DOCUMENT RESUME

ED 414 332 TM 027 863

AUTHOR Patsula, Liane N.; Pashley, Peter J.

TITLE Pretest Item Analyses Using Polynomial Logistic Regression: An Approach to Small Sample Calibration Problems Associated with Computerized Adaptive Testing.

PUB DATE 1997-04-09

NOTE 42p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Ability; *Adaptive Testing; *Computer Assisted Testing; *Item Banks; Item Response Theory; *Pretesting; *Sample Size; *Test Items

IDENTIFIERS Calibration; Large Scale Assessment; Large Scale Programs; *Logistic Regression; Polynomial Regression Models

ABSTRACT

Many large-scale testing programs routinely pretest new items alongside operational (or scored) items to determine their empirical characteristics. If these pretest items pass certain statistical criteria, they are placed into an operational item pool; otherwise they are edited and re-pretested or simply discarded. In these situations, reliable ability estimates are usually available for each examinee based on operational items, and they may be treated as fixed. If so, polynomial (in ability, theta) logistic regression analyses can be conducted using a variety of statistical software packages. In this study, a cubic logistic model (theta, theta-2, theta-3) was found to fit standard three-parameter (i.e. discrimination, difficulty, and lower asymptote) logistic item response theory (IRT) model items very well. When employing a polynomial logistic model, well-known selection routines (such as stepwise elimination) can be utilized to reduce the number of required parameters for certain items, thus reducing the sample sizes needed for reliable estimation. With this model, simultaneous confidence bands are easily calculated. As an added benefit, given that a polynomial logistic function is not necessarily monotonically increasing with ability, poor quality items and incorrect alternative responses can also be fit using the same estimation procedures. (Contains 19 figures, 4 tables, and 22 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 414 332

Pretest Item Analyses Using Polynomial Logistic Regression: An Approach to Small Sample Calibration Problems Associated with Computerized Adaptive Testing

Liane N. Patsula[*]
University of Massachusetts at Amherst

Peter J. Pashley
Law School Admission Council

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Liane Patsula

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper to be presented at the National Council on Measurement in Education Annual Meeting, New York, April 9, 1996.

TM027863

[*] This project was conducted while Liane Patsula was a summer intern at Law School Admission Council in 1995.

ERIC

BEST COPY AVAILABLE

Abstract

Many large-scale testing programs routinely pretest new items alongside operational (or scored) items to determine their empirical characteristics. If these pretest items pass certain statistical criteria, they are placed into an operational item pool; otherwise they are edited and re-pretested or simply discarded. Note that in these situations, reliable ability estimates are usually available for each examinee based on operational items and may be treated as fixed. If so, polynomial (in ability, θ) logistic regression analyses can be conducted using a variety of statistical software packages. In this study, a cubic logistic model (θ, θ^2, θ^3) was found to fit standard three-parameter (i.e., discrimination, difficulty and lower asymptote) logistic item response theory (IRT) model items very well. When employing a polynomial logistic model, well-known selection routines (such as stepwise elimination) can be utilized to reduce the number of required parameters for certain items, thus reducing the sample sizes needed for reliable estimation. With this model, simultaneous confidence bands are easily calculated. As an added benefit, given that a polynomial logistic function is not necessarily monotonically increasing with ability, poor quality items and incorrect alternative responses can also be fit using the same estimation procedures.

Pretest Item Analyses Using Polynomial Logistic Regression: An Approach to Small Sample Calibration Problems Associated with Computerized Adaptive Testing

Liane N. Patsula, University of Massachusetts at Amherst
Peter J. Pashley, Law School Admission Council

As is evident in the measurement literature of the past twenty years, the use of item response theory (IRT) by test developers to analyze test data has become increasingly prominent. For example, Law School Admission Council (LSAC) currently utilizes IRT methodology to assess the characteristics of pretest items, to direct form assemblies, and for equating purposes. The three most popular IRT models in common use are the three-parameter logistic (3PL) model (Birnbaum, 1968), the two-parameter logistic (2PL) model (Lord, 1952), and the one-parameter logistic (1PL or Rasch) model (Rasch, 1960). These models are appropriate for dichotomous item response data. The 3PL is the most general model and is defined mathematically as:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}} . \quad (1)$$

In this model, $P(\theta)$, the probability that a randomly chosen examinee with ability θ will respond correctly to an item, is defined by the discrimination (a), difficulty (b), and pseudo-guessing or lower asymptote (c) parameters. The scaling constant 1.7 (or sometimes 1.702) is often included so that the item parameters a and b correspond approximately to those that would have been obtained had a normal-ogive function been employed instead of a logistic function (see next section for more details on the logistic function). The 2PL and 1PL models are restricted cases of the 3PL model.

