Calibrating CAT Pools and Online Pretest Items Using MCMC Methods

Daniel O. Segall
Defense Manpower Data Center
Monterey Bay, CA

Abstract

This paper investigates the application of a Markov chain Monte Carlo approach (Segall, 2002) to the estimation of item response functions using sparse data matrices. These sparse data matrices are of the sort produced by computerized adaptive test sessions. The estimation approach uses a Bayesian hierarchical model, where slope and intercept parameters are assumed to be sampled from informative distributions who's moments are estimated from data, and where the ability mean and variance parameters are also estimated simultaneously with other parameters. Results of a simulation study indicate that the procedure can provide satisfactory recovery of item response functions from sparse adaptive testing data.

Each year, nearly one-half million military applicants take the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery ([CAT-ASVAB]; Segall & Moreno, 1999). This battery of tests is used to qualify applicants for military service, and for entrance into a number of jobs within the military. Data collected from applicants are also used to calibrate new items for possible use in future item pools. This paper describes work that is part of a larger effort (Krass & Willians, 2003; Nicewander, 2003; Pommerich & Segall, 2003; Thomasson, 2003) to evaluate alternate estimation procedures for calibrating new items for use in the CAT-ASVAB.

Seeding Strategy

The CAT-ASVAB uses a unique design for collecting pretest data for the evaluation of new items (Segall, Moreno, Bloxom, & Hetter, 1997). The battery is comprised of 10 separately timed, administered, and scored adaptive tests. Each of these 10 tests measure different abilities and constructs. However, the items within each test are assumed to be unidimensional. This assumption is strongly supported for most tests by empirical evidence (Segall, Moreno, & Hetter, 1997).

This paper was presented at the annual meeting of the National Council on Measurement in Education (April, 2003), Chicago, IL. The views expressed are those of the author and not necessarily those of the Department of Defense, or the United States government. Requests for copies should be sent to: Daniel O. Segall, Defense Manpower Data Center, DoD Center Monterey Bay, 400 Gigling Road, Seaside, CA 93955-6771. Email: publications@danielsegall.com

According to the data collection design, each examinee receives a single pretest item along with 10 (or 15) adaptively selected items for each of the 10 tests. The responses to the pretest items do not count towards the examinee's scores, and are collected solely for the purpose of gathering data on new items. The pretest items are positioned as the second, third, or fourth item of each adaptive test, where the position is chosen randomly for each test-taker and test.

The fact that only one pretest item response is gathered for each test means that the responses to the adaptively administered items will play an important and necessary role in the estimation of item response functions (IRFs). The approach investigated here assumes that 10 separate unidimensional calibrations will be conducted for pretest items seeded into each of the 10 adaptive tests.

Calibration Issues

Stocking (1988a; 1988b) has noted a troublesome scale-drift phenomenon when using the three-parameter logistic item response theory model in the context of online calibration. She noted that scale-drift can occur over successive generations of pretest calibrations. According to this scenario, the parameter estimates for the current set of CAT pools are assumed to come from pretest calibrations of items seeded into previous sets of CAT pools. She noted a systematic distortion of the score scale after several successive rounds of calibrations and item pool replacements.

One factor that might contribute to this distortion is the positive bias in the IRT a (discrimination) parameter. Items are selected into the pool partly on the basis of their a-parameter. In addition, CAT-ASVAB selects items largely on the basis of item-information, which is also heavily dependent on the a-parameters. Consequently, items with chance positive errors in their discrimination parameters are likely to be administered more often than items with negative estimation errors. If the calibration procedure fixes the IRFs of adaptively administered items at their previously (biased) estimated values, these positive errors among the most heavily used items are likely to further bias the values of estimated IRFs for new items. Over successive generations of item pools, this phenomenon can lead to a systematic distortion of the score-scale.

