

■ **The Impact of Item Parameter Estimation
on Computerized Adaptive Testing with
Item Cloning**

Cees A. W. Glas
University of Twente, Enschede, The Netherlands

■ **Law School Admission Council**
Computerized Testing Report 02-06
November 2005

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

© 2005 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any part or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 662 Penn Street, Box 40, Newtown, PA 18940-0040.

LSAT and LSAC are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary	1
Abstract	1
Introduction	1
Model	2
<i>First-level Model</i>	2
<i>Second-level Model</i>	2
<i>Prior for Hyperparameters</i>	2
<i>Likelihood Function</i>	3
Bayesian Estimation.	3
Adaptive Testing	3
<i>Model 1</i>	4
<i>Model 2</i>	4
<i>Model 3</i>	4
Simulation Studies	5
<i>Item Pools</i>	5
<i>The Calibration Phase</i>	5
<i>The CAT Phase</i>	5
<i>Results</i>	5
Conclusion	6
References	6

Executive Summary

Item cloning techniques involve the application of an algorithm or algorithms to generate new test items based on the characteristics of existing test items. The term *parent* is sometimes used to describe the original items, and the term *offspring* is sometimes used to describe the new items that are generated by the cloning technique. The application of item cloning techniques can greatly reduce the cost of item writing and enhance the flexibility of item presentation. Another cost-saving factor of item cloning is that the item response theory (IRT) item statistics (commonly referred to as item parameters) that are calculated for each test item and used to select items for administration in a computerized adaptive test (CAT) are often assumed to be the same, or at least very similar, for parent and offspring items.

However, a potential negative consequence of cloning is the possible variation in the values of the item parameters for items cloned from the same parent. Recent research has shown that the presence of item parameter estimation error can lead to capitalization on these errors and hence to substantial loss of precision in the ability estimates derived for test takers.

This research presents a solution that accounts for the uncertainty in the item parameter estimates, thereby resolving the problem of capitalization on item parameter estimation errors. A simulation study is presented to illustrate how the method neutralizes the effect of capitalization on errors in item parameter estimates.

Abstract

Item cloning techniques can greatly reduce the cost of item writing and enhance the flexibility of item presentation. An important consequence of cloning is that it may cause variability in the item parameters. Recently, Glas and van der Linden (in press, 2005) proposed a multilevel item response model where it is assumed that the item parameters of a 3-parameter logistic (3PL) model or a 3-parameter normal ogive (3PNO) model are sampled from a multivariate normal distribution associated with a parent item. In the sequel, the model will be referred to as the item cloning model, which will be abbreviated ICM. Several procedures for item bank calibration and computerized adaptive testing (CAT) were proposed. The latter procedures were developed under the usual assumption that the item parameters are known. However, in practice, item parameters have to be estimated, which introduces an error component that can have substantial effects. For the standard 3PL model, van der Linden and Glas (2000, 2001) show that capitalization on estimation error can lead to a substantial loss of precision. In the present report, this finding is corroborated for the ICM. It is shown that the problem can be solved by a Bayesian item selection procedure where the uncertainty about the item parameters is taken into account by implicating their posterior distributions. These posterior distributions are generated using the Gibbs Sampler. A simulation study is presented to illustrate the performance of the method.

Introduction

Item cloning is based on a formal description of a set of parent items and an algorithm to derive a larger set of operational items from them. These parent items have been known as *item forms*, *item templates*, or *item shells*, whereas the items generated from them are now widely known as item clones. Comprehensive reviews of the technology of item cloning are given in Bejar (1993) and Roid and Haladyna (1982).

Recently, Glas and van der Linden (in press, 2005) proposed a multilevel item response (IRT) model where it is assumed that the item parameters of a 3PL model are sampled from a multivariate normal distribution associated with a parent item. The model is fully Bayesian in the sense that (informative) priors are formulated for all hyperparameters describing the distributions of the item parameters within the populations. The numerical procedure used to calculate the estimates is the Markov Chain Monte Carlo (MCMC) simulation (Gibbs Sampler).