The number of item parameters required is usually dictated by the types of items being used and the responses they elicit. For example, multiple-choice items can be answered correctly through guessing and so a lower asymptote adjustment (i.e., the inclusion of a *c* parameter) is often needed to help model the behavior of low ability examinees. If this is not the case, as with constructed response items, the *c* parameter might be dropped, and a 2PL model, which accounts for differences in discrimination and difficulty across items, may be used. If a lower asymptote adjustment is not needed and the items are very similar with regard to their discriminating power (resulting in a constant *a* parameter across all items), a 1PL model may suffice; the 1PL model simply differentiates between items on the basis of difficulty.

Due to the mathematical complexity of the parameter estimation procedures associated with all of these IRT logistic models, computer programs are almost always employed to accomplish that task. The most popular packages for estimating the parameters of a 3PL model are LOGIST (Wingersky, Barton, & Lord, 1982) and BILOG (Mislevy & Bock, 1984, 1986), the latter of which is used at LSAC. LOGIST employs a joint maximum likelihood estimation procedure (Birnbaum, 1968) to estimate item (*a*, *b*, *c*) and person (θ) parameters. According to Wingersky (1983), large numbers of examinees and items (on the order of 1,000 or more examinees and 40 or more items) should be used to obtain stable 3PL model parameter estimates in LOGIST. BILOG is a more recent estimation program that utilizes marginal maximum likelihood (Bock & Aitkin, 1981) and Bayesian estimation procedures (Mislevy, 1986). Both LOGIST and BILOG can also calibrate items assuming the more restricted 2PL and 1PL models. BILOG usually requires slightly fewer numbers of examinees and items than LOGIST to obtain stable parameter estimates (Mislevy & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987). However, a general rule-of-thumb suggests that sample sizes of at least 1000, 500, and 300 examinees are needed to accurately estimate the parameters of the 3PL, 2PL, and 1PL models, respectively (Hambleton, Swaminathan, & Rogers, 1991). As with any statistical estimation procedure, if fewer parameters need to be estimated, fewer data points are usually required.

While both LOGIST and BILOG can calibrate items assuming the 3PL, 2PL or 1PL model, users of these packages must specify a priori which model should be applied to all items. This is true even though the data may suggest that some items do not require a *c* parameter, for example.

This requirement by LOGIST and BILOG probably results in an over-parameterization of datasets in many cases. Many theorists (including most Rasch modelists) believe that the 3PL is fundamentally over-specified (e.g., Wright & Stone, 1979). Note that LOGIST attempts to overcome this problem of over-specification by allowing users the option of estimating a common *c* parameter for a subset of items, and BILOG allows users to specify prior distributions for the parameters to be estimated.

While in general the 3PL may be over-specified, all three parameters are probably needed for some items. In particular, very difficult multiple-choice items with strong discriminating power will probably require something like a lower asymptote to model guessing behavior that occurs throughout the middle to low ability range. Note that guessing behavior is especially prevalent when there is no penalty for guessing in the scoring of the test.

In practice, some items may require even more than three parameters to accurately describe their behavior throughout the ability range. Barton and Lord (1981) suggested that an upper asymptote might be added to account for less than perfect responding by high ability examinees on some very difficult items. The extremely flexible non-parametric item regressions described by Ramsay (1991) and produced by his program TESTGRAF (Ramsay, 1993) often estimate item characteristic curves (ICCs) that exhibit more plateaus then can be modeled by a standard 3PL model. The item regressions of Ramsay can also model the non-monotonically increasing behavior of poor quality items. One drawback to Ramsay's approach is that users must specify, subjectively, the level of "smoothness" to be associated with individual ICCs. McDonald (1982) has also investigated alternatives to the standard IRT models, using polynomial approximations. One drawback to McDonald's nonlinear models is that they are not constrained to the probability correct space (i.e., estimated probabilities may be less than zero or greater than one).

