_{\text{obs}}) = f(\boldsymbol{y}_{\text{obs}}|\varphi) = \int f(\boldsymbol{y}|\varphi)\, d\boldsymbol{y}_{\text{mis}} \,. \tag{23}$$

However, the EM algorithm uses the complete data likelihood $f(\boldsymbol{y}|\varphi)$ instead of $f(\boldsymbol{y}_{\text{obs}}|\varphi)$.

When the complete data density (or probability) function $f(\boldsymbol{y}|\varphi)$ has an exponential-family form, it is easy to implement the EM algorithm (Dempster et al., 1977). However, the "mixed effects" IRT models (i.e., $\boldsymbol{\theta}$ is random but $\boldsymbol{\phi}$ is fixed) generally do not possess the properties of the exponential families (Ridgon & Tsutakawa, 1983). Then a more general form of the EM algorithm can be used which is as follows: First, initial values $\varphi^{(0)}$ are established. Let $\varphi^{(t)}$ be the estimate of φ at the tth iteration (or referred as an EM cycle). Then the (t+1)th EM cycle can be expressed in two steps (Dempster et al., 1977):

E-step: Compute $Q(\varphi|\varphi^{(t)})$ where

$$Q(\varphi|\varphi^{(t)}) = E[\ln f(\boldsymbol{Y}|\varphi)|\boldsymbol{y}_{\text{obs}}, \varphi^{(t)}] = \int \ln f(\boldsymbol{y}|\varphi) f(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \varphi^{(t)})\, d\boldsymbol{y}_{\text{mis}}, \tag{24}$$

M-step: Find $\varphi^{(t+1)}$ that maximizes $Q(\varphi|\varphi^{(t)})$.

In some cases maximization is hard to attain, then a generalized EM (GEM) algorithm can be used in which the M-step is to find $\varphi^{(t+1)}$ such that

$$Q(\varphi^{(t+1)}|\varphi^{(t)}) \geq Q(\varphi^{(t)}|\varphi^{(t)})\,. \tag{25}$$

The EM algorithm converges reliably in the sense that each EM cycle increases the likelihood $L(\varphi|\boldsymbol{y}_{\text{obs}})$ and if $\ln L(\varphi|\boldsymbol{y}_{\text{obs}})$ is bounded, then under certain conditions, the sequence $\ln L(\varphi^{(t)}|\boldsymbol{y}_{\text{obs}})$ converges to a stationary value (i.e., on which the first derivative is zero, but not necessarily a global/local maximum) of $\ln L(\varphi|\boldsymbol{y}_{\text{obs}})$ (see Dempster et al., 1977). However, it is well known that the EM algorithm is usually slow to converge, the rate of convergence can be painfully slow if the amount of missing data is large. Moreover, the convergence of the sequence $\ln L(\varphi^{(t)}|\boldsymbol{y}_{\text{obs}})$ by the EM or GEM algorithm may not imply the convergence of $\varphi^{(t)}$ (Wu, 1983).

How the EM Algorithm Is Applied in IRT Settings

Consider n examinees for a test of k items. The response data matrix $\boldsymbol{U}$ is the incomplete data (i.e., $\boldsymbol{Y}_{\text{obs}}$). With the ability matrix $\boldsymbol{\Theta}$ as $\boldsymbol{Y}_{\text{mis}}$, $\boldsymbol{Y} = (\boldsymbol{U}, \boldsymbol{\Theta})$ as in Equation 23 represents the complete data. Assume that $\{\underline{\Theta}_1, \underline{\Theta}_2, \ldots, \underline{\Theta}_n\}$ represents a random sample from a distribution whose density is $g(\underline{\theta}|\underline{\beta})$. Under the modeling of Equation 6, and by local independence, the

joint distribution of $\boldsymbol{Y} = (\boldsymbol{U}, \boldsymbol{\Theta})$ with parameters $\boldsymbol{\varphi} = (\boldsymbol{\phi}, \underline{\beta})$ can be expressed as,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\varphi}) = f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{\phi}, \underline{\beta}) &= \Pr(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{\phi}) g(\boldsymbol{\theta}|\underline{\beta}) \\ &= \prod_{i=1}^{k} \prod_{j=1}^{n} \Pr(U_{ji} = u_{ji}|\underline{\Theta}_j = \underline{\theta}_j, \underline{\phi}_i) \prod_{j=1}^{n} g(\underline{\theta}_j|\underline{\beta}). \end{aligned} \tag{26}$$