Glas and van der Linden (2005a) proposed several procedures for CAT with item clones. These procedures were developed under the usual assumption that the item parameters are known; however, in practice, item parameters have to be estimated, which introduces an error component that can have substantial effects. For the standard 3PL model, van der Linden and Glas (2000, 2001) show that capitalization on estimation error can lead to a substantial loss of precision. The main cause of this phenomenon is that highly discriminating items (items with a high discrimination parameter in the 3PL or 3PNO model) are both likely to be selected and are bound to have large standard errors. Paradoxically, the size of the item bank is negatively related to the precision of the ability estimates. The reason is that a large item bank contains more highly discriminating items with large standard errors. In the present report, this phenomenon is also found for the ICM. However, it is shown that the problem can be solved by a Bayesian item selection procedure where the uncertainty about the item parameters is taken into account using their posterior distributions generated with the Gibbs Sampler.

Model

Consider a set of item populations $p = 1, \dots, P$ of size K_1, \dots, K_p , respectively. The item clones in population p will be labeled $i_p = 1, \dots, K_p$. It proves convenient to introduce sampling design variables d_{ni_p} , which assumes a value equal to one if person n responded to item i_p and zero otherwise. Let X_{ni_p} be the response variable for person n and item clone i_p . If $d_{ni_p} = 1$, X_{ni_p} attains the value of one for a correct response and a value of zero for an incorrect response. If $d_{ni_p} = 0$, X_{ni_p} attains an arbitrary value of r ($r \neq 0; r \neq 1$). Notice that with this definition the design variables are completely determined by the response variables; they are only introduced to facilitate the mathematical presentation.

First-level Model

The first-level model is the 3PNO model, which describes the probability of a correct response as

$$p(x_{ni_p} = 1 | d_{ni_p} = 1, \theta_n, a_{i_p}, b_{i_p}, c_{i_p}) = c_{i_p} + (1 - c_{i_p}) \Phi(a_{i_p} \theta_n - b_{i_p}), \quad (1)$$

where a_{i_p} , b_{i_p} , and c_{i_p} are item parameters, θ_n is an examinee parameter, and $\Phi(\cdot)$ is the normal cumulative distribution function. The parameterization of the models in (1) is slightly different from the usual parameterization for the logistic and normal-ogive models, $a_{i_p} = (\theta - b_{i_p})$. The only motivation for our choice is to simplify the presentation below.

The reason for considering the 3PNO model rather than the 3PL model is that the former appears to be more tractable in an MCMC framework. However, as is well known, for an appropriately chosen scale factor both models are numerically nearly indistinguishable and either model is expected to fit only if the other does.

Second-level Model

The values of the item parameters $(a_{i_p}, b_{i_p}, c_{i_p})$ in (1) are considered as realizations of a random vector. Further, it is assumed that

$$\xi_{i_p} = (a_{i_p}, b_{i_p}, \text{logit } c_{i_p}), \quad (2)$$

has a multivariate normal distribution, that is,

$$\xi_{i_p} \sim N(\mu_p, \Sigma_p), \quad (3)$$

where μ_p is the vector with the mean values of the item parameters for population p and Σ_p , their covariance matrix. The hyperparameters (μ_p, Σ_p) are allowed to vary across the populations of items. Finally, it will be assumed that θ_n has a standard normal distribution

$$\theta_n \sim N(0, 1). \quad (4)$$

Prior for Hyperparameters

A convenient choice for the prior distribution for the hyperparameters (μ_p, Σ_p) is a normal-inverse-Wishart distribution (see, for instance, Box & Tiao, 1973, or Gelman, Carlin, Stern & Rubin, 1995). The prior follows from the specification

$$\begin{aligned} \Sigma_p &\sim \text{Inv-Wishart}_{v_0}(\Sigma_0) \\ \mu_p | \Sigma_p &\sim \text{MVN}(\mu_0, \Sigma_p / k_0) \end{aligned}$$

where Σ_0 and v_0 are the scale matrix and degrees of freedom for the prior on Σ_p and μ_0 and k_0 are the mean and weight for the prior on μ_p , respectively. The weight expresses the information in the prior distribution as the number of prior measurements it can be equated to.

Likelihood Function

The response vector of examinee n is denoted as $x_n = (x_{ni_1}, \dots, x_{ni_p}, \dots, x_{ni_p})$. Using the assumptions of (1) independence between examinees, (2) independence between items and examinees, and (3) local independence within examinees, the likelihood function associated with response data \mathbf{x} and test administration design \mathbf{d} can be written as

$$\begin{aligned} p(\theta, \xi, \mu, \Sigma; \mathbf{x}, \mathbf{d}) &= \prod_n p(x_n | \mathbf{d}_n, \theta_n, \xi, \mu, \Sigma) \\ &= \prod_n \prod_p \prod_{i_p} p(x_{ni_p} | d_{ni_p}, \theta_n, \xi_{i_p}) p(\theta_n) \\ &\quad \prod_p \prod_{i_p} p(\xi_{i_p} | \mu_p, \Sigma_p). \end{aligned} \quad (5)$$

The convention will be followed that $p(x_{ni_p} = r | d_{ni_p} = 0, \theta_n, a_{i_p}, b_{i_p}, c_{i_p}) = 1$.

