ED 475 144	TM 034 186

AUTHOR	Rizavi, Saba; Way, Walter D.; Davey, Tim; Herbert, Erin
TITLE	Tolerable Variation in Item Parameter Estimates.
INSTITUTION	Educational Testing Service, Princeton, NJ.
PUB DATE	2002-04-00
NOTE	39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Adaptive Testing; *Estimation (Mathematics); High Stakes Tests; *Test Items
IDENTIFIERS	*Item Parameters

ABSTRACT

The purpose of this study was to investigate and to quantify the tolerable error in item parameter estimates for different sets of items used in computer-based testing. The study examined items that were administered repeatedly to different examinee samples over time, examining items that were administered linearly in a fixed order each time they were used and items that appeared in different adaptive item testing pools each time they were used. In the first study, two sets of items from a high-stakes test, one set from the verbal test and one set from the quantitative test, were chosen, with sample sizes varying from 627 to 2,305 for the quantitative measure and 2,284 for the verbal measure. The second part of the study used items from another high stakes admissions test using 30 items with sample sizes larger than 500. Linearly administered items in a high-stakes testing program exhibited remarkably small variation in parameter estimates over repeated calibrations. Results also indicated that context effects played a more significant role in adaptive item parameters when comparisons were made to the parameters that were obtained from paper-and-pencil testing. This suggests that, whenever feasible, the parameter estimates obtained from paper-and-pencil administrations be replaced with computer-based testing calibrated parameters. Two appendixes contain the tables and figures that show sample sizes and parameter estimates. (Contains 4 tables, 18 figures, and 25 references.) (SLD)

Tolerable Variation in Item Parameter Estimates

Saba Rizavi

Walter D. Way

Tim Davey

Erin Herbert

Educational Testing Service

Paper presented at the Annual Meeting of the National Council on Measurement in Education,

New Orleans, LA, April 2002

Tolerable Variation in Item Parameter Estimates

Introduction

Computer-based testing, and adaptive testing in particular, typically depends upon item response theory (IRT). The advantages of IRT are well known in the testing literature and have fueled the transition of computerized adaptive testing (CAT) from a research interest to a widely used practical application. However, the introduction of computer-based testing in high volume, high stakes settings has presented new challenges to testing practitioners. In most computer-based testing programs, it is necessary to administer the same items repeatedly over time. This continuous item exposure raises security concerns that were not fully appreciated by researchers when the theory and practice of CAT was first developed.

In most CAT programs, steps are taken to protect the integrity of item pools through strategies such as item exposure control, pool rotation, and accelerated item development (Way, 1998). Despite such efforts, maintaining CAT programs remains difficult because adaptive algorithms tend to select the most highly discriminating items. As these items become exposed and are retired from use, it is very difficult to develop sufficient replacement items of the same quality. Efforts to increase item development bring increased costs and diminishing returns, as for every highly discriminating item that is retired from over-use, three or four items may need to be written to find a suitable replacement. Furthermore, the lag time between initially writing items and using the items in an operational CAT pool is usually significant, as items must be pretested, calibrated, and evaluated before they may be used operationally.

Recently, researchers at ETS have begun exploring an approach to adaptive testing that could address some of the challenges of item exposure and pool maintenance (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2002). This approach has been referred to by Bejar (1993) as generative testing and more recently has been called *item modeling*. The essence of item modeling is to create items from explicit and principled rules. The approach has roots in computer assisted instruction and domain referenced testing (Hively, 1974). The obvious vehicle for item modeling is the computer, and successful applications of automated item generation have been reported by a number of researchers (Embretson, 1999; Irvine, Dunn, & Anderson, 1990; Irvine & Kyllonen, 2001).

Although the capability to develop item models and generate items automatically is more easily established for some item types than for others, the potential utility of automated item generation for supporting computer-based testing is obvious. An effective item model provides the basis for a limitless number of items, each of which is assumed to share the same content and statistical characteristics. In CAT, the adaptive algorithm could choose an item model based on the common psychometric characteristics, and the actual instance of the item would be generated at the time of delivery. Such an approach was referred to as "on-the-fly" adaptive testing by Bejar et al. (2002). They carried out a feasibility study of a CAT application where item models were utilized, and concluded that the adaptive generative model they employed was both technically feasible and cost effective.

From a traditional IRT perspective, the use of item models with adaptive testing seems far-fetched. In fact, much of the IRT literature in recent years has centered on item parameter estimation and parameter recovery, the idea being that successful applications of IRT depend upon well-estimated parameters. The notion that one could use a single set of IRT estimates to characterize all of the items generated from a particular model directly contradicts the goal of obtaining accurate item parameter estimates. However, such a perspective does not account for the variation that may occur in student scores due to a variety of effects that influence how test items are responded to in the real world. These include context effects, item position effects, instructional effects, variable sample sizes and other sources of item parameter drift that are typically not formally recognized or controlled for in the context of CAT.

Several researchers have documented the existence and influence of such item level effects. Sireci (1991) looked at the effect of sample sizes on the stability of IRT item parameter estimates. Kingston and Dorans (1984) described such effects in equating the paper-and-pencil GRE. Leary and Dorans (1986) and Brennan (1992) reviewed literature related to context effects and provided guidelines on how such effects might be minimized. Zwick et al. (1991) described a case study of how context effects created an anomaly in Reading scores on the National Assessment of Educational Progress. Divgi and Stoloff (1986) documented changes in item parameter estimates in an early application of the CAT-ASVAB. Several researchers have investigated causes of item parameter drift in testing programs that utilized IRT in test construction and equating over time (Way, Carey, and Golub-Smith, 1992); Sykes et al., 1996).

In considering the viability of item models for CAT, we recognize that variation within models introduces a source of errors that is not present in traditional CAT. However, the repeated use of the same items across different CAT pools also introduces a source of errors that is tolerated but not accounted for. The purpose of this study was to investigate and to quantify the tolerable error in item parameter estimates for different sets of items used in computer-based testing. The study examined items that were administered repeatedly to different examinee samples over time. We examined both items that were administered linearly in a fixed order each time that they were used and items that appeared in different adaptive testing item pools each time they were used. We examined both the magnitude of variation in the item parameter estimates and the impact of this variation in the estimation of test taker scores.

Case Study 1:

Data

In order to carry out the investigation of the stability of parameter estimates in linearly administered tests, two sets of items from a high stakes admissions test were chosen. The first set comprised of 28 items from the Quantitative measure while the second set consisted of 30 items from the Verbal measure of the same test. Since ability distributions for the Quantitative measure are known to change more rapidly than the other measures, a greater variation in the parameter estimates was expected for that measure (the ability scale for all testing measures considered in this paper range from –5 to 5).

The items contained in the two sets were actually used as anchors in order to equate the linearly administered Pretest sections that are administered along with the Operational CAT sections. In every online calibration for these CAT programs, a set of anchors is administered in a similar way as pretest items. The composition of the anchor set mirrors the pretests in terms of psychometric and content characteristics. The number of items in the pretest and anchor set is the same. The verbal and quantitative anchor items evaluated in the study were used over a two-year period and were calibrated on each administration of the corresponding pretest measure. Thus 9 repeated calibrations were available for each anchor item.

The sample sizes used to calibrate each item varied from 627 to 2305 for the Quantitative measure and 839 to 2284 for the Verbal measure. The details of sample sizes used for each calibration are presented in table 1 in appendix A.

Estimation Methodology

The item parameter estimates were obtained using the software LOGIST (Wingersky, Patrick, & Lord, 1988). LOGIST uses Joint Maximum Likelihood estimation methodology to estimate item parameters, keeping the ability parameter fixed while estimating items parameter estimates. The ability parameters in this case were the actual ability estimates obtained on the operational section of the test. The estimates on the linearly administered items were then subjected to scaling using the Test Characteristic Curve methodology proposed by Lord and Stocking (1983). In this study, the stability of both sets of estimates on anchor items was looked at both before and after the scaling was carried out.

Design and Analyses

In order to look at the general trends in the variation of individual parameter estimates the *a*, *b*, and *c* parameters were plotted for each item across calibrations. The purpose of this analysis was simply to get an idea of any directional change that could occur in some items over time. In order to look at the effect of parameter estimate variation on the probability of getting an item correct, the Item Characteristic Curves were examined for each item across 9 calibrations for both measures. The Root Mean Squared Errors (RMSE) were then computed between the Item Characteristic Curves for the various calibrations in relation to the first calibration. In other words the first calibration of all was chosen as a point of reference in this case. The Root Mean Squared Error in this case is defined as,

$$RMSE_{ic} = \sqrt{\sum_{j=1}^{61} \frac{(P_{ic}(\theta_j) - P_{i1}(\theta_j))^2}{61}} \tag{1}$$

Where $P_{ic}(\theta_j)$ is the probability of getting an item (*i*) correct in a calibration (*c*) at an ability level θ_j. The 61 ability levels ranged from –3 to +3 where most of the population is known to concentrate. This index was used in similar research performed in ETS where ICCs obtained on different calibrations were compared (Guo et. al, 2001; Rizavi & Guo, 2002). The RMSEs were then plotted for each item across calibrations to capture variation for items. This analysis was repeated for both Pre- and Post- adjustment of the estimates after scaling.

Another interesting way to look at the variation is to estimate the variance-covariance matrix of item parameter estimates. Several alternatives are available for computing the sampling variances of item parameter estimates. The first is to use standard large-sample theory, which holds that the asymptotic variances of $<\hat{a},\hat{b},\hat{c}>$ are given by the inverse of the 3x3 Fisher information matrix evaluated at the true parameter values <*a,b,*c> (Lord, 1980; Hambleton, Swaminathan and Rogers, 1991) defined as,

$$\Sigma_i = \begin{bmatrix} I_{\mathrm{a}} & I_{\mathrm{ab}} & I_{\mathrm{ac}} \\ I_{ab} & I_b & I_{\mathrm{bc}} \\ I_{ac} & I_{\mathrm{bc}} & I_c \end{bmatrix} \qquad (2)$$

The problem, of course, is that the true parameters are unknown. Our best approximation is then to evaluate information at the values of the parameter estimates $<\hat{a},\hat{b},\hat{c}>$ and hope these are reasonably close to the true values. Substituting estimates for true values when computing information is likely to understate the true sampling fluctuation of the parameter estimates (Andersen, 1973). It is also true that the asymptotic sampling variances truly apply only with very large samples. Finally, item parameter estimates are often constrained to avoid their taking on inappropriate values (e.g. negative a-parameters or c-parameters outside the range $[0,1]$). Such constraints are liable to upset asymptotic theory and render the sampling variance approximations less valid.