Using the EM algorithm (also see Rigdon & Tsutakawa, 1983; Tsutakawa, 1985), the $(t+1)$th iteration chooses $\boldsymbol{\varphi}^{(t+1)}$ that maximizes the conditional expectation over $\boldsymbol{\Theta}$,

$$\begin{aligned} &Q(\boldsymbol{\varphi}|\boldsymbol{\varphi}^{(t)}) \\ &= E[\ln f(\boldsymbol{Y}|\boldsymbol{\varphi})|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\varphi}^{(t)}] \\ &= E[\ln f(\boldsymbol{u}, \boldsymbol{\Theta}|\boldsymbol{\phi}, \underline{\beta})|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}] \\ &= \sum_{j=1}^{n} E[\ln g(\underline{\Theta}_j|\underline{\beta})|\underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}] + \sum_{i=1}^{k}\sum_{j=1}^{n} E[\ln \Pr(U_{ji} = u_{ji}|\underline{\Theta}_j, \underline{\phi}_i)|\underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}] \\ &= \sum_{j=1}^{n} E[\ln g(\underline{\Theta}|\underline{\beta})|\underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}] + \sum_{i=1}^{k}\sum_{j=1}^{n} E[\ln \Pr(U_i = u_{ji}|\underline{\Theta}, \underline{\phi}_i)|\underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}]. \end{aligned} \tag{27}$$

The Equation 26 is used in going from the third to fourth line of Equation 27. In going from line 4 to line 5, the subscript j of $\underline{\Theta}_j$ can be dropped, because all $\underline{\Theta}_j$ are identically distributed. Then, starting with $\boldsymbol{\varphi}^{(0)}$, the sequence of $\boldsymbol{\varphi}^{(1)}, \boldsymbol{\varphi}^{(2)}, \boldsymbol{\varphi}^{(3)}, \cdots$ is found. If this sequence converges to $\boldsymbol{\varphi}^*$, then $\boldsymbol{\varphi}^*$ is the maximum likelihood estimate $\hat{\boldsymbol{\varphi}}_{\text{MLE}}$ of $\boldsymbol{\varphi}$, under regularity conditions. Note that theoretically the θ distribution in Equation 27 can be more general, not restricted to the discrete or continuous type.

The maximization of Equation 27 can be done by maximizing each of its terms, and each

term only involves one kind of parameter, hence both $\underline{\beta}^{(t+1)}$ and $\phi^{(t+1)}$ can be found separately. Moreover, the maximization of the second term with respect to ϕ may be performed for each $\underline{\phi}_i$ in the inner summation over j through the usual optimization procedure, for example, the Newton-Raphson method. That is, $\underline{\phi}_i^{(t+1)}$ in $\phi^{(t+1)}$ is chosen to maximize

$$\sum_{j=1}^{n} E[\ln \Pr(U_i = u_{ji}|\underline{\Theta}, \underline{\phi}_i)|\underline{u}_j, \phi^{(t)}, \underline{\beta}^{(t)}] \tag{28}$$

with respect to $\underline{\phi}_i$ for each item i.

Under regularity conditions, the convergent estimate φ^* of the EM algorithm also maximizes

$$\prod_{j=1}^{n} \Pr(\underline{U}_j = \underline{u}_j|\phi, \underline{\beta}), \tag{29}$$

that is proportional to the marginal likelihood function (see Equation 8), so φ^* is the MMLE of φ. As mentioned in the previous section, it should be noted that convergence of the EM algorithm is not guaranteed. Even if the EM algorithm converges to the EM estimate φ^*, the EM estimate still may not be the MMLE (see Tsutakawa, 1985, and Lewis, 1985, for further discussion of the convergence of the EM algorithm and the MMLE).

The Relationship Between the Bock-Aitkin Procedure and the EM Algorithm

Discrete θ

The demonstration is from a probit analysis or bioassay solution perspective (Bock & Aitkin, 1981; Harwell et al., 1988). For a discrete $\underline{\theta}$, to show the Bock-Aitkin procedure is an instance of the EM algorithm (R. K. Tsutakawa, personal communication, September 23,

1993), it is assumed that all possible ability levels are finite, $\{\underline{\theta}_l = \underline{x}_l = (x_{l1}, x_{l2}, \cdots, x_{lm})'$: $l = 1, 2, \ldots, q\}$, with a known distribution of probabilities $\beta_1 = A(\underline{x}_1)$, $\beta_2 = A(\underline{x}_2)$, $\cdots$, and $\beta_q = A(\underline{x}_q)$ with $\sum \beta_l = 1$. (By observing Equation 27, that is, from an EM algorithm perspective, the assumption of known probabilities can be relaxed as discussed later, also see Bock & Aitkin, 1981; Mislevy & Bock, 1985; and Woodruff & Hanson, 1996). Examinees are assumed to be grouped into homogeneous groups, each with a distinct ability level $\underline{x}_l$, so that the complete data likelihood function has a form that is similar to Equation 26 where each $g(\underline{x}_j|\underline{\beta})$ is replaced by $A(\underline{x}_j)$, that is,