Bayesian Estimation

An MCMC procedure will be used to sample from the posterior distribution. Only the essential steps will be outlined here, for details consult Glas and van der Linden (2005). Following Albert (1992) and Béguin and Glas (2001), two data augmentation schemes are used to create tractable posterior distributions. First, a binary variable W_{ni_p} is introduced with a conditional distribution given by

$$\begin{aligned} P(W_{ni_p} = 1 | X_{ni_p} = 1, \lambda_{ni_p}, c_{i_p}) &\propto \Phi(\lambda_{ni_p}) \\ P(W_{ni_p} = 0 | X_{ni_p} = 1, \lambda_{ni_p}, c_{i_p}) &\propto c_{i_p} (1 - \Phi(\lambda_{ni_p})) \\ P(W_{ni_p} = 1 | X_{ni_p} = 0, \lambda_{ni_p}, c_{i_p}) &= 0 \\ P(W_{ni_p} = 0 | X_{ni_p} = 0, \lambda_{ni_p}, c_{i_p}) &= 1, \end{aligned} \quad (6)$$

where $\lambda_{ni_p} = a_{i_p} \theta - b_{i_p}$. Second, the data are augmented with latent data Z_{ni_p} , which are independent and normally distributed with mean λ_{ni_p} and standard deviation equal to one.

The aim of the procedure is to simulate samples from the joint posterior distribution given by

$$p(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w} | \mathbf{x}) \propto p(\mathbf{z}, \mathbf{w} | \mathbf{x}; \xi, \theta) p(\theta) p(\xi | \mu, \Sigma) p(\mu, \Sigma | \mu_0, \Sigma_0). \quad (7)$$

The samples are generated with a Gibbs Sampler consisting of four steps.

Step 1: Draw from $p(\mathbf{z}, \mathbf{w} | \mathbf{x}; \xi, \theta)$

Step 2: Draw from $p(\theta | \mathbf{z}, \xi)$

Step 3: Draw from $p(\xi_{i_p} | \theta, \mathbf{z}_{i_p}, \mu_{i_p}, \Sigma_{i_p})$

Step 4: Draw from $p(\mu_p, \Sigma_p | \xi, \theta, \mathbf{z}, \mathbf{x})$.

Multiple MCMC chains can be started from different points to evaluate convergence by comparing the between- and within-sequence variance. Another approach is to generate a single MCMC chain and to evaluate convergence by dividing the chain into subchains and comparing between- and within-subchain variance. For these and other technical details, see Gelman et al. (1995).

Adaptive Testing

Three models for CAT with item clones will be considered. The first model uses the true item parameters ξ_{i_p} and ignores the relation with the item parents μ_p and Σ_p . The second model is analogous to the first

model, but it is based on point estimates of the item parameters ξ_{i_p} . The third model takes the uncertainty about the item parameter ξ_{i_p} and the parent item parameters μ_p and Σ_p into account by invoking their posterior distributions.

Model 1

Suppose items $1, \dots, k-1$ have been selected. The responses to which are denoted by a vector $\mathbf{y}_j^{(k-1)} \equiv (y_{j1}, \dots, y_{j(k-1)})$. The character y is used rather than the character x to distinguish the CAT responses from the responses in the calibration phase. In the present approach, the item clones are the selection objects. This is contrary to the approach studied by Glas and van der Linden (in press) where the parents were the selection objects. However, it will be assumed here that every clone selected is from a different parent.

The posterior distribution of θ_j given $\mathbf{y}_j^{(k-1)}$ is

$$f(\theta_j | \mathbf{y}_j^{(k-1)}) \propto p(\theta_j) \prod_{p=1}^{k-1} p(y_{jp} | \theta_j, \xi_{i_p}). \quad (8)$$

The posterior expectation of θ_j given $\mathbf{y}_j^{(k-1)}$ can be used as a running point estimate of ability with the posterior variance $\text{Var}(\theta_j | \mathbf{y}_j^{(k-1)})$ as a measure of precision.

