ED 421 534 TM 028 860

AUTHOR	Parshall, Cynthia G.; Kromrey, Jeffrey D.; Chason, Walter M.
TITLE	Comparison of Alternative Models for Item Parameter Estimation with Small Samples.
PUB DATE	1996-06-00
NOTE	32p.; Paper presented at the Annual Meeting of the Psychometric Society (Banff, Alberta, Canada, June 27-30, 1996).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Estimation (Mathematics); *Item Response Theory; Models; *Reliability; *Sample Size; Simulation
IDENTIFIERS	ACT Assessment; *Item Parameters

ABSTRACT

The benefits of item response theory (IRT) will only accrue to a testing program to the extent that model assumptions are met. Obtaining accurate item parameter estimates is a critical first step. However, the sample sizes required for stable parameter estimation are often difficult to obtain in practice, particularly for the more complex models. One approach is to use modified item response models, which may be constructed so additional parameters (e.g. more than one) are included in the model, while limiting estimation. This study investigated several modified IRT models across differing sample sizes and test length in terms of their relative efficiency, accuracy, and precision. Simulated data were generated from the American College Testing program mathematics test. For some of the analyses, performance of the models tended to converge at the larger sample sizes, but at the smaller samples, the modified models displayed some important performance differences relative to the unconstrained models. The strongest pattern of results was for models that displayed the best fit within samples to display the poorest stability across samples. Conversely, models that demonstrated good stability across replications tended to be associated with relatively poorer fit within replications. (Contains 3 tables, 5 figures, and 22 references.) (Author/SLD)

Comparison of Alternative Models for Item Parameter Estimation with Small Samples

Cynthia G. Parshall
Jeffrey D. Kromrey
Walter M. Chason
University of South Florida

Abstract

The benefits of item response theory will only accrue to a testing program to the extent that model assumptions are met. Obtaining accurate item parameter estimates is a critical first step. However the sample sizes required for stable parameter estimation are often difficult to obtain in practice, particularly for the more complex models. One approach is to use *modified* item response models, which may be constructed so additional parameters (e.g. more than one) are included in the model, while limiting estimation. This study investigated several modified IRT models across differing sample sizes and test length, in terms of their relative efficiency, accuracy and precision.

Paper presented at the annual meeting of the Psychometric Society, Banff, Canada, June 27-30, 1996

Comparison of Alternative Models for Item Parameter Estimation with Small Samples

The benefits of item response theory (IRT) for testing have been discussed theoretically for a number of years (Hambleton & Swaminathan, 1985; Lord, 1980). The benefits to testing programs include applications for test development, equating, and computer adaptive testing. The most popular models practice are the unidimensional 1-parameter, 2-parameter, and 3-parameter logistic models (or, 1-PL, 2-PL, and 3-PL respectively). The formulas for these models are defined as

1-PL:

$$P(\theta) = \frac{1}{1+e^{-(\theta - b)}}$$

2-PL:

$$P(\theta) = \frac{1}{1+e^{-a(\theta - b)}}$$

3-PL:

$$P(\theta) = c + \frac{1-c}{1+e^{-a(\theta - b)}}$$

The number of item parameters which must be estimated in these models determine the examinee sample sizes required for calibrating the data. Although the recommendations for minimal sample size vary somewhat, typical guidelines are: 1000 examinees for the three parameter model, 500 for the two parameter (Hulin, Lissak, & Drasgow, 1982) and 200 for the one parameter model (Wright & Stone, 1979). The advantages of IRT methods will only accrue to the extent that the assumptions of the model used are met and model-data fit is found. Among other possible problems, tests which are constructed based upon imprecise item parameters may result in an overestimate of the test information and in ability estimates which are less accurate than they appear to be (Hambleton & Jones, 1994). One source of poor parameter estimates is the use of an inadequate sample size for calibration, which can result in excessively large standard errors of the item parameter estimates (Hambleton & Jones, 1994; de Jong & Stoyanova, 1994).

Many testing programs have an interest in IRT methods for test development, item analysis, and adaptive testing. However, the sample sizes required for stable parameter estimation are often difficult to obtain in practice, particularly for the more complex models. The large sample sizes may be difficult to obtain if testing programs have small numbers of examinees per administration, sub-group analyses draw from small numbers of examinees, multiple sub-content areas are assessed separately, and/or test forms are replaced frequently.

Sample size constraints might lead testing practitioners to select the model with the least stringent requirement (e.g., one parameter). In practice, however, many testing programs consist of sets of multiple choice items which vary in discrimination and allow for guessing. This would suggest that a more general model, such as the three parameter model, might provide the best fit to typical data and that use of a more limited model would lead instead to model misspecification errors (Divgi, 1986).

Spray, Kalohn, Schulz, and Fleer (1995) conducted a simulation to investigate the effect on adaptive classification testing when the true model was the 3-PL model, but the items were calibrated according to the 1-PL model. These researchers found use of the 1-PL model under studied conditions to result in unacceptable rates of both false positive and false negative decisions (i.e., examinees classified as either passing and failing, who would have been classified otherwise according to the true 3-PL model). Yen (1981) has also pointed out problems which may arise when a 1-parameter or 2-parameter model is used inappropriately, or when truth is best modeled by a 3-parameter model. These problems include the potential for sample dependency of some item parameters, inaccurate model predictions, and attentuated correlations between actual and estimated trait values.

Modified Models

Given limitations on available examinee sample sizes, a practical concern is to obtain the most accurate item parameter estimates possible. A promising avenue of research concerns *modified* item response models (Barnes & Wise, 1991; Harwell & Janosky, 1991; Sireci, 1992; and Stone & Lane, 1991). Modifications to models may be constructed so that additional parameters are included in the model, while estimation is limited by fixing that parameter value. Other modifications may be designed such that parameter estimation is limited by allowing parameters only a limited range within which they may vary.

Sireci (1992) investigated modifications to 1-PL and 2-PL models on multiple small sample datasets, obtained over several test administrations. Part of this study was an investigation of a modified model which included a fixed *c* parameter. One analysis considered restricted conditions, in which item parameters were constrained to be equal across the multiple samples of examinees. Another analysis addressed the use of mixed models (e.g., more than one IRT model for a specific analysis). Modified IRT models were also used by Stone and Lane (1991). In this study, an unconstrained 2-parameter IRT model was compared to a model in which item parameters were constrained to be equal across pretest-posttest administrations. This modification enabled an investigation into the stability of the item parameter estimates over time. Additional alternative IRT models have also been utilized in the context of differential item functioning (DIF) analysis (Thissen, Steinberg, & Wainer, 1993).