In the current situation, we have a second means available for estimating sampling variation. The items under study were administered operationally on nine separate occasions and parameter estimates were separately obtained from each administration sample. The observed variation across these estimates is therefore an empirical estimate of the sampling fluctuation of the parameter estimates defined as,

$$\Sigma_i = \begin{bmatrix} \sigma_{\mathrm{a}}^2 & \sigma_{\mathrm{ab}} & \sigma_{\mathrm{ac}} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{\mathrm{bc}} \\ \sigma_{ac} & \sigma_{\mathrm{bc}} & \sigma_c^2 \end{bmatrix} \qquad (3)$$

In theory, and under all of the assumptions of that theory, the empirical and asymptotic estimates of sampling variation should be very similar. However, the empirical variances are only based on nine observations and so may not be very stable. Both asymptotic and empirical sampling variance estimates are therefore problematic to some extent. It was therefore decided to repeat the analyses with both.

The last and the most affirming set of analyses were performed to look at the effect of variation in the item parameter estimates on the actual reported scores. The responses of examinees on the anchor items were selected for the nine sets of calibrations on both measures. A typical ability distribution for the examinees during an administration is given in figure A.6 for both Quantitative and Verbal. The perfect response patterns were excluded from the response sets. The response set was then scored using the set of item parameter estimates obtained on it (true estimates) as well as the ones obtained on the rest of the eight sets of responses. The same process was repeated for each response set thus producing 81 sets of scores. The scoring was carried out using Maximum Likelihood Estimation methodology (Lord, 1980; Hambleton, Swaminathan and Rogers, 1991). RMSE statistics between the theta estimates or scores obtained using the true parameter estimates and the estimates from the other eight sets were then computed. The statistic was defined as,

$$RMSE_{ck} = \sqrt{\sum_{j=1}^{n_k} \frac{(\hat{\theta}_{cj} - \hat{\theta}_{kj})^2}{n_k}} \tag{4}$$

Where $\hat{\theta}_{kj}$ is the ability estimate obtained for an examinee *j* on examinee set (or response set) *k* using item parameter estimates obtained by calibrating response set *k*. $\hat{\theta}_{cj}$ on the other hand is the ability estimate for an examinee *j* on examinee set *k* using item parameter estimates obtained by calibrating response set *c*.

The ability estimates were then mapped on to the reported score scale and the distributions of differences *($Score_{cj}$-$Score_{kj}$)* for each of the 81 scenarios were plotted.

Case Study 2:

The second part of this investigation was carried out on a set of adaptively administered operational items from another high stakes admissions test. This particular program uses the item specific prior methodology with a proprietary version of computer software PARSCALE (Muraki and Bock, 1999). This methodology allows unique multivariate normal distributions to be used as prior distributions for the parameters of each item (Swaminathan and Gifford, 1986; Folk & Golub-Smith, 1996). These item specific priors are actually the mean estimates of the [b,a,c] parameters as well as the asymptotic variance-covariance matrix specified as [I, a, c] where *I* is the intercept. The mean vector and the variance-covariance matrix of the parameters for an item *i* are define as,

$$\mu_i = \begin{bmatrix} \overline{b}_i \\ \overline{a}_i \\ \overline{c}_i \end{bmatrix} \qquad (5)$$

$$\Sigma_i = \begin{bmatrix} \sigma_{\mathrm{I}}^2 & \sigma_{\mathrm{Ia}} & \sigma_{\mathrm{Ic}} \\ \sigma_{aI} & \sigma_a^2 & \sigma_{\mathrm{ac}} \\ \sigma_{cI} & \sigma_{\mathrm{ca}} & \sigma_c^2 \end{bmatrix} \qquad (6)$$

These priors are used for the CAT operational items and are different for each item, as they are item specific. On the other hand, global priors are used for the pretest items and are the same for all pretest as well as anchor items. The global priors are the weighted estimated means of the *a*, *b*, and *c* parameters for similar items in the item bank. For *a*, log of the mean is used. All pretest, anchor and CAT items are calibrated together for an administration. Unlike the previous case where a pretest or anchor set is offered as a separate section, pretest or anchor items are actually embedded in the operational test in this case. Since the priors on the CAT items are strong, their values hardly move away from their original values. The CAT items therefore set the scale thus putting all items on the same scale. Once calibrated, the operational item parameter estimates are stored as a flat file and are not used further while the pretest item estimates are actually stored in the item bank to be used in subsequent pools. This methodology has been shown to be effective in addressing the issue that sample sizes for operational items available for on-line calibration are often small across all ability levels except the ones close to the item difficulty.

Data

The data for this investigation came from the Quantitative measure of an adaptively-administered high stakes admissions test. In order to avoid the sample size issue as mentioned above, 30 slightly easy, mid-difficulty and slightly difficult items with sample sizes larger than 500 were chosen. Those items had already appeared in pools and had been calibrated originally with item specific priors on them. The item-ability regression plots were examined for these items to make sure that the range of abilities taking each of those items was not restrictive. Another reason for choosing the items was that the items had appeared in several pools and had at least 8 calibrations available on them. The number of calibrations available on items is given in table 1,

Table 1: Number of Calibrations

No. of Items	No. of Calibrations
5	8
12	9
7	10
6	11

For the purpose of this investigation, the item specific priors were removed and global priors were imposed on these CAT items thus treating them like other pretest items. The modified requests for the calibration were re-submitted using GENASYS, an ETS specific software that uses PARSCALE for calibration. Items were then calibrated and new estimates were obtained. The original estimates (used as item specific priors for the subsequent calibrations) for these items were originally obtained when they had appeared in paper-and-pencil (P & P) administrations before the introduction of CAT. The mean and standard deviations for the original *a*, *b*, and *c* parameters are give in table 2.

Table 2: Mean and Standard Deviations for the *a*, *b*, and *c* parameters (original P & P)

	a	*b*	*c*
Mean	1.07	0.23	0.16
Stdev	0.19	0.72	0.05

Design and Analyses

Similar to the previous case study, the item characteristic curves were examined for each item. The RMSEs were then computed between the ICCs for the first calibration compared with the other calibrations as discussed in the previous study. The first calibration of all was chosen as the point of comparison.

The next part of the analyses involved looking at the effect of variation in parameter estimates on ability estimation. Unlike the linear case, where a fixed number of calibrations were available on each item, the number of calibrations varied in this case. Thus 20 sets of item parameter estimates were generated for each item by drawing parameters at random from the various calibrations available for that item (except the first calibration). A response set was obtained by generating 500 examinees at 11 ability levels ranging from –2.5 to 2.5 with an increment of 0.5. The item parameter estimates used to generate the response set came from the first calibration. The response set was scored using item parameter estimates from the first calibration as well as the 20 other sets of estimates. The first set of scores was then compared to the other 20 sets of scores. RMSEs were computed between the various sets of ability estimates as in the previous study. The ability estimates were then converted to scaled or reported scores and the distributions of differences between those scores obtained using various sets of estimates were compared.

Next, the scoring analyses were repeated by generating a response data using the item bank parameters for these items. As mentioned earlier, these parameters are used as the item specific priors if it was a real calibration and was originally obtained on P&P calibrations. These analyses were expected to reveal more variation in scores due to additional P&P context effects. In real calibrations, since these estimates are used as priors for the corresponding items, hence it's important to know whether such context effects are apparent in the estimation. The response set was then scored using the same set of item parameter estimates as well as the remaining 20 sets of estimates.

Results

The results of the analyses on linearly administered items are presented in appendix A. The plots of ICCs in figures A.1 and A.2 show that the probabilities of getting an item right did not vary much across calibrations except at the very high ends of the scale. The Test Characteristic Curves (TCC) for the set of anchor items for the two measures are also shown in figure A.3 for both pre-TBLT and post-TBLT scenarios. The TCCs for both measures were extremely close under both scenarios. A slight difference between the TCCs in the middle of the curve was adjusted by applying the TBLT methodology to bring them closer to a single old form that had been administered earlier. Some variations at the tails of the curve are the characteristic of the interaction between the abilities of examinees and difficult level of the items. Those variations are also shown in the ICCs in the same figures. The investigation of the general trends did not exhibit any directional change in the estimates. In other words, none of the items exhibited a systematic decrease or increase in the parameter estimates over repeated calibrations.

The RMSEs between ICCs are shown in figure A.4. The RMSE values indicated a very small variation between calibrations for both Quantitative and Verbal measures. The differences were slightly higher for Quantitative especially for some of the items. An item that appeared to be most variant in Quantitative was an item with a very high difficulty level. Inspecting the sample sizes and ability distribution for that particular calibration of that item did not suggest any explanation beyond chance occurrence. The RMSE values for the post-TBLT results were also evaluated but are not presented, as the differences in results between Pre- and Post TBLT were miniscule.

In comparing the empirical vs. Model based variation averaged across items (figure A.5), we found that the Model based variation was larger than the empirical variation for both Quantitative and Verbal measure for the b-parameter. The Model-based variance was highly affected by the magnitude of the b-parameter; very low b-parameters resulted in large values of Model-based variance. In the case of a-parameter, Model based variance was larger than empirical for Quantitative measure while smaller for the Verbal measure. The a-parameters for Verbal were in general higher than Quantitative measure. In general, the results did indicate very small Model-Based and empirical variation in both a- and b- parameters except for the Model based variance in b-parameter for Quantitative. The extremely low bs for some of the Quantitative items caused this variance as very little information is provided by these items. In general, these results should be interpreted with caution, as the sample sizes for the analyses were smaller than suitable.