One way to avoid this distortion caused by the systematic bias of item parameters among the most heavily used items is to rely on re-estimated IRFs using new data, rather than on existing biased estimates obtained from the previous data collection round. If the IRFs for adaptively administered items are re-estimated with new data, then these IRFs should be less biased with regard to discrimination parameters (and other parameters as well). By re-estimating all (pretest and operational) parameters simultaneously, the scale-distortion problem should diminish, if not disappear entirely. However, this approach places more substantial demands on the calibration procedure, because of the sparseness of the data matrix—some items will be administered to relatively few respondents, while other items will be administered to many test-takers. Still some difficult items will be administered to only high-ability test-takers, while other easy items will be administered to primarily low-ability test-takers. Consequently, in this paper the capability of procedures to estimate the IRFs of pretest items will be investigated in the context of re-estimating the entire adaptive item pool.

If all (pretest and adaptive) IRFs are estimated simultaneously, the unit and origin of the IRT scale become arbitrary and must be fixed using some convention. To fix the score scale, it is assumed that the parameters of at least some items remain constant over time, and are known. This assumption is plausible, if these items were calibrated using large samples, and if they are administered infrequently relative to the other items. In the paradigm described below, items from Form 04D are assumed to have known IRFs, and data collected from 04D are used to fixed the unit and origin of the IRT score-scale.

Estimation Procedure

The estimation approach evaluated here is based on a Markov Chain Monte Carlo approach (Segall, 2002a) implemented by the IFACT computer program (Segall, 2002b). Markov chain Monte Carlo (MCMC) techniques have proven useful for complex estimation problems in many areas of applied statistics. Several authors have explored the applicability of MCMC estimation approaches to IRT (e.g., Albert, 1992; Baker, 1998; Béguin & Glas, 2001; Fox & Glas, 2001; Hoijtink & Molenaar, 1997; Patz & Junker, 1999a; Patz & Junker, 1999b).

The MCMC approach taken in this paper extends that taken by others in two respects. First, informative sampling distribution of item parameters are estimated for distributions of item parameters, where slope and intercept parameters are assumed to be sampled from informative distributions who's moments are estimated from data. This hierarchical approach is likely to be especially well suited to the analysis of data produced by computerized adaptive tests, where the numbers of examinees administered different items varies widely. In these instances, strength can be pooled across items, helping to increase the precision of IRF-estimates of items adaptively administered to small numbers of test-takers.

A second extension of the current application treats the mean and variance of the latent ability distribution as parameters to be estimated rather than as fixed known parameters. This allows the unit and origin of the IRT scale to be fixed by fixing some item parameters, and estimating the ability distribution mean, variance, and other item parameters simultaneously.

The objective of the MCMC approach implemented in IFACT is to obtain estimates of multidimensional item response theory item parameters. Typically, MCMC estimates are obtained by forming means of individual item-parameter values sampled from their joint posterior distribution. Similar to the approach taken by Béguin and Glas, we augment the existing item response data and item parameters with other parameters. This augmented model results in full-conditional distributions which simplify the simulation process and allows the use of Gibbs sampling. Specifically, we seek to draw samples from the joint posterior distribution:

$$\begin{aligned} p\left(\boldsymbol{\xi}=(\boldsymbol{\tau},\boldsymbol{\lambda}),\mathbf{c},\boldsymbol{\theta},\mathbf{z},\mathbf{v},\boldsymbol{\mu}_{\xi},\Delta_{\xi},\boldsymbol{\mu},\mathbf{W}|\mathbf{U}\right) &= p\left(\mathbf{z},\mathbf{v}|\mathbf{U},\boldsymbol{\theta},\boldsymbol{\xi},\mathbf{c}\right)p\left(\boldsymbol{\xi}|\boldsymbol{\mu}_{\xi},\Delta_{\xi}\right)p\left(\mathbf{c}\right) \\ &\quad \times p\left(\boldsymbol{\mu}_{\xi}\right)p\left(\Delta_{\xi}\right)p\left(\boldsymbol{\theta}|\boldsymbol{\mu},\mathbf{W}\right)p\left(\boldsymbol{\mu}\right)p\left(\mathbf{W}\right) \end{aligned}$$