While some of the studies of modified item response models have been conducted on real data (Sireci, 1992; Stone & Lane, 1991), others have been simulations (Barnes & Wise, 1991; Harwell & Janosky, 1991; Patsula & Pashley, 1996). Simulation studies have the advantage of utilizing true parameter values, which are never known in practice.

For example, Barnes and Wise (1991) conducted a simulation in which the item parameter estimates obtained under small sample conditions for typical 1-parameter and 3-parameter models were compared to two modified models. The modifications in this study involved the inclusion of a fixed, non-zero *c* parameter. These fixed *c* parameter models were based on the number of response options in the multiple choice items, *A*.

One modification fixed *c* at 1/*A*, and a second modification fixed *c* at 1/*A* - .05. Because the value of the *c* parameter was fixed, the sample size requirements for a standard one parameter model remained appropriate under the modifications. The results indicated that the modified models outperformed the more traditional 1-parameter and 3-parameter models.

Harwell and Janosky (1991) also investigated item parameter estimation with small samples. This simulation study examined several 2-parameter models in which estimation of the *a* parameter was affected by imposing different variances on the prior distribution of the *a*'s. Under the conditions in this study, item parameter estimates for small samples were recovered more accurately when a more informative prior variance was used.

An alternative approach investigated by Patsula and Pashley (1996) used polynomial logistic regression to model ICCs in pretest items (i.e., when ability estimates can be reliably computed based on operational items). This procedure included a mixed model component in that it provided a means of identifying subsets of items which could be adequately modeled with fewer parameters (i.e., 2-PL or 1-PL). Where a reduced number of parameters needs to be estimated, presumably more stable results can be obtained under smaller sample conditions.

These results suggest that modifications to popular IRT models are worthy of further investigation, and that appropriate modifications may provide more stable estimation of parameters with fewer examinees than unmodified models. This study was intended to build upon the previous research into modified item response models, under moderate and small sample size conditions. Additionally, this study included a greater number of replications than are often found in parameter estimation studies, providing for a more stable analysis of results (Robey & Barcikowski, 1992; Stone, 1992).

Purpose

Obtaining accurate parameter estimation is a critical concern, since all of the applications of IRT are based on these parameters. However practical testing applications frequently include elements which might best be modeled by more complex models (e.g., the 3-parameter model) while having only small samples of examinees to draw upon for calibration data. Determining a means for parameter estimation under these conditions is thus an important area of research.

This study investigated the relative small sample efficiency of several modifications to existing IRT models. The purpose of this study was, in part, to find a lower limit in terms of sample size needed to adequately recover item parameters. More generally, the purpose was to investigate the sample size at which sampling error causes more problems than model misspecification (and vice versa).

Methods

This study used simulated data, based on item parameters generated from data from a previous administration of the 40-item ACT Assessment Mathematics test. The 3-PL model was used to obtain empirical item parameters using the archival item responses from 2000 examinees. These empirical item parameters were then used to generate simulated data. This set of item parameters has been used previously (Parshall & Miller, 1995; Spray, 1989), and is intended to provide results more generalizable to practice than simulated data are often able to do.

Examinees' (simulees') true thetas were generated from a normal ability distribution (i.e., $N(0,1)$). These ability and item parameters were regarded as true parameters for purposes of the study. Item response vectors were then generated by determining the probability of a correct response for a given theta parameter, and then comparing that probability to a random number sampled from a uniform (0,1) distribution. If the random number was less than or equal to the probability of a correct response, then the response was scored as correct.

Item parameter estimates were obtained through the calibration program BILOG (Mislevy & Bock, 1990). Models under investigation included the typical 1-parameter, 2-parameter, and 3-parameter models as benchmarks. The additional, modified models consisted of a 2-parameter model with a restricted a (i.e., a strong prior distribution was imposed), a 3-parameter model with a restricted a parameter, and a 3-parameter model with both a restricted *a* parameter and a common *c* parameter. This yielded a total of six models, three of which were unrestricted, and three of which were restricted.

The benchmark 1-PL model constrained all *a* parameters to be equal; both the 1-PL and the 2-PL models set the *c* parameters to zero (i.e., did not estimate *c* parameters). The benchmark 2-PL and 3-PL models used BILOG's default prior distribution for *a* parameters, which is $.5^2$ in the lognormal metric (or, μ_a=1.13 and σ_a=.36 in the *a* metric). This default prior is typically imposed to avoid the extreme values sometimes estimated for *a* parameters (i.e., to prevent Heywood cases). For the benchmark 3-PL model, the default beta prior was also used for estimation of the *c* parameter. All three modified models imposed more informative priors on the *a* parameters. These modified 2-PL and 3-PL models included a prior of $.25^2$ in the lognormal metric (or, μ_a=1.03 and σ_a=.07 in the *a* metric). One modified 3-PL model also constrained the *c* parameters to be equal (but free to be non-zero). These modified models are noted as 2-PL*a*, 3-PL*a*, and 3-PL*ac*.

Each of these models was investigated with sample sizes of 1000, 500, 250, and 100. The largest sample size here is typically considered adequate for the 3-parameter model, while the smallest sample size might prove challenging for even the 1-parameter model. The full study was a 6 x 4 design, with the six models and four sample sizes yielding a total of 24 conditions. In order to control for sampling error, 100 samples of each condition were generated, and the results were analyzed across replications.

After the initial analysis of the simulated data, some samples failed to converge using BILOG's default number of EM cycles and Newton-Gauss iterations (10 cycles, followed by 2 Newton steps). This was particularly true of sample size 100, and the 3-PL and 3-PL*ac* models (see Table 1). A second phase of analysis was performed on any non-converging data files, with modified BILOG command files in which the number of iterations was increased to the arbitarily large values of 50 EM cycles and 10 Newton-Gauss iterations.

Insert Table 1 about here

Increasing the number of iterations greated increased the number of samples that converged (typically, in far fewer than the maximum number of iterations), but did not eliminate nonconvergence. A subset of nonconverging data files were examined in detail. Typically one or two items had excessively large values in one of the parameters. If removed from the analysis, the remaining items would converge successfully. Based on the design of this study and our desire to collapse results across replications, rather than deleting items we chose to resample. For those samples that did not converge after the second data analysis, new simulated data were generated from the original parameters and were used to replace the nonconverging samples in the raw data files. The samples with new data were then analyzed and all were found to converge.

Results

A variety of evaluative measures are conducted in analyses of item parameter recovery. For studies such as the present one, in which data are generated from one model, but may be estimated according to another, certain comparisons are inappropriate. For example, if the data have been generated from a 3-PL model, then those parameter estimates obtained from 2-PL and 1-PL models cannot be directly referenced back to the generating parameters (e.g., analyses such as root mean squared error), since they are not on the same scale. The relative success of the six IRT models in this study was therefore determined using indices of model-data fit and indices of the stability of the models across samples.