$$
\begin{aligned}
& f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{\phi}, \underline{\beta}) \\
&= \Pr(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{\phi}) \Pr(\boldsymbol{\Theta} = \boldsymbol{\theta}|\underline{\beta}) \\
&= \prod_{i=1}^{k} \prod_{j=1}^{n} \Pr(U_{ji} = u_{ji}|\underline{\Theta}_j = \underline{\theta}_j, \underline{\phi}_i) \prod_{j=1}^{n} A(\underline{\theta}_j) \\
&= \prod_{l=1}^{q} \prod_{i=1}^{k} \Pr(U_i = 1|\underline{\Theta} = \underline{x}_l, \underline{\phi}_i)^{r_{li}} (1 - \Pr(U_i = 1|\underline{\Theta} = \underline{x}_l, \underline{\phi}_i))^{n_l - r_{li}} \prod_{l=1}^{q} A(\underline{x}_l)^{n_l} \\
&= f(\underline{n}, \boldsymbol{r}|\boldsymbol{\phi}, \underline{\beta}), \qquad (30)
\end{aligned}
$$

where $\underline{n} = (n_1, n_2, \ldots, n_q)'$, $n_l = n_l(\boldsymbol{\theta}) = \sum_{j=1}^{n} 1_{\{\underline{\theta}_j = \underline{x}_l\}}$ denotes the number of examinees with ability $\underline{x}_l$, and $\sum_{l=1}^{q} n_l = n$, and $1_{\{\underline{\theta}_j = \underline{x}_l\}} = 1$ if examinee j's $\underline{\theta}$ value is $\underline{x}_l$, 0 otherwise; $\boldsymbol{r} = \{r_{li}\}, r_{li} = r_{li}(\boldsymbol{u}, \boldsymbol{\theta}) = \sum_{j=1}^{n} 1_{\{\underline{\theta}_j = \underline{x}_l\} \cap \{u_{ji}=1\}} = \sum_{j=1}^{n} u_{ji} 1_{\{\underline{\theta}_j = \underline{x}_l\}}$, denotes the number of examinees with ability $\underline{x}_l$ who answer item i correctly, and $\sum_{l=1}^{q} r_{li} = \sum_{j=1}^{n} u_{ji}$. Moreover,

$$
\ln f(\underline{n}, \boldsymbol{r}|\boldsymbol{\phi}) = \left\{ \sum_{i=1}^{k} \sum_{l=1}^{q} [r_{li} \ln p_i(\underline{x}_l) + (n_l - r_{li}) \ln(1 - p_i(\underline{x}_l))] \right\} + \sum_{l=1}^{q} n_l \ln A(\underline{x}_l). \qquad (31)
$$

$\underline{\beta}$ has been dropped in Equation 31, because it is known. For the complete data (i.e., assuming that the ability of each individual is known), $\{\underline{n}, \boldsymbol{r}\}$ is observed. For the incomplete data, because each randomly selected examinee's ability $\underline{\theta}$ is unknown, $\{n_l\}$ and $\{r_{li}\}$ are unobserved.

Using the EM algorithm, given some provisional value of $\boldsymbol{\phi}^{(t)}$, we want to find $\boldsymbol{\phi}^{(t+1)}$ which maximizes

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) = E[\ln f(\boldsymbol{u}, \boldsymbol{\Theta}|\boldsymbol{\phi})|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}] = E[\ln f(\underline{N}, \boldsymbol{R}|\boldsymbol{\phi})|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}],$$

that is, by Equation 31, maximizes

$$E\left[\sum_{i=1}^{k}\sum_{l=1}^{q} R_{li}\left[\ln \Pr(U_i = 1|\underline{\Theta} = \underline{x}_l, \underline{\phi}_i)\right] + \right.$$

$$\left. (N_l - R_{li})\ln(1 - \Pr(U_i = 1|\underline{\Theta} = \underline{x}_l, \underline{\phi}_i))|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}\right]$$

$$= \sum_{i=1}^{k}\sum_{l=1}^{q}\left[r_{li}^{*(t)} \ln p_i(\underline{x}_l) + (n_l^{*(t)} - r_{li}^{*(t)})\ln(1 - p_i(\underline{x}_l))\right], \tag{32}$$