where $\boldsymbol{\xi}$ is an $n\times(m+1)$ matrix of slope $\boldsymbol{\lambda}$ and intercept $\boldsymbol{\tau}$ parameters; $\mathbf{c}$ is the n-element vector of guessing parameters; $\boldsymbol{\theta}$ is the $N\times m$ matrix of ability parameters; $\mathbf{z}$ and $\mathbf{v}$ are $N\times n$ matrices of latent continuous and dichotomous item-knowledge parameters; $\boldsymbol{\mu}_{\xi}$ and Δ_{ξ} are the mean vector and covariance matrix of the distribution of slope and intercept item

parameters, and $\boldsymbol{\mu}$ and $\mathbf{W}$ are the mean vector and covariance matrix of ability parameters $\boldsymbol{\theta}$. Marginal estimates of parameters of interest can be obtained by ignoring the sampled values of the augmented parameters (i.e., $\boldsymbol{\theta}, \mathbf{z}, \mathbf{v}, \boldsymbol{\mu}_\xi, \Delta_\xi, \boldsymbol{\mu}, \mathbf{W}$), and by computing parameters means of $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, and $\mathbf{c}$ once the chain has converged (Gilks, Richardson, & Spiegelhalter, 1996). After providing starting values of item and person parameters $\boldsymbol{\tau}$, Λ, $\mathbf{c}$, and $\boldsymbol{\theta}$, the general updating process proceeds as follows:

1. Draw $\mathbf{z}$ and $\mathbf{v}$ conditional on $\mathbf{U}$, $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\mathbf{c}$, and $\boldsymbol{\theta}$.
2. Draw $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ conditional on $\boldsymbol{\theta}$, $\mathbf{z}$, $\boldsymbol{\mu}_\xi$, and Δ_ξ.
3. Draw $\boldsymbol{\mu}_\xi$ and Δ_ξ conditional on $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$.
4. Draw $\mathbf{c}$ conditional on $\mathbf{v}$ and $\mathbf{U}$.
5. Draw $\boldsymbol{\mu}$ and $\mathbf{W}$ conditional on $\boldsymbol{\theta}$.
6. Draw $\boldsymbol{\theta}$ conditional on $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\mathbf{z}$, $\boldsymbol{\mu}$, and $\mathbf{W}$.

Steps 1–6 were repeated for 1000 iterations with response data described below, until convergence had been reached. Then the average $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\mathbf{c}, \boldsymbol{\mu}$, and $\mathbf{W}$ parameter values (averaged across 1000 additional cycles) were taken as parameter estimates. Details of each of the six steps (along with additional model assumptions) are described by Segall (2002a).

Simulation Study

Three datasets of 122,400 examinees each were used to evaluate the performance of the IFACT estimation procedure[1]. The structure of the datasets, characteristics of the item pool and pretest items, and results of the IFACT estimation procedure are described below.

Item Pool Characteristics

The item pools used in the simulation study were pattered after pools used in the Arithmetic Reasoning test of the CAT-ASVAB. Estimated 3PL item parameters from previous calibrations were treated as true values for the purpose of data generation. The pretest group consisted of 100 items. See Pommerich and Segall (2003) and Thomasson (2003) for additional details regarding the characteristics of the adaptive and pretest item pools.

Dataset Structure

Three datasets of item responses were generated, each dataset based on a different normal distribution of ability: (a) low ability group with $\mu_\theta = -1, \sigma = 1.2$, (b) medium ability group with $\mu_\theta = 0, \sigma = 1$, and (c) a high ability group with $\mu_\theta = 1, \sigma = 0.8$. The simulated item selection and scoring was patterned after that used by the CAT-ASVAB (Segall et al., 1997), with adaptive test lengths of 15 items; maximum information item selection contingent on the Sympson-Hetter exposure control algorithm (Hetter & Sympson, 1997; Sympson & Hetter, 1985); and Bayesian modal ability estimates computed after the final administered item.

For each dataset, the simulation data generation design (Thomasson, 2003) closely matched the characteristics of pretest item data collected in the operational CAT-ASVAB. According to this design, data for the calibration of 100 pretest items are collected from 122,400 examinees under the following restrictions:

[1]See Krass and Willians (2003)and Pommerich and Segall (2003)for a description of the results from two other estimation approaches evaluated on the same data.