The results of scoring using the different sets of parameter estimates are presented in figures A.7 to A.10. The results indicated that the RMSEs in ability estimates ranged from 0.11 to 0.22 in most of the cases where a response set used in a calibration was scored by an item parameter set obtained on other response sets. Results of two such scenarios are shown in figure A.7 where Quantitative response sets 1 and 9, respectively, are scored using item parameter estimate sets obtained on other response sets. The figures show that the error in estimates when scored using different sets of parameter estimates remained consistent across calibrations.

The differences in examinee reported scores when scored using different sets of item parameter estimates remained limited to a 0 to 2 point difference on the reported score scale for majority of the examinees. Figure A.8 illustrates the same fact. The first part of the figure shows the response set 2 scored using item parameter estimates obtained on response set 1. The second part shows the response set 9 scored using item parameter estimates obtained on response set 5. The x-axis in the chart represents the score points; 10 points mean a1-point difference on the reported score scale, 20 means a 2-point difference and so on. In the first scenario, 93% of the examinees exhibited a 0 to 2 point difference in their reported scores, while 91% showed this difference for the second scenario. Similar results for Verbal are shown in figures A.9 and A.10. The percentages of examinees exhibiting 0 to 2 point score difference ranged from 89 to 95% (92% on average) for Quantitative and from 91 to 98% (95% on average) for Verbal.

The results for the adaptively administered items are presented in appendix B. The investigation of the actual ICCs (figure B.1 --- the thick black line labeled VAT indicates the original P& P parameters for those items) and the RMSEs between ICCs for the 20 adaptive calibrations in comparison with first adaptive calibration revealed remarkably small variation. The RMSEs of the ICCs are shown in figure B.2. Table B.1 and the plot in figure B.2 signify that the RMSEs for those items were very close in magnitude to the linearly administered items. The values remained in the range of 0.1 and 0.2 for all items for all calibrations. The consistency of the RMSEs across calibrations is depicted by figure B.3. When investigated per ability level (figure

B.4), a large portion of the error seemed to concentrate in the low ability levels. The differences in reported scores for the adaptive case ranged from 0 to 2 points for 91% to 97% (94% on average) of examinees for 20 sets of calibrations when compared with scores based on first calibration.

When P&P calibrated estimates of the items were used in place of the first calibration for comparison between calibrations, the results were quite different. Figure B.6 shows the RMSEs between theta estimates obtained on the P&P calibrated sets of parameter estimates and 20 sets of estimates obtained on CBT calibrations. The results indicate an increase of overall RMSEs, when abilities obtained using P&P estimates were used for comparison. While for the scenario where comparisons were based on 1st calibration and the overall RMSEs between scores ranged from 0.10 to 0.17, here the errors ranged from 0.17 to 0.22. The errors due to calibration in the scores when compared across ability levels remained significantly small at the middle ability levels, higher for the high ability levels and highest for the low ability levels. The errors were as high as 0.3 at the lower ability levels. The percentage of examinees that exhibited reported score differences between 0 to 2 points on the reported score scale ranged from 87% to 92% (89% on average). This percentage was considerably smaller than the previous scenarios where most of the calibrations resulted in more than 89% of the examinees exhibiting 0 to 2 point difference. In other words, the percentage of examinees whose scores changed by more than 2 points was significantly large in this case.

Conclusions

The studies discussed in this paper aimed to investigate the effect of stability of item parameter estimation in the current CBT calibrations. The results of the study will serve as a baseline for the design work involved in creating models for automated item generation. The concept of having a single model to generate a family of items should be informed by knowing the relative stability of the parameter estimates when calibrated on-line.

Several conclusions can be drawn from the results of this study. The linearly administered items in a high stakes testing program exhibited remarkably small variation in parameter estimates over repeated calibrations. Although the sample sizes upon which the calibrations were performed varied considerably, the results were not affected. As long as the sample sizes are large enough to calibrate, the calibrations produce stable results. The stability was observed in the items even before they were scaled to an old form. Similar findings with adaptively administered items in another high stakes testing program were also found when initial adaptively-based item parameter estimates were compared with estimates from repeated subsequent use. These findings have implications for research on item modeling because they suggest that the use of item modeling with operational CAT programs will clearly introduce variation in ability estimation beyond what is currently present due to item context effects. It will be important to quantify and account for these sources of variation as this research progresses.

The results of this study also indicated that context effects played a more significant role in adaptive item parameters when the comparisons were made to the parameters that were obtained from paper-and-pencil testing. This suggests that whenever feasible, the parameter estimates obtained on paper-and-pencil administrations be replaced with the CBT calibrated parameters. The approach employed for this paper, that is, freeing the item specific priors that constrain item parameter estimates for selected operational items during the process of pretest item calibration, is one possible alternative for this kind of updating. However, further research with this approach would be necessary to determine if it would feasible in the context of an ongoing, operational CAT program.

References

Anderson, B. E. (1973). Conditional inference for multiple-choice questionnaires. The British Journal of Mathematical and Statistical Psychology, 26, 31-44.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., & Bennet, R. E. (2002). A feasibility study of On-the-Fly Adaptive Testing. GRE Report # 98-12. Princeton, NJ: Educational Testing Service.

Brennan, R. (1992). The context of context effects. Applied Measurement in Education, 5, 225-264.

Divgi, D. R. (1986). Determining the sensitivity of CST-ASVAB scores to changes in item response curves with the medium of administration. Alexandria, VA: Center for Naval Analyses.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. Psychometrika, 64, 407-433.

Folk, V. G., & Golub-Smith, M. (April, 1996). Calibration of On-Line Pretest Data Using Bilog. Paper presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Golub-Smith, M. (April, 1996). Challenges of On-Line Calibration and Scaling with Multilingual Examinee Population. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Guo, F., Stone, E., & Cruz, D. (April, 2001). On-line calibration using PARCALE Item Specific Prior method: changing test population and sample size. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Hambleton, R., Swaminathan, H. and Rogers, J. (1991). Fundamentals of item response theory. London: SAGE.

Hively, W. (1974). Introduction to domain-reference testing. Educational Technology, 14 (6), 5-10.

Irvine, S. H., & Kyllonen, P. (Eds.). (2001). Item generation for test development. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Irvine, S. H., Dunn, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. British Journal of Psychology, 81, 173-195.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. Review of Educational Research, 55, 387-413.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Muraki, E. & Bock R. (1999). PARSCALE 3.5: IRT item analysis and test scoring for rating-scale data. Scientific Software, Inc.

Rizavi, S., & Guo, F. (2001). Investing the stability of current GRE anchors (work in progress). Princeton, NJ: Educational Testing Service.

Sireci, S. G. (October, 1991). "Sample-Independent" Item Parameters? An investigation of the stability of IRT item parameters estimated from small data sets. Paper presented at the annual meeting of Northeastern Educational Research Association, Ellenville NY. (ERIC Document Reproduction Service No. Ed 338 707)

Stocking, M., & Lord, F. M. (1983). Developing a common metric in Item Response Theory. Applied Psychological Measurement, 7, 201-210.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. Psychometrika, 51, 589-601.

Sykes, R.C., & Fitzpatrick, A.R. (1992). The stability of IRT b-values. Journal of Educational Measurement, 29, 201-211.

Way, W. D., Carey, P. A., & Golub-Smith, M. L. (1992). An exploratory study of characteristics related to IRT item parameter invariance with the Test of English as a Foreign Language (TOEFL Technical Report # TR-6). Princeton, NJ: Educational Testing Service.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and Practice, 17, 17-26.

Wingersky, M., Patrick, R. & Lord, F. M. (1988). LOGIST: Computer software to estimate examinee abilities and item parameters. Educational Testing Service, Princeton, NJ.

Zwick, R. (1991). Effects of Item Order and Context on Estimation of NAEP Reading Proficiency. Educational Measurement: Issues and Practice, 10, 10-16.

Appendix A

Table A.1: Sample Sizes for each calibration

Calibration	Total Sample	Act. Sample Size per anchor item	# of perfect scores	Final Sample
1	6656	1299	8	1291
2	10178	1420	15	1405
3	16311	1182	7	1175
4	20018	1115	11	1104
5	6038	833	8	825
6	17949	1432	6	1426
7	19863	2323	18	2305
8	16493	858	14	844
9	20422	636	9	627
1	13632	2287	3	2284
2	8774	1066	2	1064
3	13329	992	0	992
4	14697	1118	4	1114
5	15151	1047	2	1045
6	11026	876	3	873
7	2130	1569	2	1567
8	5869	834	4	830
9	24945	939	2	937

Table A.2: Average Item Parameter Estimates (*a*, *b*)

Calib	Pre-TBLT				Post-TBLT			
	a-parm		b-parm		a-parm		b-parm	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
1	0.9160	0.372	-0.1903	1.070	0.842	0.342	-0.0199	1.163
2	0.9013	0.368	-0.1940	1.083	0.852	0.348	-0.0076	1.145
3	0.8728	0.359	-0.2175	1.103	0.826	0.340	-0.0280	1.166
4	0.8260	0.395	-0.2744	1.241	0.764	0.365	-0.0415	1.341
5	0.8331	0.324	-0.2286	1.153	0.764	0.297	-0.0300	1.256
6	0.8804	0.307	-0.1488	1.069	0.779	0.272	0.0157	1.208
7	0.8762	0.363	-0.2265	1.153	0.818	0.339	-0.0465	1.235
8	0.8432	0.374	-0.2937	1.168	0.755	0.335	-0.1815	1.305
9	0.8353	0.322	-0.1695	1.090	0.775	0.299	-0.0355	1.175
1	0.8707	0.254	0.0261	1.324	1.003	0.292	-0.0903	1.150
2	0.8799	0.259	0.0625	1.347	0.983	0.290	-0.0342	1.205
3	0.8301	0.243	0.0157	1.405	0.954	0.279	-0.0652	1.222
4	0.8306	0.232	0.1517	1.430	0.967	0.270	0.0354	1.229
5	0.8358	0.237	0.0663	1.378	1.024	0.291	0.0449	1.124
6	0.9442	0.262	0.1060	1.265	1.129	0.313	0.0464	1.059
7	0.8230	0.235	0.0637	1.314	0.970	0.278	-0.0095	1.114
8	0.8453	0.250	0.0876	1.335	1.020	0.301	0.0099	1.107
9	0.8256	0.238	-0.0301	1.371	0.954	0.275	-0.0607	1.186