where $\underline{N}$ and $\boldsymbol{R}$ are random, and

$$\begin{aligned} n_l^{*(t)} &= E[N_l|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}] = E\left[\sum_{j=1}^{n} 1_{\{\underline{\Theta}_j = \underline{x}_l\}}|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}\right] \\ &= \sum_{j=1}^{n} \Pr(\underline{\Theta}_j = \underline{x}_l|\underline{u}_j, \boldsymbol{\phi}^{(t)}) = \sum_{w=1}^{s} r_w \Pr(\underline{\Theta} = \underline{x}_l|\underline{u}_w, \boldsymbol{\phi}^{(t)}) \\ &= \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}^{(t)})A(\underline{x}_l)}{\Pr(\underline{U} = \underline{u}_w|\boldsymbol{\phi}^{(t)})}, \end{aligned} \tag{33}$$

$$\begin{aligned} r_{li}^{*(t)} &= E[R_{li}|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}] = E\left[\sum_{j=1}^{n} u_{ji} 1_{\{\underline{\Theta}_j = \underline{x}_l\}}|\boldsymbol{u}, \boldsymbol{\phi}^{(t)}\right] \\ &= \sum_{j=1}^{n} u_{ji} \Pr(\underline{\Theta}_j = \underline{x}_l|\underline{u}_j, \boldsymbol{\phi}^{(t)}) = \sum_{w=1}^{s} u_{wi} r_w \Pr(\underline{\Theta} = \underline{x}_l|\underline{u}_w, \boldsymbol{\phi}^{(t)}) \end{aligned}$$

$$= \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}^{(t)}) A(\underline{x}_l)}{\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}^{(t)})} u_{wi}\,, \tag{34}$$

$$\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}^{(t)}) = \sum_{l=1}^{q} \Pr(\underline{U} = \underline{u}_w | \underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}^{(t)}) A(\underline{x}_l). \tag{35}$$

In the M-step of the $(t + 1)$th EM cycle, the maximization of $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})$ with respect to $\boldsymbol{\phi} = (\underline{\phi}_1, \cdots, \underline{\phi}_k)$ reduces to maximizing the inner summation of Equation 32 for each item i, separately. For each item, the (complete data) likelihood equations have the form

$$\sum_{l=1}^{q} \frac{r_{li}^{*(t)} - n_l^{*(t)} p_i(\underline{x}_l)}{p_i(\underline{x}_l)(1 - p_i(\underline{x}_l))} \frac{\partial p_i(\underline{x}_l)}{\partial \phi_{i\nu}} = 0, \tag{36}$$

where $\phi_{i\nu}$ is an element of $\underline{\phi}_i$. The maximization of the M-step, then, gives $\boldsymbol{\phi}^{(t+1)}$.

For the unidimensional case, comparing Equations 33 with 20, 34 with 21, 35 with 22, there are, without the superscript (t),

$$\Pr(\underline{U} = \underline{u}_w | \boldsymbol{\phi}) = \tilde{P}(\underline{U} = \underline{u}_w | \boldsymbol{\phi}),$$

$$r_{li}^* = \overline{r}_{li},$$

and

$$n_l^* = \overline{n}_l.$$

Hence, $\overline{r}_{li}$ can be interpreted as the conditional expected frequency of correct response to item i at $\underline{\theta} = \underline{x}_l$, and $\overline{n}_l$ as the conditional expected sample size at level $\underline{x}_l$ (see also Bock & Aitkin, 1981). Comparing Equations 19 and 36, a relationship is observed that when the ability values are finite with known distribution (e.g., if ability levels are assumed to be equal to those nodes in the Gauss-Hermite quadrature used in Equation 19, and probabilities of ability levels are assumed

to be equal to the weights associated with nodes), then the Bock-Aitkin procedure (or the BA1 procedure) and the EM algorithm give the same result.

Note that instead of Equation 30, Harwell et al. (1988) showed the relationship using

$$\Pr(\underline{N} = \underline{n}, \boldsymbol{R} = \boldsymbol{r} | \boldsymbol{\phi}, \underline{\beta})$$

$$= \Pr(\boldsymbol{R} = \boldsymbol{r} | \underline{N} = \underline{n}, \boldsymbol{\phi}, \underline{\beta}) \Pr(\underline{N} = \underline{n} | \boldsymbol{\phi}, \underline{\beta})$$

$$= \left\{ \prod_{l=1}^{q} \prod_{i=1}^{k} \frac{n_l!}{(n_l - r_{li})! r_{li}!} \Pr(U_i = 1 | \underline{\Theta} = \underline{x}_l, \underline{\phi}_i)^{r_{li}} (1 - \Pr(U_i = 1 | \underline{\Theta} = \underline{x}_l, \underline{\phi}_i))^{n_l - r_{li}} \right\}$$

$$\cdot \left\{ \frac{n!}{n_1! \cdots n_q!} \prod_{l=1}^{q} A(\underline{x}_l)^{n_l} \right\}$$

$$= \left\{ \left[\prod_{l=1}^{q} \prod_{i=1}^{k} \frac{n_l!}{(n_l - r_{li})! r_{li}!} \right] \frac{n!}{n_1! \cdots n_q!} \right\} f(\underline{n}, \boldsymbol{r} | \boldsymbol{\phi}, \underline{\beta}),$$

which is proportional to $f(\underline{n}, \boldsymbol{r} | \boldsymbol{\phi}, \underline{\beta})$. The constant of proportionality is positive and does not depend on model parameters $\boldsymbol{\phi}$ and $\underline{\beta}$.