Table 1: True and Estimated Ability Distribution Moments

	Population		Estimated	
Ability Group	μ	σ	μ	σ
Low	-1	1.20	$-.88$	1.12
Medium	0	1.00	.03	.97
High	1	.80	.98	.82

• Randomly equivalent (in terms of latent ability) groups of test-takers are assigned to one of four alternate items pools (denoted by 01D, 02D, 03D, and 04D). Each item pool consists of unique sets of items (containing 94 items for Form 01D, and 137 items each for the remaining 3 forms). These forms are designed to produce interchangeable test-scores.

• About one-third of the test-taker group ($N = 40,000$) are assigned to each of the three Pools 01D, 02D, and 03D. The remainder of the sample ($N = 2,400$) is assigned to Pool 04D.

• For each examinee, a pretest item is randomly selected from the pool of 100 pretest items. This process results in 1,224 responses to each pretest item, with randomly equivalent (with regard to θ-distribution) groups of respondents answering each pretest item.

Results

Table 1 summarizes the estimated ability distribution moments from the IFACT estimation procedure. As indicated, the estimated mean μ and standard deviation σ are very close to the true values used to generate the data in each of the three datasets.

Table 2 summarizes root mean squared difference (RMSD) statistics between IRFs calculated from true (a, b, c) and estimated $\left(\hat{a}, \hat{b}, \hat{c}\right)$ parameters for each of the three conditions. The squared differences between IRFs were evaluated at the true θ for each examinee receiving the item. Separate summaries were computed for adaptive and pretest items:

$$\text{RMSD(Adaptive)} = \left\{ (15N)^{-1} \sum_{j=1}^{N} \sum_{i=1}^{15} \left[P\left(a_{A_{ij}}, b_{A_{ij}}, c_{A_{ij}}; \theta_j\right) - P\left(\hat{a}_{A_{ij}}, \hat{b}_{A_{ij}}, \hat{c}_{A_{ij}}; \theta_j\right) \right]^2 \right\}^{1/2}$$

$$\text{RMSD(Pretest)} = \left\{ N^{-1} \sum_{j=1}^{N} \left[P\left(a_{S_j}, b_{S_j}, c_{S_j}; \theta_j\right) - P\left(\hat{a}_{S_j}, \hat{b}_{S_j}, \hat{c}_{S_j}; \theta_j\right) \right]^2 \right\}^{1/2}$$

where $P(a, b, c; \theta)$ denotes the three-parameter logistic function; A_{ij} denotes the item index of the ith adaptively administered item to the jth examinee; S_j denotes the item index of the single seed (pretest) item administered to the jth examinee; θ_j denotes the true ability parameter for examinee j; and $N = 120,000$. As expected, the RMSDs for adaptive items were slightly larger than those for pretest items. This is most likely due to the reduced sample sizes for some adaptively administered items. In general, the RMSDs were acceptably small.

Table 3 summarizes the moments of the true and estimated parameters. As indicated by a comparison of the true and estimated parameter means, there appeared to be a slight

Table 2: RMSDs Between True and Estimated IRFs

Ability Group	Adaptive Items	Pretest Items
Low	.047	.030
Medium	.024	.023
High	.021	.020

Table 3: Characteristics of True and Estimated Parameters

	Adaptive			Pretest		
Statistic	a	b	c	a	b	c
True Parameters						
μ	1.53	.02	.18	1.51	.08	.19
σ	.46	1.17	.06	.46	1.21	.06
Estimated – Low Ability Group						
μ	1.77	.18	.23	1.46	.16	.19
σ	.39	1.13	.11	.31	1.12	.06
Estimated – Medium Ability Group						
μ	1.70	.24	.24	1.48	.15	.21
σ	.37	1.07	.14	.42	1.13	.09
Estimated – High Ability Group						
μ	1.68	.47	.30	1.45	.26	.30
σ	.32	.80	.20	.42	1.00	.20

positive bias for the a and b parameters for the Adaptive items. This bias was less evident however for the Pretest items.

Table 4 displays item parameter recovery statistics for each condition, including the correlation r and root-mean-squared-difference RMSD between true and estimated parameters. As expected, parameter recovery for the pretest items was superior to that of the adaptive items. Also as expected, b parameters were better estimated than other parameters; a parameters were more precisely estimated than the guessing parameters c; and the c parameter was more precisely estimated in the low-ability group than in the medium or high ability groups.