This paper investigates another alternative ICC modeling technique that could be used in conjunction with standard IRT calibrations; namely, polynomial logistic regression. These are referred to here as "regression" models because ability estimates are not concurrently derived, but rather are assumed to be known a priori. This situation may occur in practice, for example, when analyzing pretest items. Typically, in this case the pretest item characteristics are unknown, but examinee abilities are known (or at least very well estimated) from the operational (or scored) sections of a test. In practice, some of these pretest items might be poorly behaved. Pretest item analyses often identify questionable items that sometimes cannot even be calibrated by BILOG, for example. Some of these items exhibit negative biserial correlations with test scores, suggesting that the ICCs are not necessarily monotonically increasing functions of ability. The polynomial logistic regression models investigated here are not necessarily monotonically increasing functions and thus have the added flexibility to model very unusual items. While these unusual items might not be used as is, modeling them correctly may relay important information to test development staff that could be used to edit the items. As an added benefit, because they are not necessarily monotonically increasing, these more general logistic regression functions can be used to model responses to individual distractors (or alternatives).

Three different datasets will be used in this study to empirically evaluate the capacity of polynomial logistic functions to fit item response data. First, a dataset comprised of real responses to operational multiple-choice items was obtained. Second, after calibrating the real dataset with BILOG, resulting item and ability estimates were used to simulate standard 3PL model item responses. Third, a dataset comprised of real responses to pretest items was obtained. The simulated dataset was used to assess the fit of the polynomial logistic function to item responses that are known to be standard 3PL. The original operational dataset was used to assess the fit of polynomial logistic functions to real and well-behaved item responses based on various examinee sample sizes. Finally, the pretest dataset was used to illustrate the fit of polynomial logistic functions to real, but not necessarily well-behaved, item responses.

The Polynomial Logistic Model

A general form of the logistic function is

$$\Psi(\lambda) = \frac{e^{\lambda}}{1 + e^{\lambda}} \equiv \frac{1}{1 + e^{-\lambda}} , \tag{2}$$

ERIC Full Text Provided by ERIC

where λ is called a logit. One example of a logit is $\lambda = 1.7a(\theta - b)$, which is a 2PL logit. Thus, the 2PL model can be written as

$$P(\theta) = \Psi\left(1.7a(\theta - b)\right) . \tag{3}$$

If the *a* parameter in this 2PL logit was assumed constant across all items calibrated, the result defines a logit for a 1PL model.

The 3PL model (Equation 1) can be re-written as

$$P(\theta) = c + (1 - c)\Psi\left(1.7a(\theta - b)\right) . \tag{4}$$

To avoid confusion with other logistic functions discussed below, this standard IRT 3PL will be referred to as the Birnbaum 3PL in the rest of this paper. Note that the Birnbaum 3PL does not take the form of a natural logistic function (as defined in Equation 2) because the *c* parameter is not incorporated in the logit. When it is, the Birnbaum 3PL can be defined as:

$$P(\theta) = \Psi\left(\ln\left\{1 + e^{[1.7a(\theta - b) - \ln c]}\right\} + \ln\left\{\frac{c}{1 - c}\right\}\right) . \tag{5}$$

Clearly, the complete logit (i.e., where all parameters are incorporated into the logit) of a Birnbaum 3PL is not linear, as is the case for the 2PL and 1PL models. This fact may explain some of the estimation difficulties associated with the Birnbaum 3PL model. Figure 1 illustrates the shape of a typical Birnbaum 3PL in both the probability correct and logit spaces. Upper and lower asymptotes, defined respectively by $1.7a(\theta - b)$ and $\ln[c/(1 - c)]$, are displayed in the logit space plot. These asymptotes can be intuitively derived from Equation 5 by letting θ tend to negative or positive infinity. Note that 2PL models define straight lines in the logit space and 1PL models define straight parallel lines.