Continuous θ

An EM algorithm perspective is taken. Although a continuous θ scale is used for demonstration, it is also true for the discrete case (the integration is changed to the summation). With a continuous θ scale, following the EM algorithm, for each item i Equation 28 becomes

$$A_i^{(t)} = \sum_{j=1}^{n} E[\ln \Pr(U_i = u_{ji} | \underline{\Theta}, \underline{\phi}_i) | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}]$$

$$= \sum_{j=1}^{n} \int \left[\ln \Pr(U_i = u_{ji} | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i) \right] \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}) \, d\underline{\theta}$$

$$= \int \sum_{j=1}^{n} \left[u_{ji} \ln \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i) + (1 - u_{ji}) \ln(1 - \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i)) \right]$$

$$\cdot \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)})\, d\underline{\theta}$$

$$= \int \ln \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i) \sum_{j=1}^{n} \left[u_{ji} \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}) \right] d\underline{\theta}$$

$$+ \int \ln(1 - \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i)) \sum_{j=1}^{n} \left[(1 - u_{ji}) \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}) \right] d\underline{\theta}$$

$$= \int \ln \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i) r_i(\underline{\theta})^{(t)}$$

$$+ \ln(1 - \Pr(U_i = 1 | \underline{\Theta} = \underline{\theta}, \underline{\phi}_i))(n(\underline{\theta})^{(t)} - r_i(\underline{\theta})^{(t)})\, d\underline{\theta}, \tag{37}$$

where

$$n(\underline{\theta})^{(t)} = \sum_{j=1}^{n} \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}), \tag{38}$$

$$r_i(\underline{\theta})^{(t)} = \sum_{j=1}^{n} u_{ji} \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}), \tag{39}$$

and the conditional density function $\pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)})$ is given by

$$\begin{aligned} \pi(\underline{\theta} | \underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}) &= \frac{\Pr(\underline{U} = \underline{u}_j | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}^{(t)}) g(\underline{\theta} | \underline{\beta}^{(t)})}{\Pr(\underline{U} = \underline{u}_j | \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)})} \\ &= \frac{\Pr(\underline{U} = \underline{u}_j | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}^{(t)}) g(\underline{\theta} | \underline{\beta}^{(t)})}{\int \Pr(\underline{U} = \underline{u}_j | \underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}^{(t)}) g(\underline{\theta} | \underline{\beta}^{(t)})\, d\underline{\theta}}. \end{aligned}$$

In going from line 2 to line 3 of Equation 37, Equation 2 is used.

The first derivative of $A_i^{(t)}$ with respect to ϕ_{iv} is

$$\frac{\partial A_i^{(t)}}{\partial \phi_{iv}} =$$

$$\int \frac{r_i(\underline{\theta})^{(t)} - n(\underline{\theta})^{(t)} \Pr(U_i = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}_i)}{\Pr(U_i = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}_i)(1 - \Pr(U_i = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}_i))} \frac{\partial \Pr(U_i = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}_i)}{\partial \phi_{iv}} d\underline{\theta}. \quad (40)$$

Let $B^{(t)}$ be the first term of the right hand side of line 5 of Equation 27,

$$\begin{aligned} B^{(t)} &= \sum_{j=1}^{n} E[\ln g(\underline{\Theta}|\underline{\beta})|\underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)}] \\ &= \int \sum_{j=1}^{n} \left[\ln g(\underline{\theta}|\underline{\beta})\pi(\underline{\theta}|\underline{u}_j, \boldsymbol{\phi}^{(t)}, \underline{\beta}^{(t)})\right] d\underline{\theta} \\ &= \int (\ln g(\underline{\theta}|\underline{\beta}))n(\underline{\theta})^{(t)} d\underline{\theta}. \end{aligned}$$

The first derivative of $B^{(t)}$ with respect to β_ν, an element of $\underline{\beta}$, is

$$\frac{\partial B^{(t)}}{\partial \beta_\nu} = \int \frac{n(\underline{\theta})^{(t)}}{g(\underline{\theta}|\underline{\beta})} \frac{\partial g(\underline{\theta}|\underline{\beta})}{\partial \beta_\nu} d\underline{\theta}. \quad (41)$$

Note that Equation 40 (from the EM algorithm) and Equation 19 (from the Bock-Aitkin procedure) are similar; they are the same, if $\underline{\Theta}$ has the standard multivariate normal distribution and the m-dimensional Gauss-Hermite quadrature is used (also compared with Equation 12 in Bock and Aitkin's 1981 paper).