Fit Indices

Two indices of fit were calculated for each item in each sample. These fit indices represent the extent to which each model was able to predict the generating data in the sample. First, raw residuals from ability groups (Hambleton & Swaminathan, 1985) were calculated for each sample and each model. In this method, the range of estimates of theta in the sample is divided into ten equal intervals. Within each interval, the squared difference between the actual proportion of examinees who answered the item correctly and the expected proportion based on the IRT model of the item is calculated. The sum of these squared residuals, across the ten intervals, is calculated as the index of fit for the item in the sample, and the mean of these fit indices across the 100 samples was used:

$$r_i = \frac{\left\{ \sum_k \left[\sum_j (P_{ijk} - E_{ijk})^2 \right] \right\}}{100}$$

where

r_i = raw residual for item i

P_{ijk} = observed proportion of correct responses for item i, interval j, and sample k,

E_{ijk} = predicted proportion of correct responses for item i, interval j, and sample k.

For the second index of fit, individual person residuals were calculated. These are the residuals between the observed item data and the obtained probabilities (X_{ij} - P_{ij}) calculated for each item and each examinee. The average of these residuals across examinees is used as the fit index for the item:

$$XP_i = \frac{\left\{ \sum_k \left[\frac{\sum_j (X_{ijk} - P_{ijk})}{40} \right] \right\}}{100}$$

where

XP_i = mean person residual for item i

X_{ijk} = observed response for item i, examinee j, and sample k,

P_{ijk} = estimated probability of correct responses for item i, examinee j, and sample k.

Stability Indices

Estimates of the stability of the item parameter estimates and the item response functions were obtained by calculating the standard deviations of the estimates of the *a* and *b* parameters, and the standard deviations of the entire item response curve over the 100 samples. The standard deviations of the item parameter estimates were obtained using the usual formula for the sample estimate of a population standard deviation:

$$\sigma_{ij} = \sqrt{\left[\frac{\sum_{k}\left(X_{ijk} - \mu_{ij}\right)^{2}}{99}\right]}$$

where

σ_{ij} = standard deviation of parameter i, for item j,

X_{ijk} = estimate of parameter i, for item j, in sample k, and

μ_{ij} = mean of parameter i, for item j in the 100 samples.

The standard deviation of the entire item response curve was obtained by dividing the theta scale into 31 equally spaced intervals (spanning a theta range from -3.0 to 3.0) and calculating the expected proportion of correct responses within each interval (P_{mn}, for interval m and item n), given the item parameter estimates obtained from the sample data. The standard deviation was then obtained as

$$\sigma_{n} = \frac{\left\{\sum_{m}\sqrt{\frac{\sum_{o}\left(P_{mno} - \mu_{mn}\right)^{2}}{99}}\right\}}{31}$$

where

σ_{n} = standard deviation of item response curve for item n,

P_{mno} = estimate of proportion of correct responses for interval m, item n, and sample o,

μ_{mn} = mean of estimates for interval m, item n, in the 100 samples.

Model-Data Fit Across Samples

The fit indices obtained from each model are presented in Table 2. Reported in the table are average fit indices across the 40 items. The standard deviations in the table are the average standard deviation in item fit across the 100 samples. These fit indices are graphed in Figures 1 and 2.

Insert Table 2 and Figures 1 and 2 about here

An examination of the raw residuals (Figure 1), shows that the unmodified 3-PL and the *3-PLa* models evidenced the smallest residuals for the six models examined. The differences in fit between the unmodified 3-PL and the modified, 3-PL*a* were negligible with samples as small as 250 (in which the residuals were 0.1177 for the 3-PL*a*, and 0.1188 for the unconstrained 3-PL model). However, with samples of size 100, the constrained model showed a better fit to the data (0.2168 for the 3-PL*a* model and 0.2489 for the 3-PL). The only exception to the superior fit of these two models is with samples of size 100, in which the residuals from the 2-PL model (0.2430) were slightly smaller than those of the unconstrained 3-PL (0.2489). However, in this small sample condition, the 3-PL*a* model was notably better fitting than any of the other five models.

For the individual person residuals (Figure 2), both the unmodified 3-parameter model and the 3-parameter model with constrained *a* (3-PL*a*) provided the best fit to the data across all of the sample sizes examined. With samples of size 100, the constrained *a* model was substantially superior to the unmodified model (with residuals of 0.0107 and 0.0140, respectively). As sample size increased, the difference in fit of these two models became negligible (with residuals of 0.0180 and 0.0183, for the constrained and unconstrained models, respectively, with samples of size 1000), but the fit of both of these models remained substantially better than that of the other four models examined. Least well fitting to the data were the 1-PL and 2-PL models (including both constrained and unconstrained versions of the latter). The 3-PL model with both constrained *a* and fixed *c* (3-PL*ac*) did not fit the data as well as the less constrained versions of the 3-PL, but the fit of this model was better than that of the 1-PL or 2-PL models.

The variation in the fit of the models across the samples is reported in Table 2. The standard deviations reported in this table are the average standard deviations of the fit indices in the 100 samples. Small values of this statistic reflect consistency in fit across the samples, while large values reflect greater amounts of variation in fit with different samples of examinees. As seen in Table 2, the standard deviations for both indices of model-data fit (i.e., (X_{ij} - P_{ij}) and model residuals) were smallest for the 3-PL and 3-PL*a* models, and were largest for the 1-PL and 2-PL models. The smaller variability for the 3-PL and 3-PL*a* models was consistent across the sample sizes examined in this study. Thus, these data suggest that the 3-PL and 3-PL*a* models not only provide better average fit than the other models, but the fit is more consistent across samples.

Stability Across Samples

The second general method for evaluating the success of the six models was the stability of item parameter estimates across samples. Estimates of such stability were obtained by calculating the standard deviations of the estimates for the *a* and *b* parameters for each item, then averaging these standard deviations across the 40 items on the test. In addition, an overall measure of the stability of the item curves was obtained by calculating the standard deviation of P_{ij} at each of 31 theta values. These stability estimates are presented in Table 3 and are graphed in Figures 3 through 5.

Insert Table 3 and Figures 3 - 5 about here

In an examination of the stability of the estimates of the *b* parameter (Figure 3), all of the estimates became more stable as sample size increased, and the stabilities across models became more similar with large sample sizes. For example, with samples of size 100, the most stable estimates of *b* were obtained with the 1-PL model (standard deviation = 0.2543), while the least stable estimates were obtained with the 3-PL*a* model (standard deviation = 0.3627), giving a range in stability of 0.1084. In contrast, with samples of size 1000, the most stable estimate (obtained with the 1-PL model, standard deviation = 0.0762) was only 0.0467 lower than the least stable estimate (obtained with the 3-PL model, standard deviation = 0.1229). However, the most stable estimates of *b*, across the four sample sizes examined in this study were obtained with the 1-PL and the 2-PL*a* models. Conversely, the least stable estimates were obtained with the 3-PL and 3-PL*a* models.

