Online Appendix A

In this appendix, we review the properties of three popular IRT trait estimation methods. Each of these trait estimation algorithms considers the likelihood of a response vector $\boldsymbol{y}_j$ by person j, which for dichotomous items equals

$$L(\theta|\boldsymbol{y}_j, \boldsymbol{\xi}) = \prod_{i=1}^{M} P(y_{ij} = 1|\boldsymbol{\xi}_i, \theta)^{y_{ij}} (1 - P(y_{ij} = 1|\boldsymbol{\xi}_i, \theta))^{1-y_{ij}}, \tag{A1}$$

where P denotes the probability of a keyed item response ($y_{ij} = 1$) for item i, $i = 1, \ldots, M$, θ is the latent trait parameter, and $\boldsymbol{\xi}$ is a vector of item parameters. The log of the likelihood function equals

$$l(\theta\ |\boldsymbol{y}_j, \boldsymbol{\xi}) = \sum_{i=1}^{M} \left[y_{ij} \log(P(y_{ij} = 1|\boldsymbol{\xi}_i, \theta)) + (1 - y_{ij}) \log(1 - P(y_{ij} = 1|\boldsymbol{\xi}_i, \theta)) \right], \tag{A2}$$

and the test information function, $I(\theta)$, equals

$$I(\theta) = \sum_{i=1}^{M} \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{A3}$$

where $P_i(\theta)$ is condensed notation for $P(y_{ij} = 1|\boldsymbol{\xi}_i, \theta)$, and $P_i'(\theta) = \frac{\partial P_i(\theta)}{\partial \theta}$. The maximum likelihood (ML) trait estimate equals the θ value that maximizes the log likelihood in Equation A2, and the standard error of the ML trait estimate equals the inverse square root of the test information function evaluated at the ML trait estimate. Assuming correct model specification and conditional independence, the ML trait estimate is asymptotically unbiased and normally distributed with a standard deviation equal to the standard error of the estimate (Birnbaum, 1968, p. 457). However, this asymptotic result holds as the number of test items increases, and most psychometric tests are too short to produce unbiased ML trait estimates. Lord (1983) demonstrated that ML trait estimates are biased outwards, that is, high trait scores are positively biased and low trait scores are negatively biased. Consequently, ML trait estimates can be infinite (for non-mixed response patterns) or otherwise implausibly extreme. Thus, it is usually necessary to constrain the ML trait estimate, for example, between -4 and +4. Additionally, ML trait estimates may be biased even when there is high conditional test information (Samejima, 1993), that is, if the standard error of the estimate is small. A Bayesian alternative to ML trait estimation is the *expected a posteriori* (EAP; Bock & Mislevy, 1982) ability estimate, which equals the mean of the posterior distribution of ability after observing a subject's responses to one or more items. In quadrature form, the EAP trait estimate equals

$$\hat{\theta}^{\text{EAP}} = \frac{\sum_{q=1}^{Q} X_q L(X_q) W(X_q)}{\sum_{q=1}^{Q} L(X_q) W(X_q)}, \tag{A4}$$

where $X_1, X_2, \ldots, X_Q$ are the set of Q quadrature nodes, L is the likelihood in Equation A1, and $W(X_q)$ is the quadrature weight associated with node q. Quadrature weights are calculated from the prior distribution of the latent trait and scaled such that $\sum_{q=1}^{Q} W(X_q) = 1$. The standard deviation of the posterior distribution equals

$$se(\hat{\theta}^{\text{EAP}}) = \left(\frac{\sum_{q=1}^{Q} (X_q - \hat{\theta})^2 L(X_q) W(X_q)}{\sum_{q=1}^{Q} L(X_q) W(X_q)} \right)^{1/2}. \tag{A5}$$

Although the $se(\hat{\theta}^{\text{EAP}})$ notation is technically incorrect, we use it here to emphasize that the posterior standard deviation is often substituted for the standard error of the estimated trait scores when computing reliability coefficients (Bock & Mislevy, 1982) or when constructing confidence intervals (Waller & Reise, 1989). Although Bayesian estimates are known to be biased toward the middle of the prior distribution, EAPs may be preferred because they always produce finite trait estimates in short tests or during the early stages of a computerized adaptive test (CAT; Weiss, 1982). Additionally, Bock and Mislevy (1982) showed that after approximately 20 items are administered, the posterior distribution approaches normality and closely traces the likelihood function. As a result, EAP and ML trait estimates are very similar in sufficiently long tests. An alternative and unbiased trait estimation method is weighted likelihood estimation (WL; Warm, 1989). Based on Lord's (1983) derivation for the finite-test bias of the ML trait estimator, the WL trait estimate is unbiased in tests of finite length and equals the θ value that satisfies

$$\frac{J(\theta)}{2I(\theta)} + \frac{\partial l(u_j|\boldsymbol{\xi}, \theta)}{\partial \theta} = 0, \tag{A6}$$

where

$$J(\theta) = \sum_{i=1}^{M} \frac{P_i' P_i''}{P_i(1 - P_i)}, \tag{A7}$$

$\frac{\partial l(u_j|\boldsymbol{\xi}, \theta)}{\partial \theta}$ equals the first derivative of the log likelihood in Equation A2 with respect to θ, $P_i' = \frac{\partial P_i}{\partial \theta}$, and $P_i'' = \frac{\partial^2 P_i}{\partial \theta^2}$. The standard error of the WL $\hat{\theta}$ equals

$$se(\hat{\theta}^{\text{WL}}) = \left(\frac{I'(\hat{\theta}) J(\hat{\theta}) - I(\hat{\theta}) J'(\hat{\theta})}{2I(\hat{\theta})^2} + I(\hat{\theta}) \right)^{-1/2} \tag{A8}$$

where $I(\theta)$ is the test information function in Equation A3, $I'(\theta) = \frac{\partial I(\theta)}{\partial \theta}$, and $J'(\theta) = \frac{\partial J(\theta)}{\partial \theta}$. In a Monte Carlo simulation study, Warm (1989) found that the WL estimator outperformed both ML and a Bayesian estimator in terms of average absolute error, bias, and mean squared error. In that study, the advantages of WLE over the other estimators were most prominent for extreme θ values.

References

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

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, of of their parallel-forms reliability. *Psychometrika*, *48*, 233–245.

Samejima, F. (1993). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika*, *58*, 195–209.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology*, *57*, 1051–1058.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

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