Table A.3: RMSEs in ICCs for Quantitative measure

	2	3	4	5	6	7	8	9
1	0.04	0.02	0.02	0.01	0.01	0.03	0.03	0.03
2	0.04	0.01	0.05	0.02	0.05	0.02	0.02	0.04
3	0.02	0.03	0.05	0.02	0.03	0.03	0.03	0.08
4	0.04	0.03	0.03	0.03	0.01	0.02	0.02	0.03
5	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.05
6	0.04	0.01	0.03	0.05	0.02	0.00	0.02	0.05
7	0.03	0.03	0.04	0.02	0.03	0.03	0.06	0.03
8	0.02	0.02	0.00	0.02	0.04	0.01	0.04	0.03
9	0.01	0.02	0.03	0.05	0.04	0.02	0.01	0.03
10	0.04	0.03	0.06	0.03	0.04	0.04	0.04	0.05
11	0.01	0.04	0.06	0.07	0.03	0.03	0.05	0.03
12	0.02	0.04	0.03	0.04	0.03	0.03	0.10	0.03
13	0.02	0.02	0.04	0.06	0.02	0.02	0.03	0.03
14	0.03	0.02	0.03	0.04	0.03	0.03	0.03	0.03
15	0.04	0.03	0.04	0.07	0.03	0.03	0.06	0.03
16	0.05	0.06	0.07	0.03	0.05	0.03	0.06	0.07
17	0.04	0.02	0.05	0.03	0.02	0.01	0.07	0.02
18	0.08	0.05	0.04	0.05	0.03	0.05	0.04	0.03
19	0.04	0.07	0.19	0.05	0.04	0.04	0.03	0.04
20	0.03	0.03	0.06	0.02	0.04	0.03	0.06	0.04
21	0.02	0.02	0.04	0.04	0.03	0.04	0.03	0.04
22	0.01	0.02	0.01	0.03	0.03	0.01	0.06	0.03
23	0.04	0.04	0.02	0.04	0.02	0.03	0.03	0.03
24	0.04	0.01	0.02	0.03	0.05	0.02	0.04	0.05
25	0.02	0.03	0.03	0.04	0.03	0.03	0.02	0.03
26	0.01	0.02	0.01	0.01	0.04	0.01	0.01	0.03
27	0.02	0.06	0.04	0.11	0.06	0.03	0.03	0.02
28	0.01	0.02	0.07	0.03	0.03	0.03	0.03	0.03

Table A.4: RMSEs in ICCs for Verbal measure

	2	3	4	5	6	7	8	9
1	0.01	0.05	0.04	0.02	0.02	0.02	0.02	0.01
2	0.03	0.08	0.03	0.03	0.05	0.05	0.05	0.04
3	0.03	0.13	0.05	0.04	0.03	0.03	0.04	0.04
4	0.04	0.10	0.07	0.06	0.07	0.07	0.05	0.05
5	0.01	0.07	0.02	0.09	0.07	0.07	0.05	0.02
6	0.04	0.06	0.10	0.05	0.05	0.05	0.08	0.06
7	0.06	0.01	0.02	0.03	0.02	0.02	0.02	0.04
8	0.07	0.09	0.02	0.07	0.03	0.03	0.02	0.10
9	0.03	0.03	0.03	0.06	0.07	0.07	0.04	0.04
10	0.03	0.03	0.04	0.04	0.05	0.05	0.01	0.04
11	0.00	0.02	0.03	0.07	0.03	0.03	0.04	0.03
12	0.05	0.06	0.04	0.10	0.05	0.05	0.06	0.09
13	0.01	0.04	0.02	0.05	0.00	0.00	0.02	0.03
14	0.02	0.04	0.03	0.05	0.03	0.03	0.04	0.04
15	0.02	0.02	0.07	0.02	0.02	0.02	0.03	0.04
16	0.02	0.05	0.02	0.04	0.03	0.03	0.05	0.04
17	0.05	0.04	0.04	0.05	0.09	0.09	0.03	0.04
18	0.03	0.00	0.02	0.04	0.01	0.01	0.03	0.03
19	0.04	0.03	0.03	0.03	0.04	0.04	0.04	0.03
20	0.05	0.03	0.04	0.06	0.05	0.05	0.07	0.03
21	0.03	0.03	0.04	0.05	0.06	0.06	0.05	0.03
22	0.03	0.01	0.02	0.02	0.04	0.04	0.04	0.03
23	0.02	0.02	0.03	0.05	0.05	0.05	0.02	0.06
24	0.04	0.06	0.06	0.04	0.06	0.06	0.03	0.02
25	0.02	0.03	0.04	0.05	0.03	0.03	0.04	0.04
26	0.04	0.02	0.02	0.04	0.02	0.02	0.03	0.05
27	0.05	0.04	0.06	0.03	0.02	0.02	0.03	0.04
28	0.03	0.04	0.04	0.02	0.05	0.05	0.03	0.06
29	0.04	0.09	0.11	0.05	0.08	0.08	0.05	0.07
30	0.15	0.05	0.06	0.12	0.11	0.11	0.08	0.09

Figure A.1: Item Characteristic Curves for 4 Quantitative items over 9 calibrations