Moreover, the θ-distribution parameters can be estimated along with item parameters by simultaneously solving EM-likelihood equations 40 and 41. For example, a Bock-Aitkin procedure (BAHAF) is described as follows, with the standard multivariate normal distribution of $\underline{\Theta}$: Using multidimensional Gauss-Hermite quadrature,

$$\begin{aligned} \frac{\partial A_i^{(t)}}{\partial \phi_{iv}} &= \int \frac{\left[\sum_{j=1}^{n} \frac{u_{ji} p_j^{(t)}(\underline{\theta})}{\int p_j^{(t)}(\underline{\theta}) g(\underline{\theta}|\underline{\beta}^{(t)})\, d\underline{\theta}} - \sum_{j=1}^{n} \frac{p_j^{(t)}(\underline{\theta}) p_j(\underline{\theta})}{\int p_j^{(t)}(\underline{\theta}) g(\underline{\theta}|\underline{\beta}^{(t)})\, d\underline{\theta}}\right] g(\underline{\theta}|\underline{\beta}^{(t)})}{p_i(\underline{\theta})(1 - p_i(\underline{\theta}))} \frac{\partial p_i(\underline{\theta})}{\partial \phi_{iv}} d\underline{\theta} \\ &\approx \sum_{l=1}^{q} \frac{\overline{r}_{li}^{(t)} - \overline{n}_l^{(t)} p_i(\underline{x}_l)}{p_i(\underline{x}_l)(1 - p_i(\underline{x}_l))} \frac{\partial p_i(\underline{x}_l)}{\partial \phi_{iv}}, \end{aligned} \quad (42)$$

and

$$\frac{\partial B^{(t)}}{\partial \beta_\nu} = \int \frac{\left[\sum_{j=1}^{n} \frac{p_j^{(t)}(\underline{\theta})}{\int p_j^{(t)}(\underline{\theta}) g(\underline{\theta}|\underline{\beta}^{(t)})\, d\underline{\theta}}\right] g(\underline{\theta}|\underline{\beta}^{(t)})}{g(\underline{\theta}|\underline{\beta})} \frac{\partial g(\underline{\theta}|\underline{\beta})}{\partial \beta_\nu}\, d\underline{\theta}$$

$$\approx \sum_{l=1}^{q} \frac{\overline{n}_l^{(t)}}{g(\underline{x}_l|\underline{\beta})} \frac{\partial g(\underline{x}_l|\underline{\beta})}{\partial \beta_\nu}, \tag{43}$$

where $p_j^{(t)}(\underline{\theta}) = \Pr(\underline{U} = \underline{u}_j|\underline{\Theta} = \underline{\theta}, \boldsymbol{\phi}^{(t)})$, and $p_i(\underline{\theta}) = \Pr(U_i = 1|\underline{\Theta} = \underline{\theta}, \underline{\phi}_i)$. As Equations 20 and 21,

$$\overline{n}_l^{(t)} = \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}^{(t)}) A(\underline{x}_l)}{\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi}^{(t)})}, \tag{44}$$

$$\overline{r}_{li}^{(t)} = \sum_{w=1}^{s} \frac{r_w \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}^{(t)}) A(\underline{x}_l) u_{wi}}{\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi}^{(t)})}, \tag{45}$$

$$\tilde{P}(\underline{U} = \underline{u}_w|\boldsymbol{\phi}^{(t)}) = \sum_{l=1}^{q} \Pr(\underline{U} = \underline{u}_w|\underline{\Theta} = \underline{x}_l, \boldsymbol{\phi}^{(t)}) A(\underline{x}_l)\,. \tag{46}$$

1. Start with pre-specified nodes and provisional values of item parameters and distribution parameters.

2. For each EM cycle, compute $\overline{n}_l^{(t)}$ and $\overline{r}_{li}^{(t)}$ of Equations 44 and 45, respectively, for each item i and each quadrature node in the E-step.

3. In the M-step, use the Newton-Raphson iteration to solve EM-likelihood equations for distribution parameters $\underline{\beta}$ via Equation 43, and obtain the item parameter estimates by solving the system of EM-likelihood equations for all item parameters $\boldsymbol{\phi}$ via Equation 42.