The stability of the estimates of the *a* parameter showed a different pattern from that obtained with the *b* parameter (Figure 4). Less convergence was evident as sample size increased, and the difference in stabilities between small and large sample sizes was less striking. As should be expected, the most stable estimates of the *a* parameter were obtained with those models that provided a constraint on *a* (i.e., the 1-PL, 2-PL*a*, 3-PL*a*, and 3-PL*ac*). The models that do not constrain the *a* parameter (2-PL and 3-PL) evidenced substantially more variability across samples. As sample size increased, the variability in the *a* parameter estimates obtained with the 2-PL model was similar to that obtained with the models that impose a constraint on *a*, but the variability obtained with the 3-PL model remained notably larger even with samples of size 1000.

Finally, the estimates of the stability of the item curves (Figure 5) are consistent with the previous stability indices. The least stable curves were obtained with the 3-PL and 3-PL*a* models, while the most stable curves were obtained with the 1-PL model. With samples of size 100, the standard deviation of the curves with the 1-PL was only 0.0421, while the standard deviations for the 3-PL and 3-PL*a* were 0.0564 and 0.0547, respectively. As with the stability of the *b* parameter estimates, the stability of the overall item response curve increased substantially with increasing sample size, however, convergence of stability across models was not evident in these data. With the largest sample sizes examined (n = 1000), the standard deviation of the curves for the 1-PL model was 0.0124, while those for the 3-PL and 3-PL*a* models were 0.0239 and 0.0219, respectively.

Discussion

This study was designed to investigate the relative effects of sample size and model misspecification on item parameter estimation, and whether various modifications to typical models might improve estimation under these conditions. While for some of the analyses, performance of the models tended to converge at the larger sample sizes, at the smaller samples the modified models displayed some important performance differences relative to the unconstrained models.

The strongest pattern of results in this study was the tendency for models which displayed the best fit within samples, to display the poorest stability across samples. Conversely, models which demonstrated good stability across replications tended to be associated with relatively poorer fit within replications. For conditions examined in this research, it appears that a trade-off between these two criteria may need to be made.

One anticipated effect of sample size was the general improvement in fit for all models as sample size increased. Stability also improved with increasing sample size. For the *b* parameter, this increased stability was associated with a tendency for the performance of the various models to become more similar to one another. For the overall item response curves an improvement in stability as sample size increased was still present, although somewhat less marked, but convergence across models was not evident. A more complex result was found for the *a* parameter, due to constraints some of the models placed on this parameter. Little convergence across models appeared, and few models demonstrated improvement in stability with increased sample sizes. Interestingly, the models with constrained *a* parameters performed about as well (or poorly) at the smallest sample sizes as they did at the largest. Only minimal changes in performance across sample size can be noted for these constained models, with the 2-PL*a* and 3-PL*ac* showing slight improvement, while the 3-PL*a* worsened slightly.

Truth for this research was defined as the 3-PL model, and the data were generated according to this model. This would suggest that models which incorporated 3 parameters (the 3, 3-PL*a,* and 3-PL*ac* models) should demonstrate better fit than those models with fewer parameters. In general, this expectation is realized, along with a tendency for these models to display poorer stability than the 2- and 1-parameter models. In comparison to the unconstrained 3-PL model, the 3-PL*a* generally displayed better fit, with results converging at N=1000. At N=100, a sample size far below recommendations for use of a 3-parameter model, the 3-PL*a* model yielded the best fit to the observed data. The stability of estimation of the *a* parameter and of the overall ICC were also improved under the 3-PL*a* model; however, the stability of the *b* parameter was slightly worse for the 3-PL*a* as compared to the unconstrained 3 parameter model when N<500. The 3-PL*ac* showed relatively poorer fit and somewhat better stability in comparison to the

3-PL and 3-PL*a* models, while displaying better fit and poorer stability than the remaining 2- and 1-parameter models.

The 2-parameter model with a constrained *a* parameter (2-PL*a*) displayed better fit than the unconstrained 2-PL model under the individual person residual analysis, but not under the analysis of raw residuals obtained from ability groups, with results converging at N=1000 for both analyses. The 2-PL*a* demonstrated better stability than the 2-parameter model for the *a* and *b* parameters, as well as for the ICC as a whole. This improved stability can be noted across all sample sizes, but is especially evident at sample sizes less than 500.

For this simulation, imposing a more informative prior on the variance of the *a* parameter seemed to improve both fit and stability in comparison to the unconstrained models with the same number of parameters, especially at the smaller sample sizes. This effect can be noted for both the 2-PL*a* and 3-PL*a* models (in comparison to the 2-PL and 3-PL models), but is most marked for the 3-PL*a*. These results are in line with those noted by Harwell and Janosky (1991), who investigated the effect of differing prior variances on the *a* parameter in a 2-PL model, and found more informative priors to improve parameter recovery with small samples and short tests.

When an additional constraint is placed on the *c* parameter, as well as on the *a* parameter (3-PL*ac*), fit was worsened markedly, while stability was somewhat improved. In fact, the 2-PL*a* and the 3-PL*ac* may be viewed as "compromise" models for this set of results. These two models, which differ only by the inclusion of a common, non-zero *c* for the 3-PL*ac*, tended to produce results in the center of the set of 6 models investigated in this study.

Other research on model modifications related to the *c* parameter used a fixed value for *c* (Barnes & Wise, 1991; Sireci, 1992), rather than the common *c* investigated in this study. Barnes and Wise noted improved results for modified models including a fixed *c*, over both the 3- and 1-parameter models under small sample conditions and multiple-choice data. The fixed *c* method could have yielded improvements over the common *c*, since use of that method reduces the number of parameters which need to be estimated.

One limitation of this study was the use of only a single set of parameters for generating data. Another, was the fact that the data were generated according to a single model. This may have implications for the performance of modifications such as the 3-PL*a* and the 3-PL*ac* relative to other models under investigation.

What the general performance of any of these modified models may be under a variety of condition remains to be investigated. For constraints on both the *c* and the *a* parameter, the relative truth of those constraints may hold strong implications for the effectiveness of the models which make use of them. For example, some datasets may be inappropriately modelled by a more restrictive prior on the distribution of the *a* parameter. Additionally, the range of *c* values in a given dataset may be poorly represented by a common value. An informative prior, or a fixed value, which are far from truth could lead to worse results than those obtained without constraints.