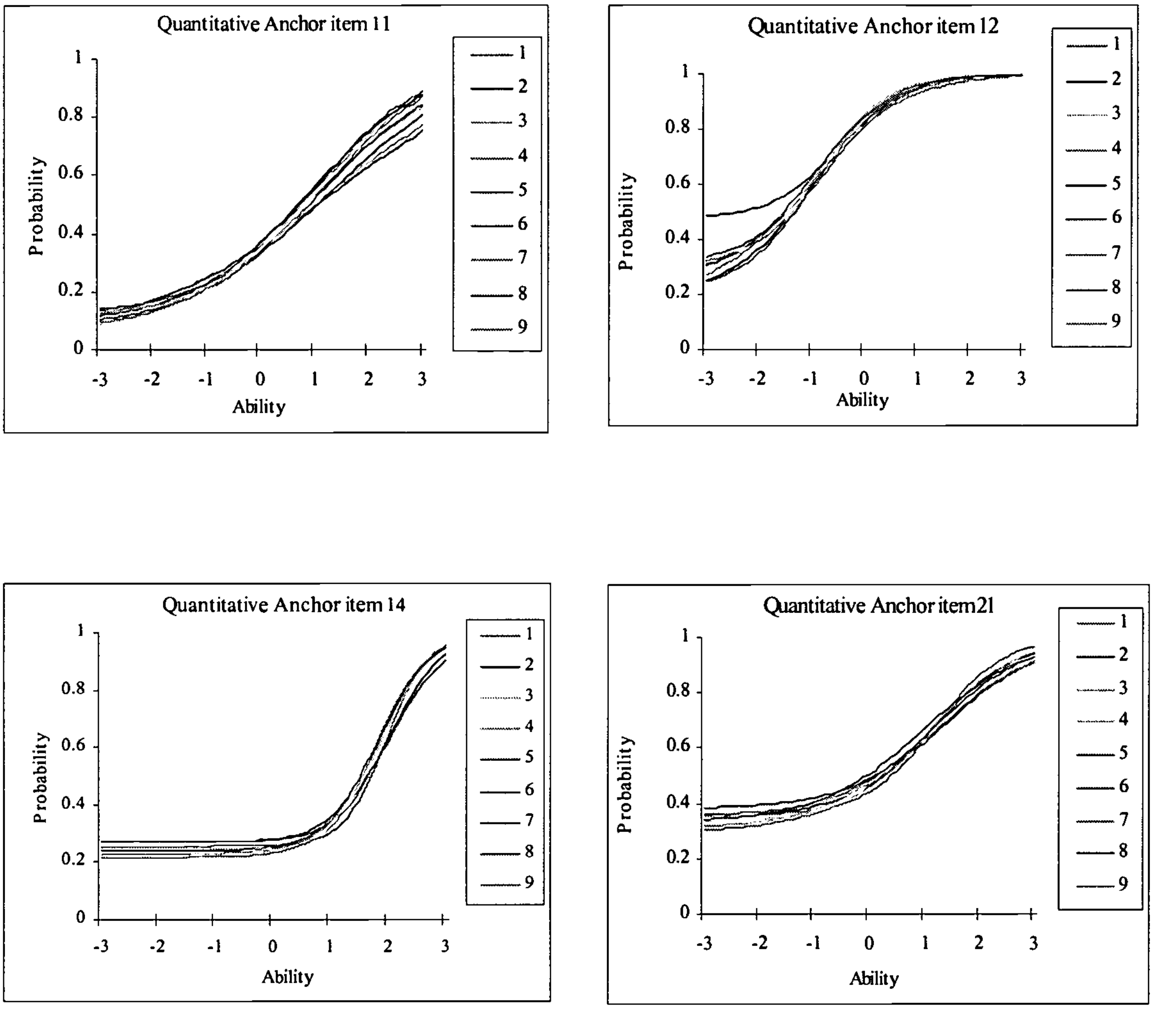


Figure A.2: Item Characteristic Curves for 4 Verbal items over 9 calibrations

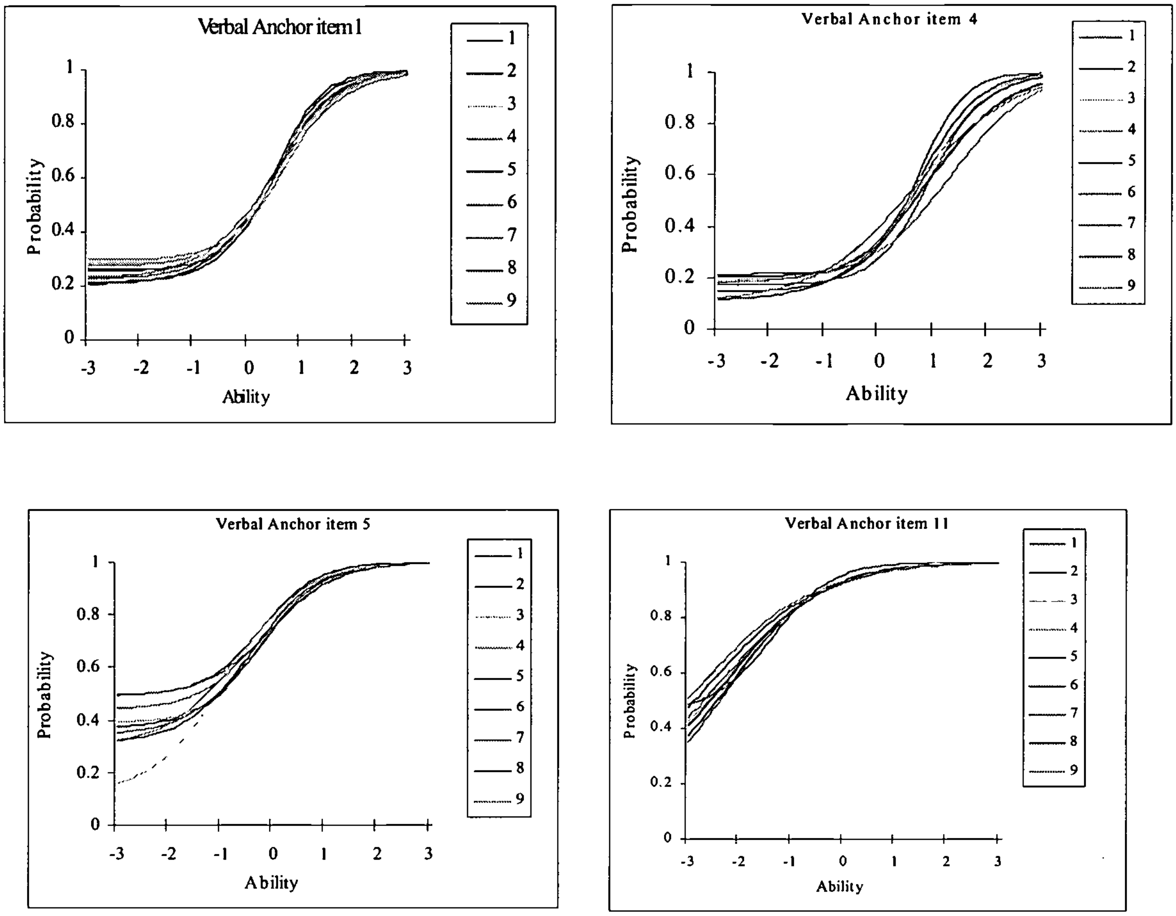
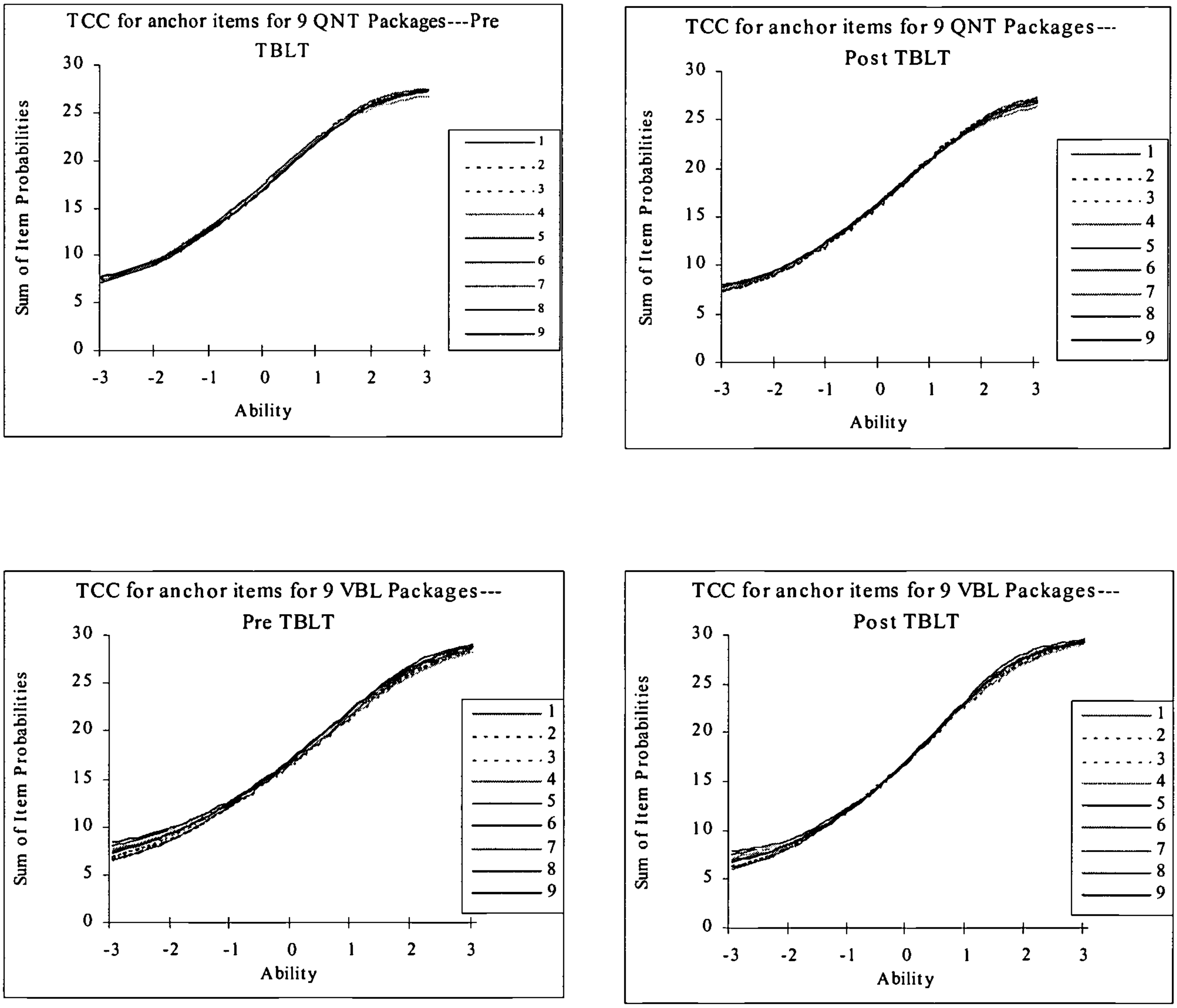


Figure A.3: Test Characteristic Curves for Quantitative and Verbal anchor items over 9 calibrations