4. The EM algorithm stops after convergence criteria are reached.

Note again that the maximization in the M-step can be performed for $\boldsymbol{\phi}$ and $\underline{\beta}$ separately, and the maximization for $\boldsymbol{\phi}$ can be performed item by item. If θ is discrete, then the BAHAF procedure

is similar to those described by Mislevy and Bock (1985; i.e., the BAM procedure) or Woodruff and Hanson (1996), where $\underline{\beta}$ is the vector of discrete probabilities. That is, the quadrature points and weights are changed to the discrete points and probabilities. One application of BAHAF is for the one-parameter logistic (Rasch) model. With a Rasch model in which the mean of θ is set to zero, BAHAF can be used to estimate the variance of θ as well as item difficulty parameter (Mislevy & Bock, 1985; Rigdon & Tsutakawa, 1983).

Therefore, theoretically the Bock-Aitkin procedure (see Equations 40 and 41) is a reformulation of the EM algorithm (see Equation 27) for continuous and discrete θ, and can include estimation of the θ-distribution parameters. Lewis (1985) had noticed that the Bock-Aitkin procedure is a special case of the EM algorithm.

Discussion

In this paper, it was shown that the Bock-Aitkin procedure is an instance of EM/MMLE when $\underline{\theta}$ is either continuous or discrete. The Bock-Aitkin procedures such as BA1, BA2, and BAM (see the introduction section) can be thought of as numerical implementation of the EM algorithm. Consequently, the procedure theoretically inherits the statistical properties of the EM algorithm. For example, the convergence of the Bock-Aitkin procedure is not guaranteed theoretically. Even if the procedure does converge, the value it converges to may not be the MML estimate. To date, among IRT models only the one-parameter logistic model with the normal θ distribution has been shown to converge to the MML estimate under the EM algorithm, assuming that the algorithm converges to a finite value (Tsutakawa, 1985).

As Bock and Aitkin (1981) pointed out, the M-step of the Bock-Aitkin procedure performs a probit (or logit) analysis on $\overline{n}_l$ and $\overline{r}_{li}$ (see Equations 20 and 21, or 44 and 45). Because of this connection to the probit/logit analysis, the computation load of EM/MMLE is reduced. The key points are (a) interchange of integration and summation on the third line of Equation 37 and (b) the formation of $n(\underline{\theta})^{(t)}$ and $r_i(\underline{\theta})^{(t)}$, so that the integral is not evaluated for each examinee for each item at each EM cycle.

The assumption of a continuous θ should be retained (unless θ is intended to be discrete). The θ scale can be discrete, because θ is unobserved and thus the scale is arbitrary. In addition, the test length is finite (a test can only be constructed with a finite number of items) and the number of distinct observed response vectors (or patterns) is finite (for item calibration the sample size of examinees usually is not large), so the number of ability levels that can be estimated with response patterns is limited. (Note that examinees with different θ values can produce the same response patterns, and different response patterns may be produced by the examinees with the same θ values.) However, when θ is estimated for the individual examinees by their response patterns, the estimated θ values may not match those values that were used in the earlier item calibration. Moreover, the Bock-Aitkin procedure was originally intended to solve the implementation difficulties of Bock and Lieberman's MML estimation for a continuous θ scale. Therefore, the assumption of a continuous θ variable can be adopted. If evaluation of the integral is desired, then the concern of how many ability values should be used is a question of how many (quadrature) points are needed for the precision of a numerical integral.

On the other hand, Lewis (1985) agreed with Mislevy and Bock (1985, also see Bock,

1997) that discrete distributions should be used to approximate continuous ability distributions, because "without such a simplification, the practical applications of the [EM/]GEM algorithm would be limited to relatively small samples of persons" (Lewis, p.208). Besides, in the Bock-Aitkin procedure the probit/logit analysis is linked to the M-step; that is, the complete-data likelihood is the likelihood of the probit analysis. This is what the interpretations of $\overline{n}$ and $\overline{r}$ (Equations 20 and 21) are based on. Since the number of levels of dosage (stimuli) is finite in the probit analysis, so is the number of θ values in the Bock-Aitkin (implementation) procedure. Although BA1 is described first where Gauss-Hermite quadrature is used, BAM where a discrete representation is used should be the typical Bock-Aitkin procedure. In other words, the use of numerical quadrature is a preprocess to obtain a discrete representation. Tsutakawa (1984) mentioned that the Bock-Aitkin procedure differs from EM/MMLE by using a discrete representation with predetermined ability levels. It should be noted that using numerical quadrature to evaluate the integrals (when there is no closed form to compute the integrals), and using a discrete distribution/representation with a finite number of points to approximate the continuous distribution are different approaches. Moreover, even the θ values are not necessarily prespecified, although the number of θ levels is prespecified. Bock (1997) mentioned that "if both the probabilities and the locations of the points are estimated jointly, in the manner of Kiefer and Wolfowitz, the latent distribution is characterized semiparametrically" (p.29). The practical meaning is that to use BAM we do not need the assumption about the underlying θ distribution which can be estimated by BAM, except for the number of θ levels. Keep in mind that item parameter estimates are with respect to the sample of examinees. The number of θ levels affects MML estimation precision,

but how well this "nonparametric" estimation of the θ distribution influences MML estimation may need further study.