Future research on modified models should include the use of additional data sets and generating parameters. Tests of differing lengths and test characteristic curves should be utilized, along with differing examinee ability distributions and sample sizes. Alternative model modifications should also be considered, including a fixed *c* and additional prior distributions for the *a* parameter. Finally, the accuracy and stability of the item parameter estimates obtained on one sample could be evaluated through use of a second sample and a cross-validation approach.

References

Barnes, L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4*, 143-157.

de Jong, J. H. A. L., & Stoyanova, F. (1994, March). *Theory building: Sample size and data-model fit*. Paper presented at the annual Language Testing Research Colloquium.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*, 171-186.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*, 143-155.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*, 249-260.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mislevy, R. J., & Bock, R. D. (1990). *Bilog 3: Item analysis and test scoring with binary logistic models, 2nd Edition* [Computer program]. Chicago: Scientific Software.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement*, *32*, 302-316.

Patsula, L. N., & Pashley, P. J. (1996, April). *Pretest item analyses using polynomial logistic regression: An approach to small sample calibration problesm associated with computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Sireci, S. G. (1992, August). *The utility of IRT in small-sample testing applications*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Spray, J., Kalohn, J. C., Schulz, M., & Fleer, P., Jr. (1995, June). *The effect of model misspecification on classification decisions made using a computerized test: 3-PL versus 1-PL logistic item response models*. Paper presented at the annual meeting of the Psychometric Society, Minneapolis.

Spray, J. A. (1989). *Performance of three conditions DIF statistics in detecting differential item functioning on simulated tests* (Research Rep. No. 89-7). Iowa City, IA: American College Testing.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16,* 1-16.

Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4*, 125-141.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item repsonse models. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52, 275-291.

Table 1

Number of Nonconverging Samples in the First Analysis/Second Analysis

	Sample Size			
Model	100	250	500	1000
1-PL	1/1	0	0	0
2-PL	27/6	8/0	3/0	1/0
2-PL*a*	3/0	2/0	3/0	0
3-PL	42/14	31/4	17/1	6/0
3-PL*a*	7/0	2/0	1/0	0
3-PL*ac*	92/0	100/0	99/0	100/0

Note: 0 in a cell indicates all samples converged for a condition

Table 2

Indices of Model-Data Fit for Six Models and Four Sample Sizes.

		Fit Index			
		X-P		Residual	
Sample Size	Model	Mean	SD	Mean	SD
100	1-PL	0.0276	0.4259	0.2629	0.2298
	2-PL	0.0271	0.4201	0.2430	0.2130
	2-PL*a*	0.0257	0.4231	0.2553	0.2237
	3-PL	0.0140	0.4178	0.2489	0.2078
	3-PL*a*	0.0107	0.4178	0.2168	0.1540
	3-PL*ac*	0.0215	0.4210	0.2585	0.2310
250	1-PL	0.0276	0.4264	0.1650	0.1269
	2-PL	0.0288	0.4213	0.1435	0.1124
	2-PL*a*	0.0276	0.4235	0.1554	0.1466
	3-PL	0.0162	0.4183	0.1188	0.0938
	3-PL*a*	0.0146	0.4209	0.1177	0.0945
	3-PL*ac*	0.0240	0.4232	0.1774	0.2121
500	1-PL	0.0281	0.4258	0.1265	0.0856
	2-PL	0.0295	0.4230	0.1204	0.1497
	2-PL*a*	0.0281	0.4225	0.1082	0.0725
	3-PL	0.0177	0.4196	0.0744	0.0458
	3-PL*a*	0.0167	0.4193	0.0698	0.0362
	3-PL*ac*	0.0251	0.4218	0.1106	0.0842
1000	1-PL	0.0280	0.4263	0.1084	0.0639
	2-PL	0.0295	0.4221	0.0862	0.0594
	2-PL*a*	0.0293	0.4225	0.0891	0.0622
	3-PL	0.0183	0.4195	0.0505	0.0201
	3-PL*a*	0.0180	0.4198	0.0507	0.0191
	3-PL*ac*	0.0254	0.4214	0.0761	0.0407

Table 3

Indices of Stability of Estimates.

		Stability		
Sample Size	Model	b	a	Curve
100	1-PL	0.2543	0.0876	0.0421
	2-PL	0.3090	0.2518	0.0560
	2-PL*a*	0.2719	0.1300	0.0463
	3-PL	0.3486	0.3001	0.0564
	3-PL*a*	0.3627	0.1118	0.0547
	3-PL*ac*	0.2944	0.1355	0.0482
250	1-PL	0.1570	0.0513	0.0260
	2-PL	0.1992	0.1816	0.0365
	2-PL*a*	0.1852	0.1280	0.0329
	3-PL	0.2155	0.2760	0.0399
	3-PL*a*	0.2278	0.1317	0.0383
	3-PL*ac*	0.1932	0.1382	0.0334
500	1-PL	0.1114	0.0394	0.0184
	2-PL	0.1451	0.1335	0.0266
	2-PL*a*	0.1373	0.1108	0.0249
	3-PL	0.1678	0.2340	0.0317
	3-PL*a*	0.1616	0.1331	0.0289
	3-PL*ac*	0.1467	0.1212	0.0259
1000	1-PL	0.0762	0.0216	0.0124
	2-PL	0.1080	0.0989	0.0195
	2-PL*a*	0.0983	0.0869	0.0181
	3-PL	0.1229	0.1812	0.0239
	3-PL*a*	0.1143	0.1252	0.0219
	3-PL*ac*	0.1065	0.1044	0.0190