TCC for anchor items for 9 QNT Packages---Pre TBLT

TCC for anchor items for 9 QNT Packages---Post TBLT

TCC for anchor items for 9 VBL Packages---Pre TBLT

TCC for anchor items for 9 VBL Packages---Post TBLT

Figure A.4: RMSEs between ICCs between first calibrations and the others

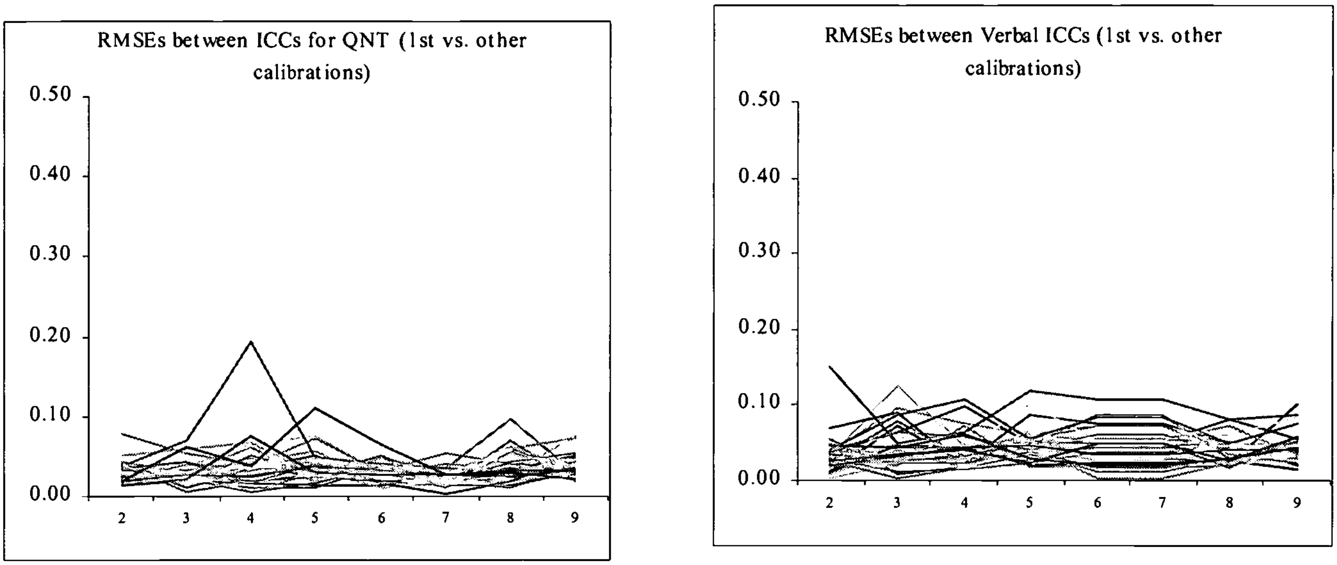


Figure A.5: Model-based vs. Empirical average variance for a- and b- parameters

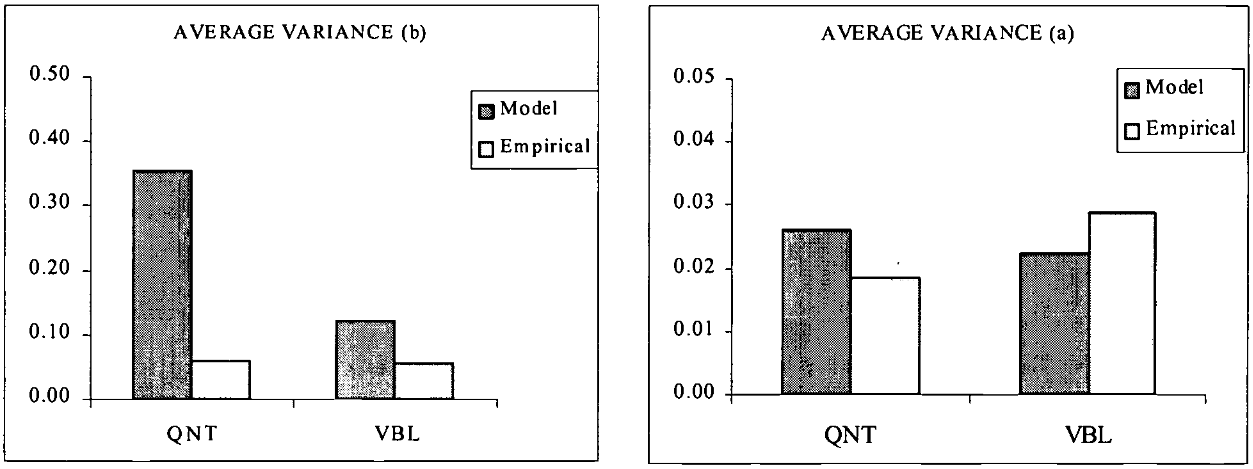
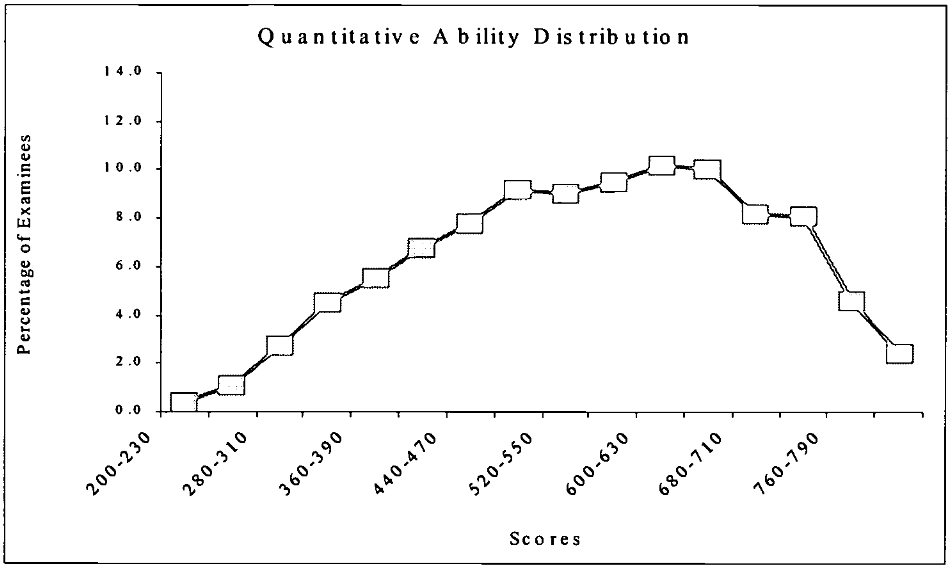


Figure A.6: Typical ability distributions for Quantitative and Verbal measures



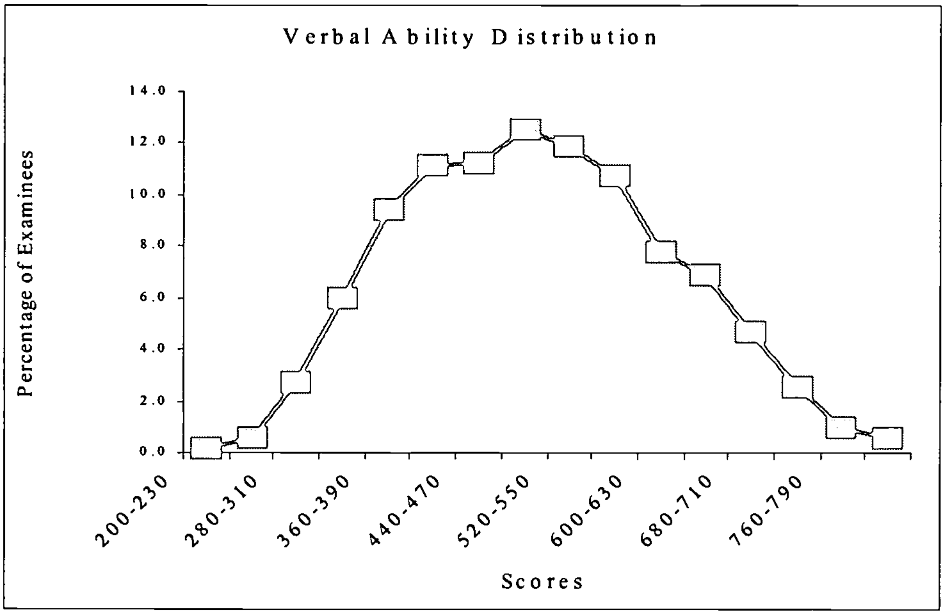


Figure A.7: RMSEs between ability estimates on a response set scored by its own and other sets of item parameter estimates---Quantitative

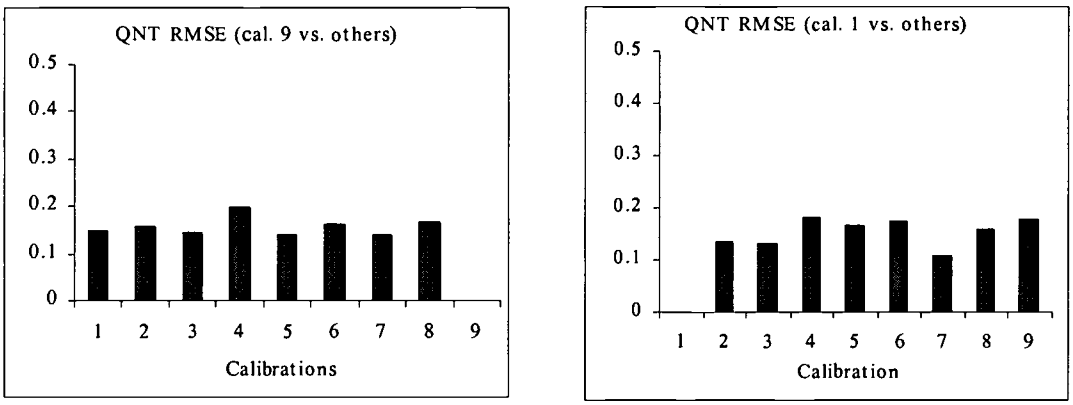


Figure A.8: Percentage of differences in the reported scores for Quantitative

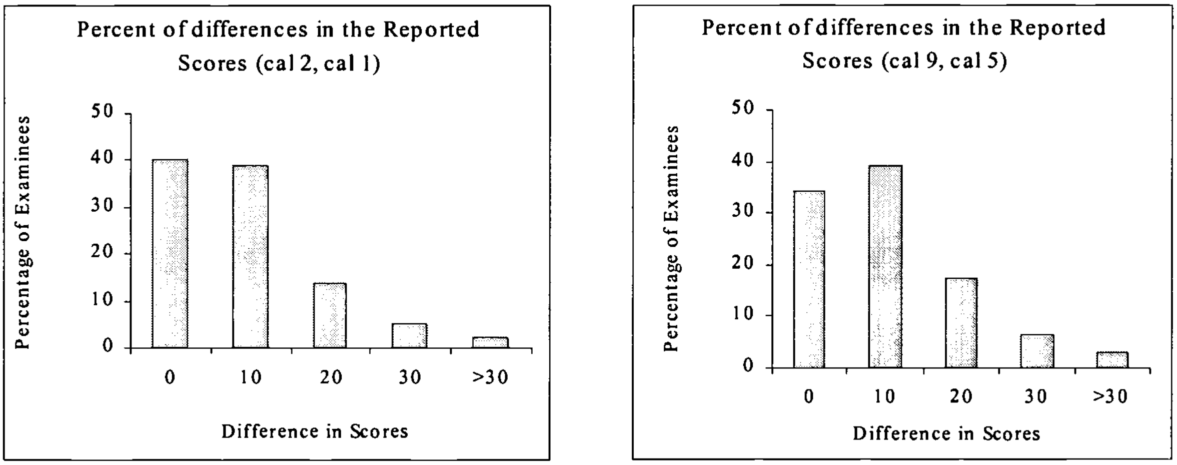


Figure A.9: RMSEs between ability estimates on a response set scored by its own and other sets of item parameter estimates---Verbal

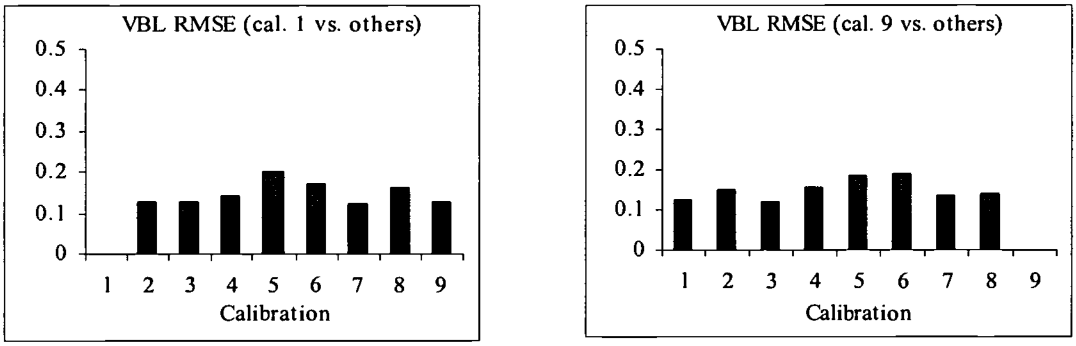
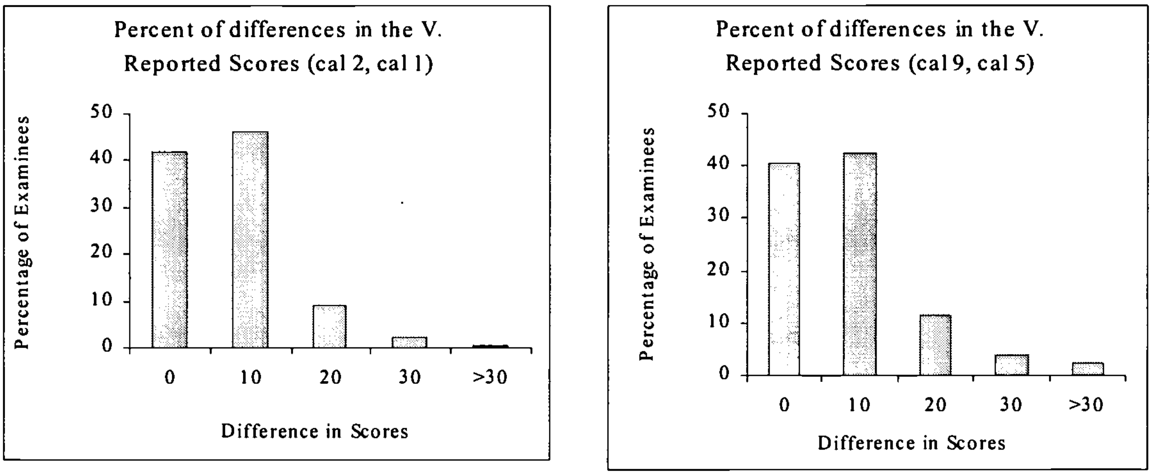