The k th item is selected to be optimal at this posterior distribution. As in Glas and van der Linden (in press), the criterion for item selection will be the minimum expected posterior variance (also see, van der Linden, 1998, and van der Linden & Pashley, 2000).

The posterior predictive distribution of the response of examinee j to item i_p given the previous responses $\mathbf{y}_j^{(k-1)}$ is given by

$$f(y_{jp} | \mathbf{y}_j^{(k-1)}) = \int p(y_{jp} | \theta, \xi_{i_p}) f(\theta | \mathbf{y}_j^{(k-1)}) d\theta. \quad (9)$$

The two possible responses lead to updates of the posterior variance, which we denote as $\text{Var}(\theta_j | \mathbf{y}_j^{(k-1)}, X_{j\tilde{i}_p} = 0)$ and $\text{Var}(\theta_j | \mathbf{y}_j^{(k-1)}, X_{j\tilde{i}_p} = 1)$. The proposed criterion for the selection of the k th parent is the expected value of this update. That is,

$$i_p \equiv \arg \min_r \left\{ \text{Var}(\theta_j | \mathbf{y}_j^{(k-1)}, X_{j\tilde{i}_r} = 0) f(0 | \mathbf{y}_j^{(k-1)}) + \text{Var}(\theta_j | \mathbf{y}_j^{(k-1)}, X_{j\tilde{i}_r} = 1) f(1 | \mathbf{y}_j^{(k-1)}) \right\}; r \in R_k \quad (10)$$

where R_k is the set of parents in the pool in which no items are selected yet.

Model 2

Model 2 is analogous to Model 1, but the posterior expectation of ξ_p is used as a point estimate.

Model 3

In this case, the posterior distribution of θ_j given $\mathbf{y}_j^{(k-1)}$ is averaged over all values drawn from the posterior distribution $p(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w} | \mathbf{x})$ by the Gibbs Sampler. So the posterior distribution of θ_j is given by

$$f(\theta_j | \mathbf{y}_j^{(k-1)}) \propto \int \dots \int \left[p(\theta_j) \prod_{p=1}^{k-1} p(x_{jp} | \theta_j, \xi_{i_p}) h(\xi_{i_p} | \mu_p, \Sigma_p) \right] p(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w} | \mathbf{x}) d(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w}).$$

In the same manner, the posterior predictive distribution becomes

$$f(y_{jp} | \mathbf{y}_j^{(k-1)}) = \int \dots \int \left[\int p(y_{jp} | \theta, \xi_{i_p}) h(\xi_{i_p} | \mu_p, \Sigma_p) f(\theta_j | \mathbf{y}_j^{(k-1)}) d\theta \right] p(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w} | \mathbf{x}) d(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w}).$$

Using the draws from the Gibbs Sampler, evaluation of these two expressions boils down to a form of Monte Carlo integration. That is, $p(\xi, \theta, \mu, \Sigma, \mathbf{z}, \mathbf{w} | \mathbf{x})$ is not analytically evaluated but sampled from. Further, the draws of \mathbf{z} , \mathbf{w} , and θ do not play a role; only the draws of ξ , μ , and Σ are relevant. As above, item selection can be done using (10).

Simulation Studies

Simulation studies were conducted to compare the performance of the three CAT models.

Item Pools

The parameters of the item parent parameters μ_p and Σ_p were drawn from a prior distribution with mean

$$\mu_0 = (1.0, .0, \text{logit}(.2)) \quad (11)$$

and covariance matrix

$$\Sigma_0 = \begin{bmatrix} 0.20 & 0.05 & -0.05 \\ 0.05 & 1.00 & 0.10 \\ -0.05 & 0.10 & 0.10 \end{bmatrix} \quad (12)$$

and these draws were in turn used to generate values $(a_p, b_p, \text{logit } c_p)$.

The Calibration Phase

The simulations in the calibration phase were made with either $P = 20$ or $P = 40$ item parents. In all simulations, there were $K_p = 10$ operational items per parent. So the total number of operational items equaled $K = 200$ and $K = 400$. Every simulee responded to P operational items, one from each parent. The number of simulees responding to an operational item had three conditions: 50, 100 and 200. This resulted in total calibration sample sizes of 500, 1,000 and 2,000. Posterior distributions were generated with the Gibbs Sampler using 10,000 iterations.