Discussion

The IFACT procedure appeared to provide satisfactory results for the estimation of item parameters from sparse data matrices. The estimates of pretest items appear to be estimated sufficiently well for use in adaptive item pools, where these parameters will be used to select items and generate test scores.

One advantage of the IFACT procedure over BILOG (Zimowski, Muraki, Mislevy, & Bock, 2003) is that the estimates can be placed on the proper scale by fixing some IRFs at previously estimated values, rather than through a separate linking step. This might reduce scale drift over time. However, additional study would be required to confirm this

Table 4: Item Parameter Recovery

	Adaptive			Pretest		
Statistic	a	b	c	a	b	c
Low Ability Group						
r	.81	.98	.46	.84	1.00	.80
RMSD	.36	.31	.11	.26	.16	.04
Medium Ability Group						
r	.80	.91	.29	.87	.99	.36
RMSD	.33	.54	.15	.23	.19	.09
High Ability Group						
r	.71	.72	.06	.86	.92	$-.05$
RMSD	.35	.85	.24	.24	.52	.24

assertion.

The ForScore (Levine, in press) procedure is possibly disadvantaged by the requirement of initial estimates of the adaptive item parameters. These initial estimates are required for both re-estimating adaptive items, and for estimating IRFs of pretest items. Results suggest however, that misspecification of these adaptive IRFs (by an amount expected from a conventional calibration of these items) will have only a small negative effect on the precision of the estimates of adaptive and pretest items.

One potential disadvantage of both the IFACT and BILOG procedures is their heavy reliance on the unidimensional 3PL assumptions. The ForScore procedure evaluated by Krass and Willians (2003) might be less susceptible to model violations than either the BILOG or IFACT approaches. However, additional simulation studies are required to confirm this hypothesis.

The MCMC approach investigated here will be evaluated using additional data to examine the outcome of parameter estimates and scale drift over successive rounds of pretest item calibration and adaptive item pool construction phases. In each round, the pretest items calibrated in the previous round will be used to construct adaptive item pools for the next round. In these future simulations, item response data will be generated which violates (to a realistic degree) assumptions of unidimensionality and monotonicity of item response functioning. Results from the evaluation of these data will provide additional useful information for evaluating the relative merits of the IFACT, BILOG, and ForScore approaches for use in CAT-ASVAB item pool maintenance, and will help answer important questions regarding scale drift.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269.

Baker, F. B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*, 153–169.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–562.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chan Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1–19). New York: Chapman & Hall.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*, 171–189.

Krass, I. A., & Willians, B. (2003, April). *Using nonparametric and adjusted marginal maximum likelihood methods for online calibration.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Levine, M. V. (in press). Dimension in latent variable models. *Journal of Mathematical Psychology.*

Nicewander, W. A. (2003, April). *Issues in maintaining scale consistency for the CAT-ASVAB.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

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

Pommerich, M., & Segall, D. O. (2003, April). *Calibrating CAT pools and online pretest items using marginal maximum likelihood methods.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Segall, D. O. (2002a, April). *Confirmatory item factor analysis using Markov chain Monte Carlo estimation with applications to online calibration in CAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Segall, D. O. (2002b). IFACT computer program Version 2.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [Computer program]. Seaside, CA: Defense Manpower Data Center.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 131–140). Washington, DC: American Psychological Association.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117–130). Washington, DC: American Psychological Association.

Stocking, M. L. (1988a). *Scale drift in on-line calibration.* (Tech. Rep. No. ERIC ED389710). Educational Testing Service, Princeton, N.J.

Stocking, M. L. (1988b). *Some considerations in maintaining adaptive test item pools.* (Tech. Rep. No. ERIC ED391814). Educational Testing Service, Princeton, N.J.

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

Thomasson, G. L. (2003, April). *Evaluating stability of online item calibrations under varying conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. D. (2003). BILOG-MG [Computer program]. Lincolnwood, IL: Scientific Software International, Inc.