Figure A.10: Percentage of differences in the reported scores for Verbal

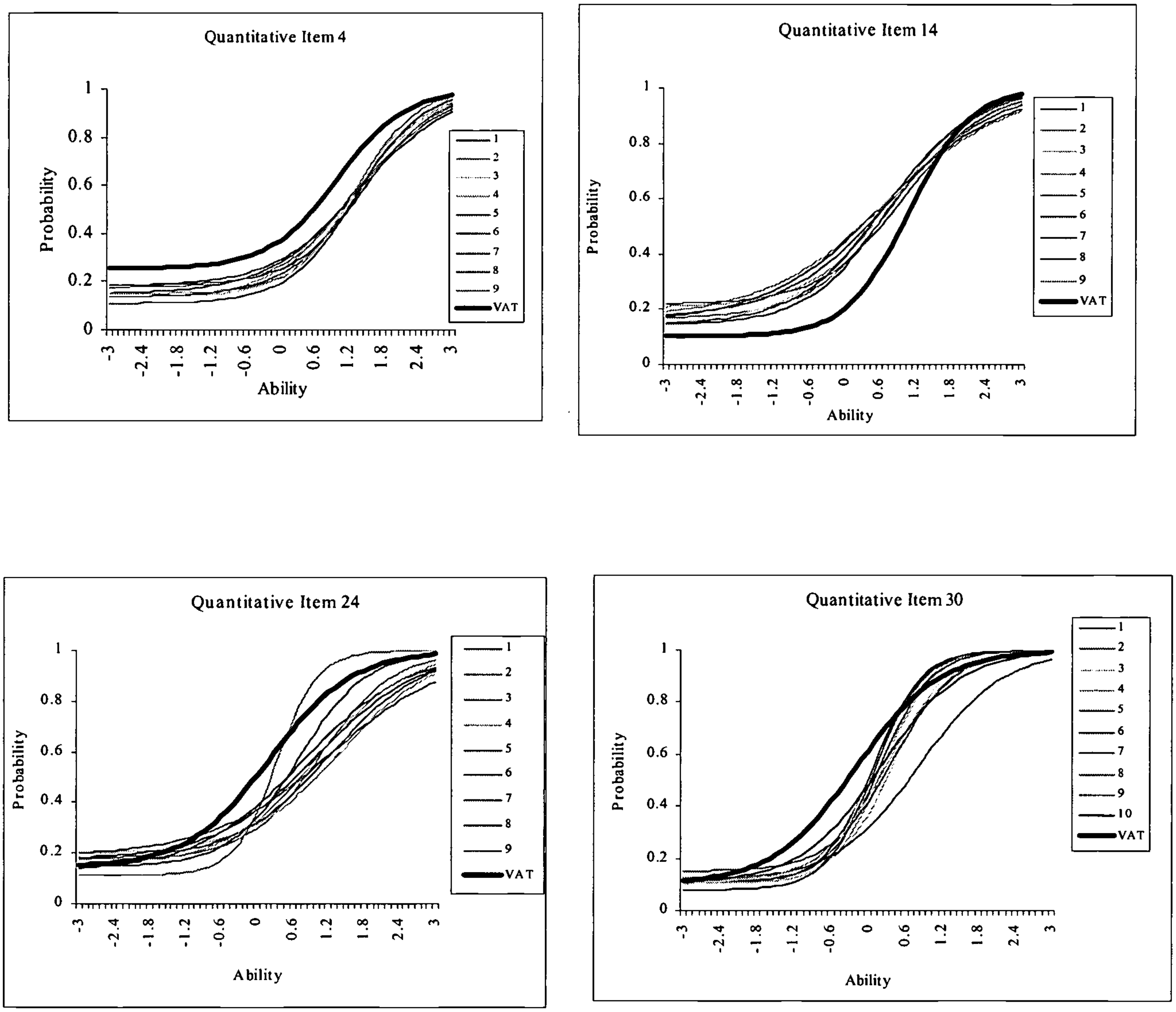


Appendix B

Table B.1: RMSEs in ICCs for CAT items on Quantitative measure

	1	2	3	4	5	6	7	8	9	10
1	0.03	0.04	0.03	0.06	0.03	0.05	0.04	0.05	0.02	0.06
2	0.02	0.01	0.02	0.05	0.03	0.05	0.04			
3	0.00	0.03	0.07	0.04	0.02	0.00	0.01	0.05		
4	0.03	0.05	0.03	0.02	0.06	0.03	0.03	0.05		
5	0.08	0.03	0.02	0.05	0.01	0.06	0.05	0.06		
6	0.02	0.02	0.02	0.04	0.01	0.02	0.02	0.04	0.04	0.06
7	0.04	0.06	0.04	0.03	0.04	0.06	0.02	0.04	0.05	0.04
8	0.04	0.04	0.03	0.04	0.02	0.03	0.02	0.02	0.03	
9	0.05	0.04	0.04	0.05	0.03	0.05	0.02	0.05	0.05	0.05
10	0.03	0.03	0.02	0.03	0.03	0.01	0.02	0.03		
11	0.06	0.05	0.06	0.08	0.07	0.06	0.06	0.05	0.05	
12	0.03	0.04	0.04	0.04	0.04	0.05	0.04	0.05	0.09	
13	0.05	0.03	0.05	0.08	0.07	0.06	0.07			
14	0.02	0.03	0.04	0.05	0.06	0.05	0.02	0.07		
15	0.02	0.05	0.04	0.04	0.05	0.04	0.05	0.04	0.04	0.02
16	0.06	0.02	0.05	0.01	0.04	0.02	0.07	0.02		
17	0.02	0.02	0.01	0.03	0.02	0.04	0.07	0.05		
18	0.02	0.01	0.03	0.02	0.03	0.00	0.04	0.03	0.02	
19	0.03	0.02	0.01	0.02	0.02	0.03	0.03	0.05	0.02	
20	0.01	0.03	0.06	0.02	0.05	0.03	0.03			
21	0.05	0.04	0.05	0.12	0.03	0.01	0.02			
22	0.03	0.02	0.05	0.03	0.03	0.01	0.04	0.02	0.04	
23	0.01	0.03	0.04	0.02	0.05	0.02	0.02	0.04	0.02	
24	0.04	0.16	0.05	0.07	0.03	0.04	0.03	0.03		
25	0.02	0.09	0.03	0.03	0.03	0.03	0.02	0.01	0.03	
26	0.04	0.04	0.01	0.03	0.02	0.02	0.06	0.06		
27	0.06	0.04	0.04	0.03	0.05	0.02	0.05			
28	0.01	0.04	0.02	0.03	0.03	0.00	0.01	0.07		
29	0.05	0.03	0.04	0.04	0.03	0.05	0.04	0.04	0.04	0.06
30	0.01	0.07	0.04	0.16	0.02	0.07	0.03	0.05	0.06	

Figure B.1: Item Characteristic Curves for 4 Quantitative CAT items



Quantitative Item 4

Quantitative Item 14

Quantitative Item 24

Quantitative Item 30

Figure B.2: RMSEs between CAT ICCs between first calibration and the others

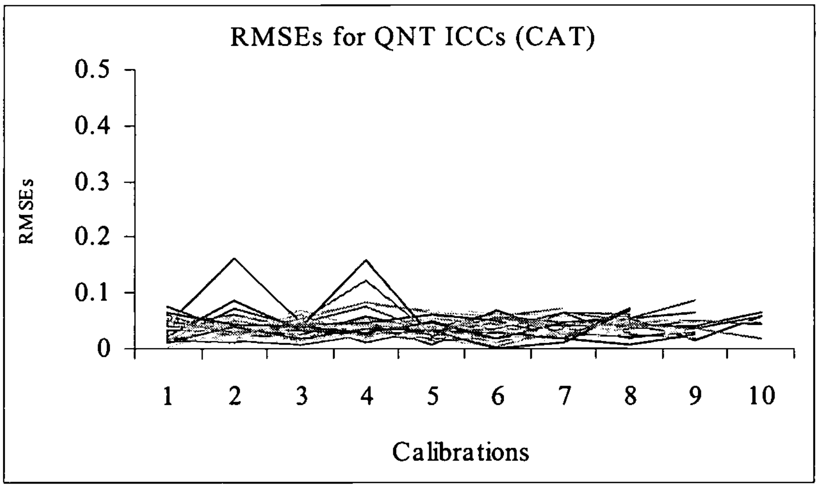


Figure B.3: RMSEs between ability estimates on a response set scored by it's own and other sets of item parameter estimates---own set=1st calibration