The CAT Phase

The CAT phase had a factorial design. The three factors were the CAT model (Model 1, 2, or 3), the calibration sample size (500, 1,000 or 2,000), and the total item bank size (200 and 400), respectively. The fourth factor entailed the true ability value, which was fixed at -2.0 , -1.0 , 0.0 , 1.0 , or 2.0 , or drawn from a standard normal distribution. For each combination of calibration sample size and the total item bank size, five replications of the MCMC procedure were made, and within each replication, 200 CATs were simulated for each true ability condition and each CAT model. The test length was always equal to 20 items. Item selection was under the restriction that all items administered were from different parents. Model 3 was executed with 100 values from the 10,000 MCMC draws chosen at equally spaced iterations.

Results

The mean absolute errors (MEAs) in the ability estimates are shown in Table 1 ($K = 200$) and Table 2 ($K = 400$). The impact of estimation error in the item parameters can be assessed by comparing the first row (Model 1) with the other rows. As expected, the MEA decreases with increased sample size. Further, the MEAs of Model 3 are always smaller than the MEAs of Model 2.

TABLE 1
Mean absolute error in ability estimates

N	CAT Model	Item bank size $K = 200$					Standard Normal
		θ	-2	-1	0	1	
—	1	.526	.266	.263	.268	.432	.291
500	2	.621	.323	.325	.302	.538	.349
	3	.614	.311	.299	.287	.489	.333
1000	2	.616	.309	.306	.295	.497	.328
	3	.578	.296	.297	.282	.475	.310
2000	2	.566	.308	.298	.285	.473	.321
	3	.552	.291	.275	.282	.450	.306

TABLE 2
Mean absolute error in ability estimates

N	CAT Model	Item bank size $K = 400$					Standard Normal
		θ	-2	-1	0	1	
—	1	.501	.245	.228	.233	.401	.276
500	2	.810	.456	.274	.202	.484	.332
	3	.577	.356	.257	.202	.468	.312
1000	2	.596	.325	.261	.225	.465	.326
	3	.558	.266	.247	.202	.448	.296
2000	2	.576	.275	.255	.175	.437	.311
	3	.528	.256	.232	.202	.428	.282

The finding in the study by Glas and van der Linden (in press, 2005) that an increase in the size of the item bank can have a detrimental effect on the MEAs if the item parameters are poorly estimated is also corroborated here. In Model 2, the item clones are treated as fixed effects; that is, their nesting under parent items is not taken into account. Further, every item clone is only responded to by 10% of the calibration sample. Especially for a sample size of 500 simulees (with 50 simulees responding to each item clone), the effect is dramatic: the MEA is 0.810. The effect decreases with sample size. Further, Model 3 is far less vulnerable to this effect.

Conclusion

Statistical models for CAT with item clones that properly take all dependencies in the data into account are definitely more complex than the standard models for CAT. Also the computational effort is substantially greater. However, the advantages of using item clones in terms of item development cost may far outweigh the statistical complications.

As already demonstrated by van der Linden and Glas (2000, 2001), one of the main pitfalls in studies of the psychometric advantages of CAT is disregarding the fact that item parameters are estimated. Also in the present study, it is shown that the presence of highly discriminating but poorly estimated items threatens the precision of CAT and may even nullify the benefit of the availability of a large item bank. As van der Linden and Glas show, this effect of capitalization on error can be partially checked by cross-validation techniques. The present report shows that this can also be achieved by using Bayesian techniques that allow for modeling uncertainty with a full posterior distribution and where nesting of item clones under item parents results in shrinkage estimators that counteract spuriously high discrimination parameters.

References

- Albert, J. H. (1992). Bayesian estimation of normal-ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 17, 261–269.
- Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541–562.
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. I. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–357). Hillsdale, NJ: Lawrence Erlbaum.

-
- Box, G., & Tiao, G. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison Wesley.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.
- Glas, C. A. W., & van der Linden, W. J. (in press). *Modeling adaptive testing using item cloing* (Computerized Testing Report 01-03). Newtown, PA: Law School Admission Council.
- Glas, C. A. W., & van der Linden, W. J. (2005). *Modeling variability in item parameters in educational measurement* (Computerized Testing Report 01-07). Newtown, PA: Law School Admission Council.
- Roid, G., & Haladyna, T. (1982). *A technology for test-item writing*. New York: Academic Press.
- van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, 63, 201–216.
- van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, 13, 35–53.
- van der Linden, W. J., & Glas, C. A. W. (2001). Cross-validating item parameter estimation in adaptive testing. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 205–219). New York: Springer.
- van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Norwell, MA: Kluwer Academic.