Mean Residual Values (MNR) for Six Models by Sample Size

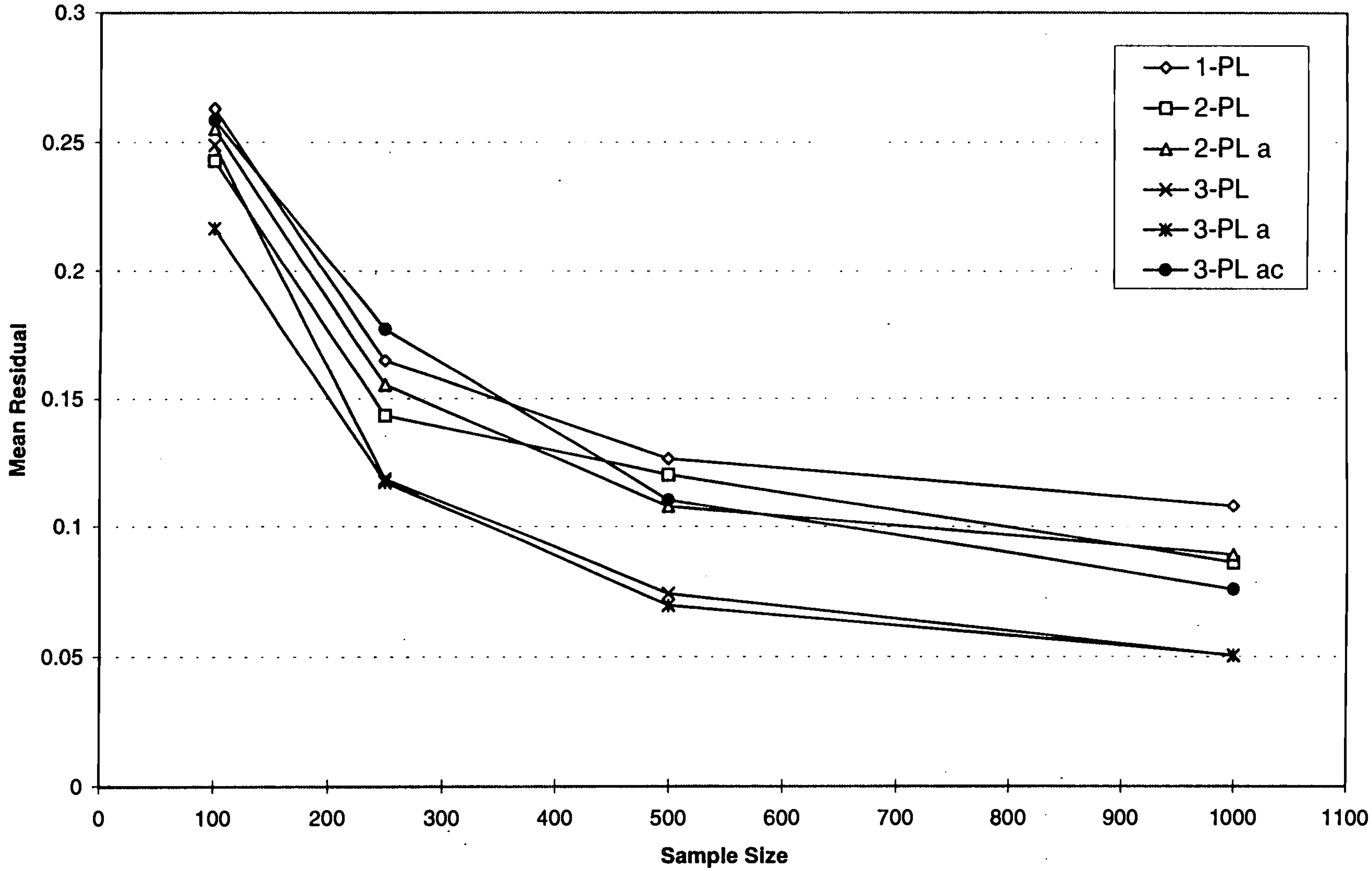

Figure 1

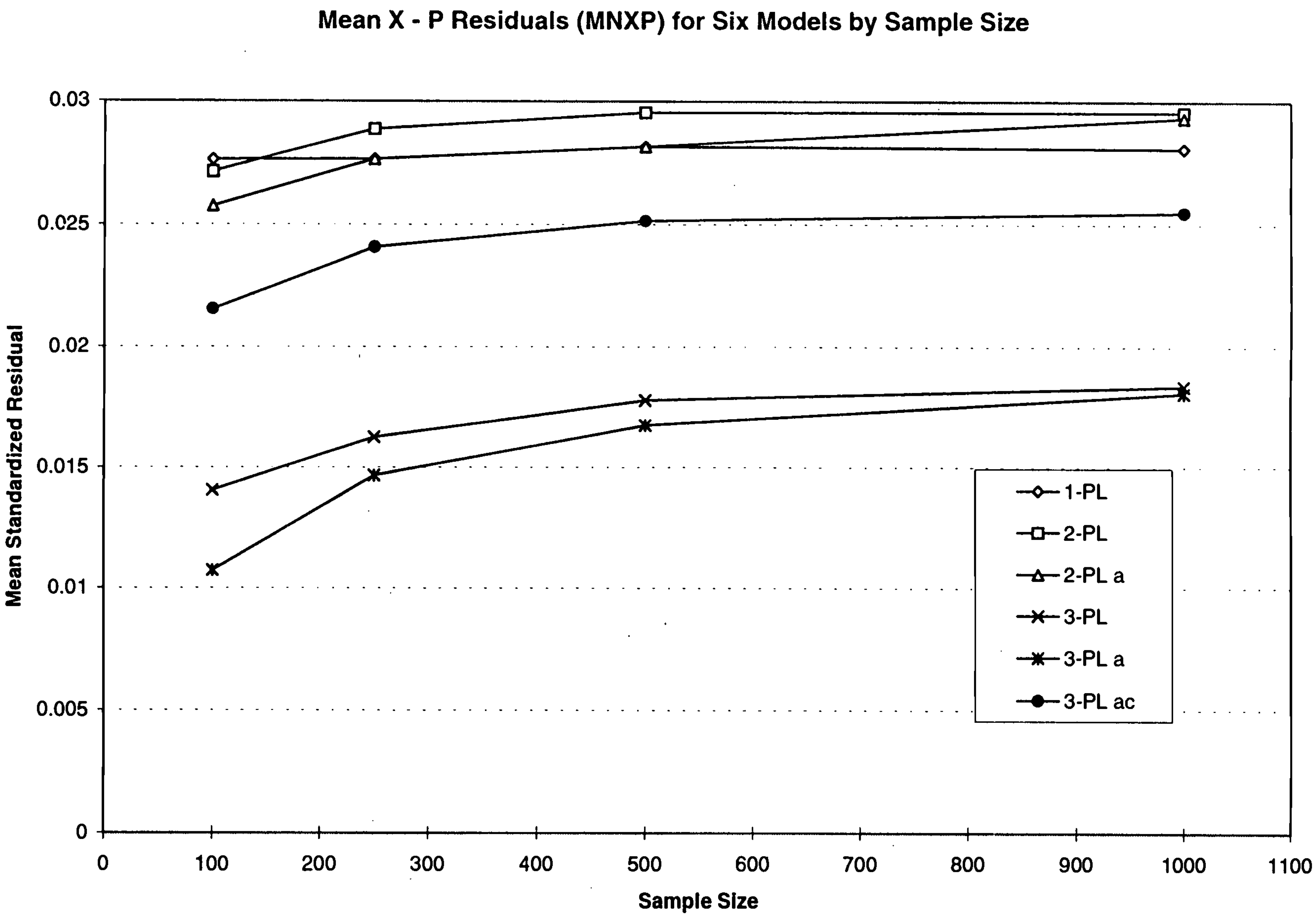
Mean X - P Residuals (MNXP) for Six Models by Sample Size
Mean Standardized Residual
0.03
0.025
0.02
0.015
0.01
0.005
0
0
100
200
300
400
500
600
700
800
900
1000
1100
Sample Size
◇ 1-PL
□ 2-PL
△ 2-PL a
✕ 3-PL
✱ 3-PL a
● 3-PL ac