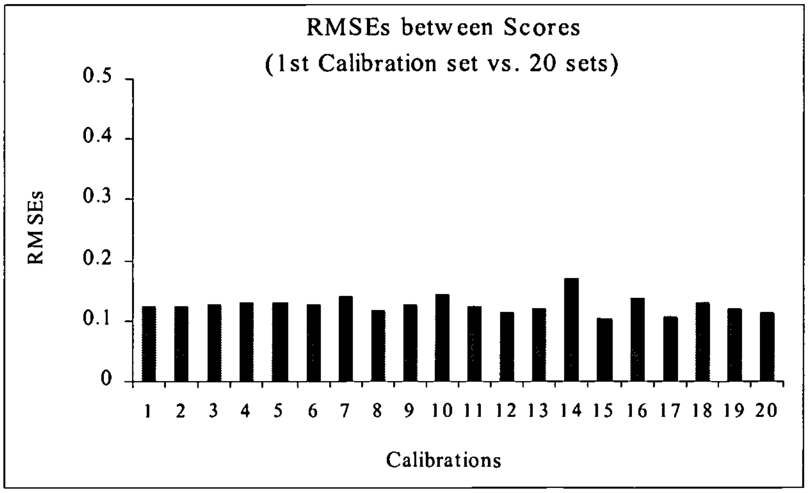


Figure B.4: RMSEs between ability estimates on a response set scored by it's own and another set of item parameter estimates by ability level---own set=1st calibration

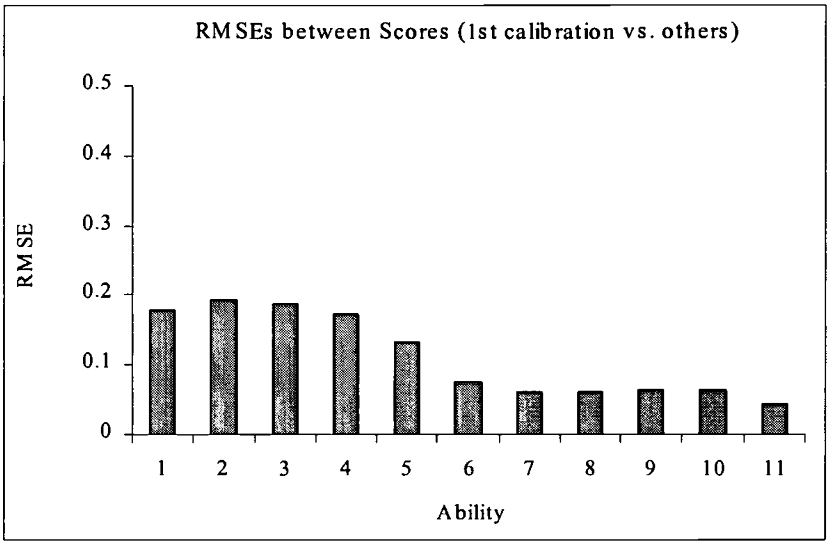


Figure B.5: Percentage of differences in the reported scores—comparison with 1st CBT calibration

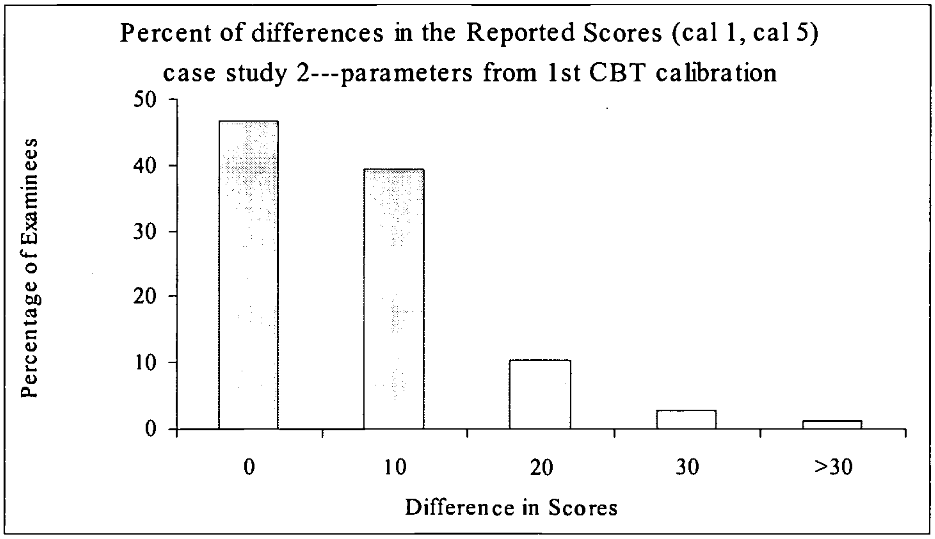


Figure B.6: RMSEs between ability estimates on a response set scored by it's own and other sets of item parameter estimates---own set=P&P bank parameters

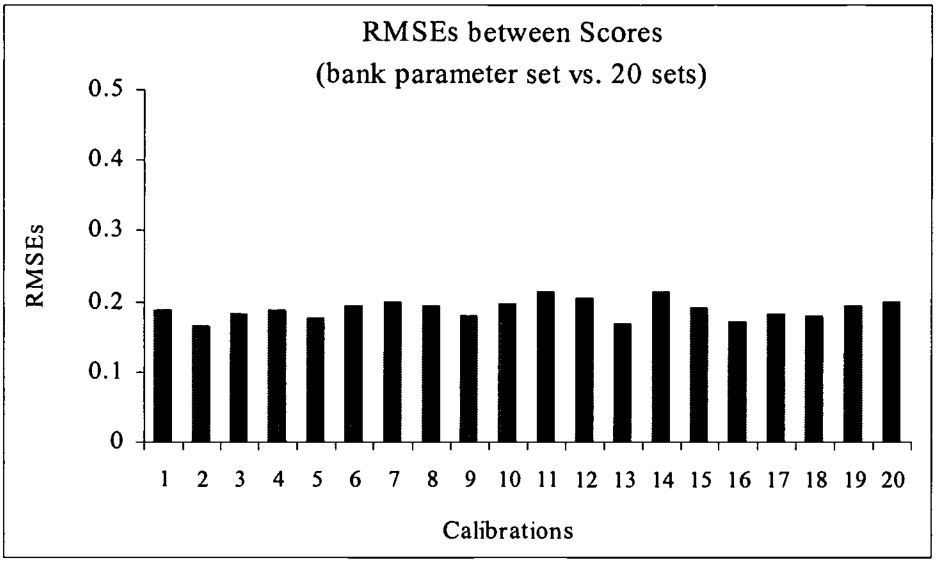


Figure B.7: RMSEs between ability estimates on a response set scored by it's own and other sets of item parameter estimates by ability level---own set=P&P bank parameters

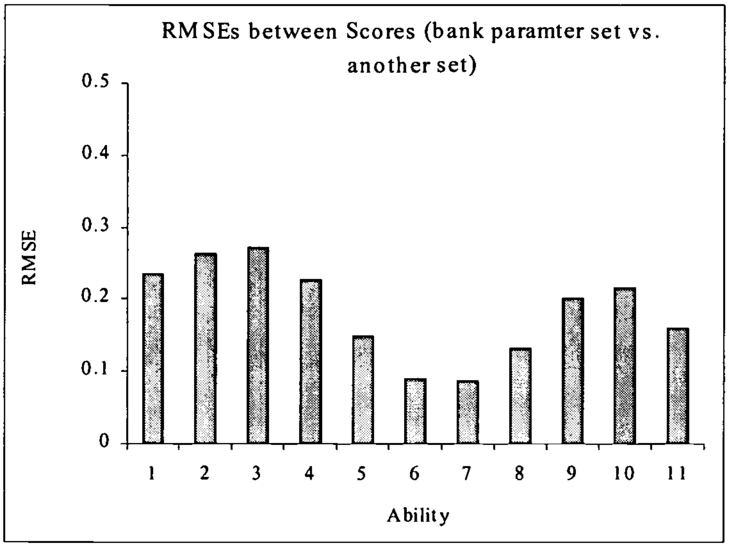
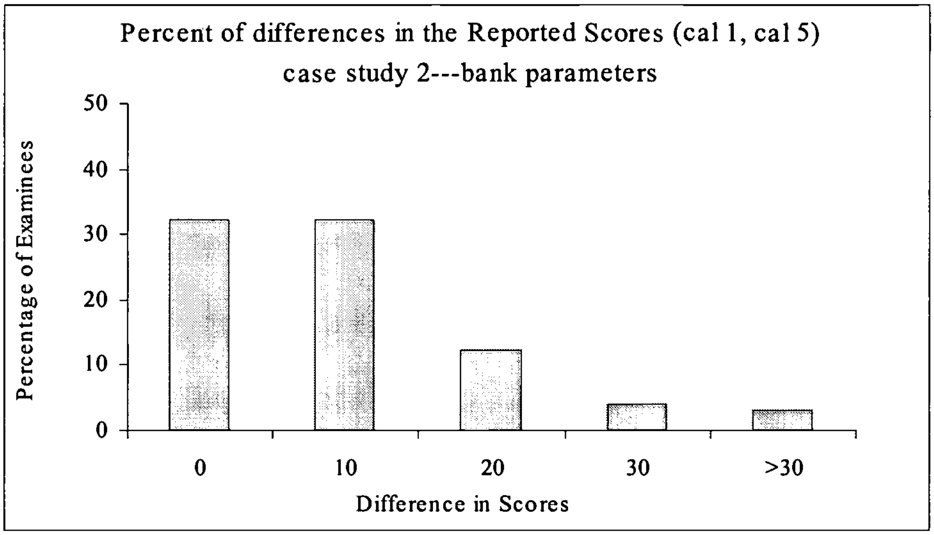


Figure B.8: Percentage of differences in the reported scores—comparison with P&P calibration



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
NCME

Reproduction Release

(Specific Document)

TM034186

I. DOCUMENT IDENTIFICATION:

Title: Tolerable Variation in Item Parameter Estimates	
Author(s): Saba Rizavi, Walter D. Way, Tim Davey and Erin Herbert	
Corporate Source:	Publication Date: APRIL, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
↑ ☑	↑ □	↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Saba Rizavi	Printed Name/Position/Title: SABA RIZAVI/MEASUREMENT STATISTICIAN/DR.	
Organization/Address: SABA RIZAVI EDUCATIONAL TESTING SERVICE, 13-L PRINCETON, NJ 08541	Telephone: 609-683-2496	Fax:
	E-mail Address: SRIZAVI@ETS.ORG.	Date: 5/20/2002.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation **1129 Shriver Laboratory (Bldg 075)** **College Park, Maryland 20742**	**Telephone: 301-405-7449** **Toll Free: 800-464-3742** **Fax: 301-405-8134** **ericae@ericae.net** **http://ericae.net**