A general form of a polynomial logistic function in terms of ability, θ, is

$$P(\theta) = \Psi(\beta_0 + \beta_1\theta + \beta_2\theta^2 + \beta_3\theta^3 + \beta_4\theta^4 + \ldots) , \tag{6}$$

where the β_i's are item parameters. As indicated above, the θ values will assumed to be fixed, and the β_i's need to be estimated. As with all standard logistic functions, the $P(\theta)$ in Equation 6 is restricted to be greater than zero and less than one. If the logit in Equation 6 is reduced to its linear component (i.e., $\beta_0 + \beta_1\theta$), the result is mathematically equivalent to the 2PL given in Equation 3.

Preliminary investigations indicated that a third-order (or cubic) polynomial logit is probably sufficient to model a wide variety of multiple-choice items. To demonstrate the fit of polynomial logistic functions to 3PL data, figure 2 illustrates the ICC of a typical Birnbaum 3PL item, along with best fitting linear (i.e., 2PL), quadratic and cubic logistic functions, using a uniform [-3, 3] ability distribution. The residual plot included in Figure 2 subtracts the true Birnbaum 3PL probabilities from those calculated from the other logistic functions. Clearly, cubic, and even quadratic, logistic functions can mimic this typical Birnbaum 3PL item. The 2PL does not do as well in this case. However, when the fitting is performed using a standard normal ability distribution, the 2PL does fit the middle and upper portions of this particular Birnbaum 3PL better, but still not as well as the quadratic and cubic logistic functions.

Simultaneous Confidence Bands: Confidence bands for response functions are good indicators of the precision with which data are being modeled. Due to the non-linear nature of the logit associated with the Birnbaum 3PL model, analytical solutions are not available for deriving associated ICC simultaneous confidence bands, even when ability estimates are assumed fixed (Lord & Pashley, 1988). In contrast, Hauck (1983) provided a very simple analytical approach to obtaining simultaneous confidence bands for polynomial logistic models. This approach is now briefly presented without details.

In the most general case, let the vectors

$$\beta' = (\beta_0, \ \beta_1, \ \beta_2, \ ...) \tag{7}$$

and

$$\gamma' = (1, \ \theta, \ \theta^2, \ ...) \tag{8}$$

contain the item and person parameters, respectively. Then in matrix notation, Equation 5 can be expressed as

$$P(\theta) = \Psi\left(\beta'\gamma\right) . \tag{9}$$

In the logit space, a 1 - α simultaneous confidence band for $\lambda = \beta'\gamma$ is given by

$$\left[\lambda_L, \ \lambda_U\right] = \left[\beta'\gamma - \left(\chi^2_{k+1,\alpha}\gamma' V^{-1}\gamma\right)^{\frac{1}{2}}, \ \beta'\gamma + \left(\chi^2_{k+1,\alpha}\gamma' V^{-1}\gamma\right)^{\frac{1}{2}}\right], \tag{10}$$

where $\mathbf{V}^{-1}$ is the large-sample covariance matrix for β, which is usually estimated by most statistical packages that perform logistic regression, and the chi-square value has $k + 1$ degrees of freedom, which equals the order of the polynomial used plus one.

The associated simultaneous confidence band in the probability-correct space is then

$$\left[P_L(\theta), \ P_U(\theta)\right] = \left[\Psi(\lambda_L), \ \Psi(\lambda_U)\right] . \tag{11}$$

Empirical Analyses

To investigate the cubic logistic regression model empirically, real and simulated data were analyzed. The real dataset was obtained by randomly sampling 2,000 examinees from the approximately 42,000 test takers who took a recent administration of the Law School Admission Test (LSAT)[1]. The first 100 of the 101 scored items were used for the 2,000 randomly sampled examinees. All omits (i.e., no responses) were scored as incorrect. The computer program BILOG was used to estimate item and ability parameters for this dataset, using all default settings and assuming a Birnbaum 3PL model. As can be seen from the ICCs in Figure 3, many items require a Birnbaum 3PL model to adequately fit the data. However, some items could be modeled using the 1PL or 2PL. In practice, though, a single model is usually employed for all items, rather than different models for different items. The ability distribution from this real dataset was found to be fairly normal between -2 and 2, as shown in the normal plot in Figure 4. The mean, standard deviation, and skewness values for the sample of 2,000 examinees were approximately 0, 1 and -.27, respectively.