Figure 2

Standard Deviation of B Parameter for Six Models by Sample Size

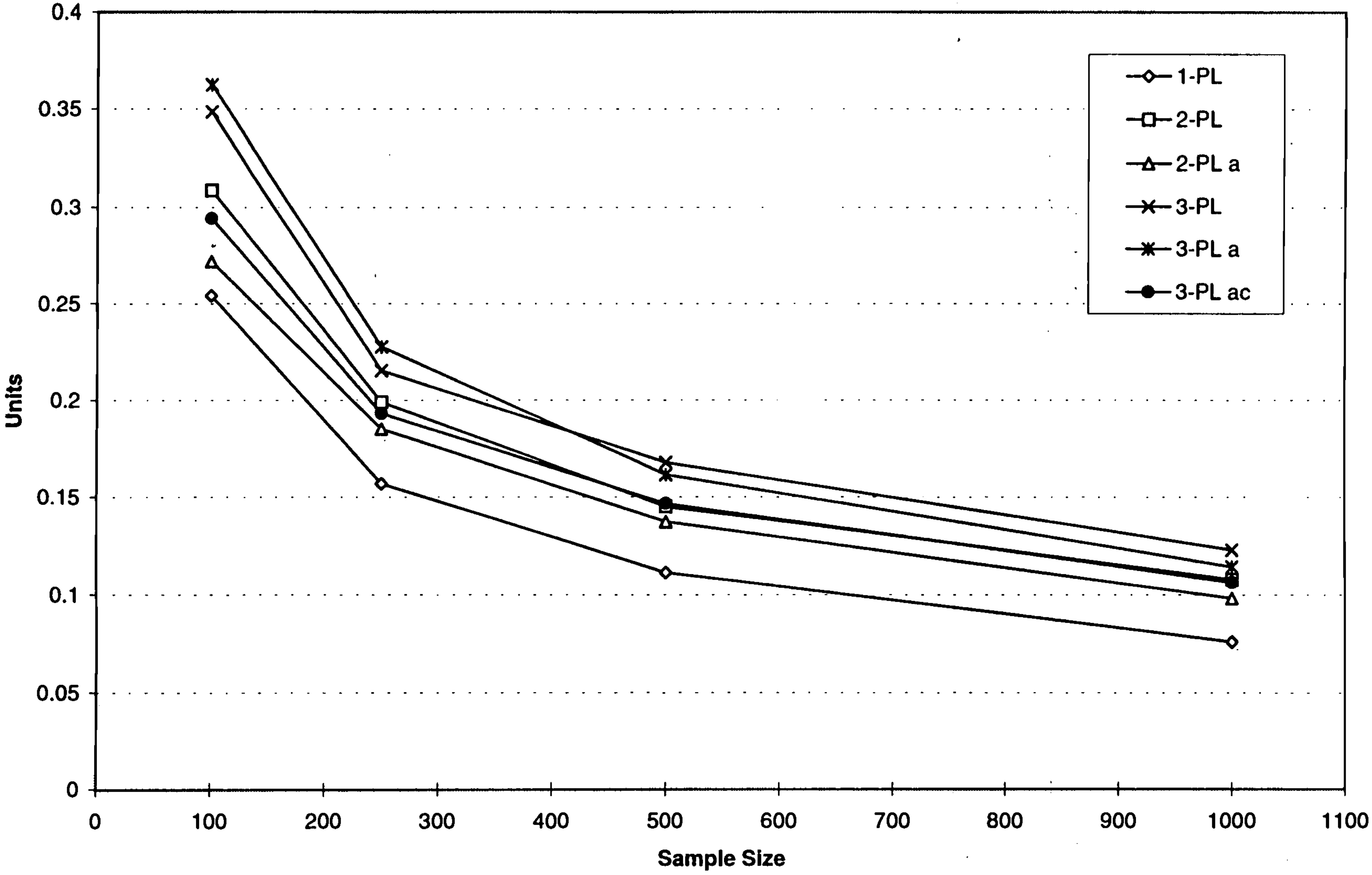

Figure 3

Standard Deviation of the A Parameter (SDA) for Six Models by Sample Size

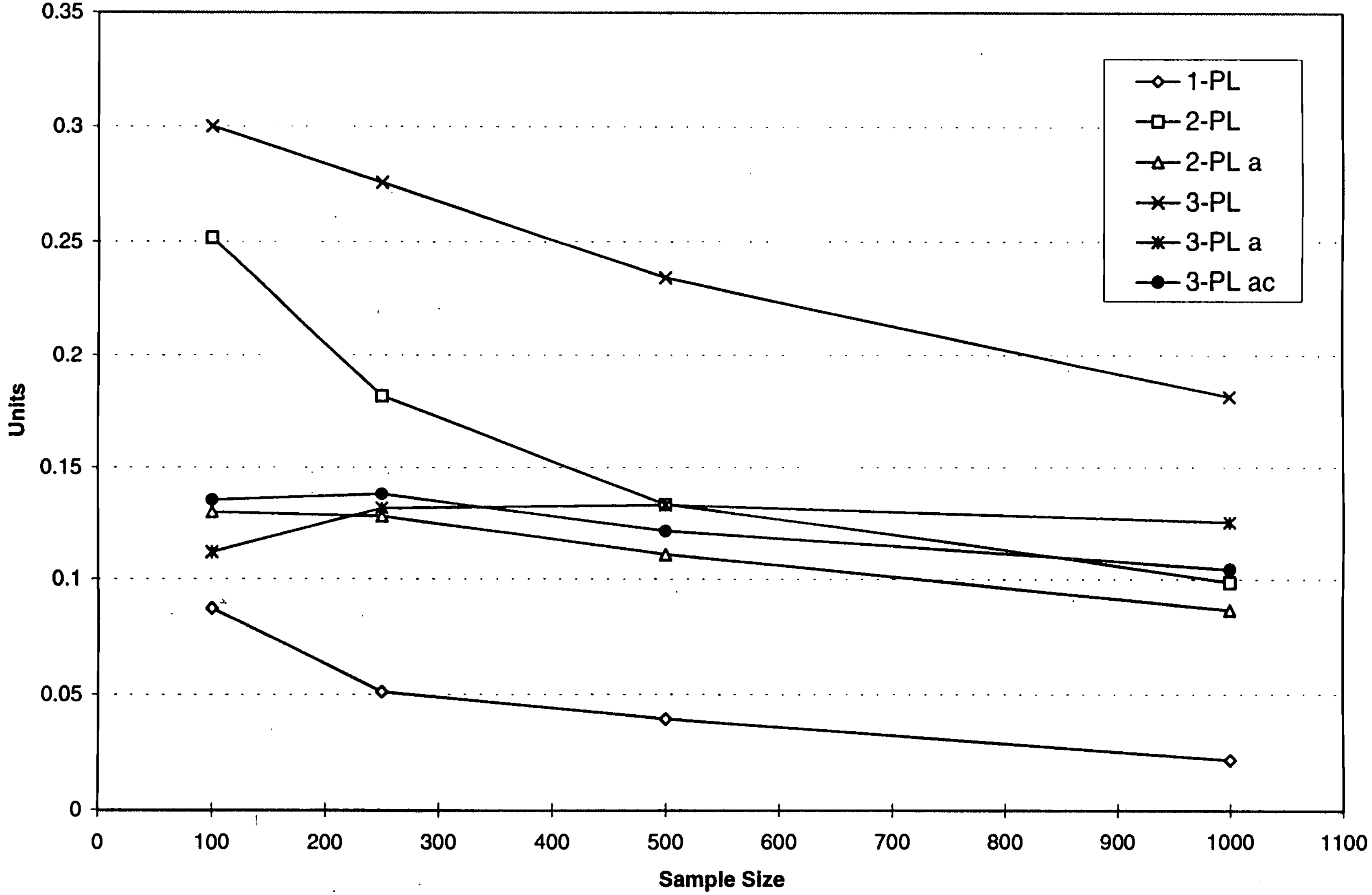

Figure 4

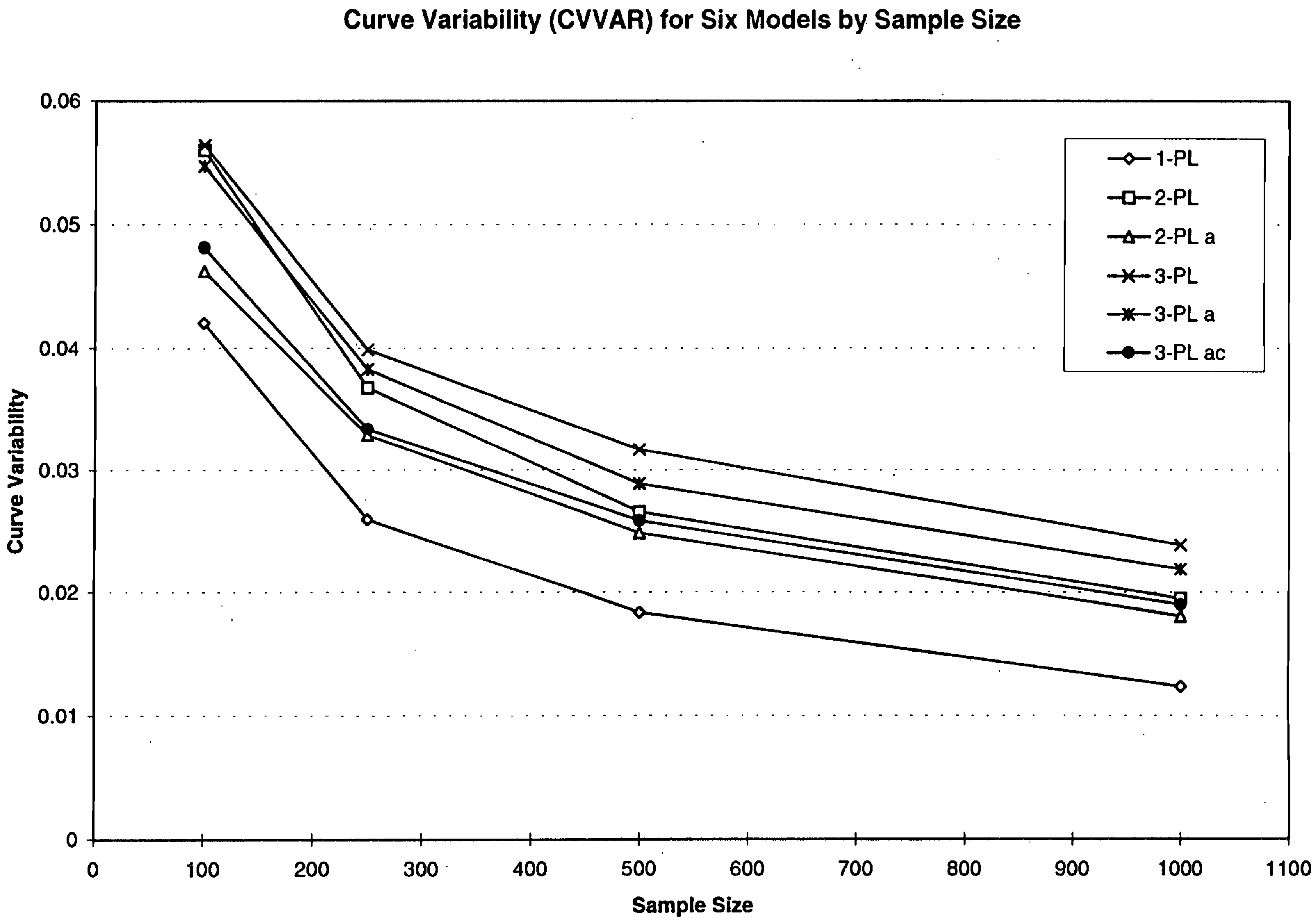
Curve Variability (CVVAR) for Six Models by Sample Size
Curve Variability
0.06
0.05
0.04
0.03
0.02
0.01
0
0
100
200
300
400
500
600
700
800
900
1000
1100
Sample Size
1-PL
2-PL
2-PL a
3-PL
3-PL a
3-PL ac

Figure 5

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

ERIC®

TM028860

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Comparison of alternative models for item parameter estimation with small samples

Author(s): Parshall, C. G., Kromrey, J. D., & Chason, W. M.

Corporate Source: Psychometric Society

Publication Date: June 1996

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 at the bottom of the page.

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) 1	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) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
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.

Sign here,→ please

Signature:

Printed Name/Position/Title: C G Parshall Psychometrician

Organization/Address: HMS 401, USF, Tampa, FL 33620

Telephone: 813/974-1256

FAX: 813/974-5132

E-Mail Address: parshall@seaweed.coedu.usf.edu

Date: 5-11-98

(over)