Because the Bock-Aitkin procedure is an instance of the EM algorithm, the θ-distribution parameters can be estimated along with item parameters, as described in the last section. However, there are two constraints needed to identify the two-parameter logistic model family, which includes the three-parameter logistic model (Birnbaum, 1968) with the lower asymptote parameter fixed (see Hambleton & Swaminathan, 1985; Tsutakawa, 1984, 1985). One way is to set the mean of θ to be zero and the variance to be one. If θ is assumed as discrete with a set of q *predetermined* θ values, then in addition to item parameters, there are q distribution parameters (i.e., the probabilities associated with each predetermined θ level) to be estimated subject to two constraints. The q probabilities can be re-estimated at each EM cycle as described in the above extended procedure, but the mean and variance should be adjusted to meet the two constraints at each cycle or after the final cycle. The new transformed θ values may not be the same as the predetermined values. The item parameter estimates also need to be adjusted correspondingly. The necessity of this re-standardization for identifying item parameters when θ is discrete has been questioned by Lewis (1985, p.205). But if both the probabilities and the locations of the points of a discrete representation are estimated jointly, Bock (1997, p.29) regarded the (underlying) θ distribution as characterized semiparametrically. On the other hand, if θ, throughout the study, is assumed to be fixed as $N(0, 1)$, then it is not necessary to use a discrete representation, but Gauss-Hermite quadrature can be used, and the quadrature points and associated weights do not need to be changed during the EM iteration, because the θ distribution is known and fixed.

The EM algorithm perspective is more flexible than the Bock-Aitkin procedure perspective. If the quadrature-point technique is adopted, then other kinds of numerical quadrature techniques than Gauss-Hermite quadrature can be considered. Moreover, in addition to the Newton-Raphson method, other maximization methods can be considered in the M-step.

For some IRT models, the maximization (for example, using the Newton-Raphson iteration) in the M-step may not be attained. When this occurs, a generalized EM algorithm (GEM, see Equation 25) can be considered.

Whereas the EM algorithm produces MLEs, Dempster et al. (1977) also described a modification of the EM algorithm for Bayesian modal estimates (for applications of the Bayesian-approach EM algorithm to IRT see Harwell & Baker, 1991; Mislevy, 1986; Tsutakawa & Lin, 1986). These two algorithms have fundamentally different philosophies of statistical inference (Lewis, 1985). In the Bayesian approach, item parameters are also treated as random variables with prior distributions, that is, as a random sample from a large population (item pool) (Rigdon & Tsutakawa, 1987). In IRT, the Bayesian approach is used, as a numerical tool, to prevent parameter estimates from becoming indefinitely large (e.g., Mislevy & Bock, 1993) or to control the parameter estimates within a desired range. The choice of prior distributions for item parameters is often made in an arbitrary manner. The procedure is for Mislevy's marginalized Bayesian estimation (see Harwell & Baker, 1991), not MML estimation.

References

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

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

Bock, R. D. (1995). Full-information item factor analysis by fractional quadrature. In R. D. Bock (Organizer), *Marginalization methods for two-stage measurement models [Abstract].* Symposium conducted at the annual meeting of the Psychometric Society, Minneapolis.

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: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., & Schilling, S. G. (1997). High-dimensional maximum marginal likelihood item factor analysis. Manuscript in preparation.

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

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

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48,* 129-141.

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

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

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15,* 375-389.

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

Holland, P. W. (1981). When are item response theory models consistent with observed data? *Psychometrika, 46,* 79-92.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577-601.

Lewis, C. (1985). Discussion. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 203-209). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McKinley, R. L., & Reckase, M. D. (1983, August). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR 83-2). Iowa City, IA: American College Testing Program.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 189-202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R. J., & Bock, R. D. (1993). BILOG 3: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

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

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73-90.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567-574.

Rigdon, S. E., & Tsutakawa, R. K. (1987). Estimation for the Rasch model when both ability and difficulty parameters are random. *Journal of Educational Statistics, 12,* 76-86.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175-186.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics, 9,* 263-276.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51,* 251-267.

Tsutakawa, R. K. (1985). Estimation of item parameters and the GEM algorithm. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 180-188). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Woodruff, D. J., & Hanson, B. A. (1996, September). *Estimation of item response models using the EM algorithm for finite mixtures* (ACT Research Report 96-6). Iowa City, IA: ACT.

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