The first evaluation of the polynomial logistic function involved fitting data known to be 3PL (i.e., simulated data). BILOG parameter estimates from the real dataset were treated as "true" parameters and used to generate a "simulated" dataset of responses of 2,000 simulees to 100 Birnbaum 3PL items. The simulated dataset was then re-calibrated with BILOG. The ability estimates produced from this second BILOG run, rather than the "true" values, were used as fixed values for subsequent polynomial logistic regression analyses. (The rationale for this is that the "true" values are not known in real test data. The ability estimates from the operational items would be used as the fixed values when performing polynomial logistic regression analyses

[1] The LSAT is large-scale paper-and-pencil assessment of reading comprehension and reasoning skills that is generally administered to college seniors and college graduates for use as an aid in making law school admission decisions. The LSAT currently consists of four sections: Reading Comprehension, Analytical Reasoning, and two Logical Reasoning, usually contributing to a 101 five-choice item test. An unscored variable fifth section is also administered to all examinees in addition to the scored items; the variable section is used to pretest or pre-equate items.

on pretest items. Calibrating the simulated data and using the resulting ability estimates is somewhat analogous to using the ability estimates from operational data while investigating pretest items, under the assumption of unidimensionality.)

The second evaluation of the polynomial logistic function involved fitting the original "real" dataset that contained responses that did not necessarily adhere to the Birnbaum 3PL model. When these items were modeled with polynomial logistic functions, the original BILOG ability estimates (i.e., from the first BILOG calibration of the real dataset) were used as fixed values. Again, this procedure is somewhat different from using operational score item ability estimates as fixed while analyzing separate pretest items. However, the purpose here is to evaluate the modeling capabilities of the polynomial logistic function, rather than assessing the influence of fixed ability distributions on fitting items.

Finally, 24 pretest items were analyzed. In this case, ability estimates from the operational (or scored portion) of the test were assumed fixed for the 1,000 examinees who were given these 24 pretest items and used to fit polynomial logistic functions to all 24 pretest item alternatives.

Large-Sample Results. In general, the cubic logistic regression model was found to mimic simulated Birnbaum 3PL data very well. The polynomial logistic regressions were performed using the LOGISTIC procedure in the statistical package SAS. A stepwise selection procedure was evoked, and reduced models resulted for many items. Table 1 provides counts of the number of items which required a linear (2PL), quadratic, or cubic polynomial for the simulated and real datasets. Interestingly, almost half the simulated and real responses were fit well using a 2PL model.

Figures 5-9 illustrate how well the estimated polynomial logistic functions and Birnbaum 3PL BILOG estimates matched the true Birnbaum 3PL curves. The 95% simultaneous confidence bands shown are for the polynomial logistic functions. Figure 5 illustrates a typical item from these analyses; the polynomial logistic function fits very well. Figure 6 provides an example where the polynomial logistic function does not perform as well as BILOG in the lower and upper tails, while Figure 7 provides an example where it actually does better in the tails but not as well in the middle range. Figure 8 illustrates the wider confidence bands that result from using the full cubic logistic model. Finally, Figure 9 gives a rare example of fitting (or misfitting) a non-monotonically increasing function. Note, however, that the simultaneous confidence bands given in Figure 10 always include the true Birnbaum 3PL throughout the displayed ability range.

Evaluating Birnbaum 3PL Fits. The polynomial logistic function can also be used to assess the fit of Birnbaum 3PL models to real data. To accomplish this, the polynomial logistic function calibrations of the real dataset were re-run with an added term. This added term was the Birnbaum 3PL complete logit (see Equation 5) with parameters estimated by BILOG. If the Birnbaum 3PL was the best fitting model, one would expect the stepwise selection algorithm to eliminate all terms but the added logit term and the intercept, and their estimated β coefficients would be expected to be approximately one and zero, respectively. In fact, this occurred for only 56 of the 100 items investigated. Of the rest, 35 were modeled as 2PL and nine others were a mix of specifications. This analysis indicates that, for at least this data, the Birnbaum 3PL generally does a good job of modeling real multiple-choice item responses, but could be reduced to a 2PL in many cases.

Can the 2PL also be reduced to a 1PL for many items? Figure 10 plots the slope versus the intercept terms for the 48 real items that were modeled well by a 2PL (see Table 1). Clearly there is a wide dispersion of slope parameters. Some could be grouped to develop small Rasch (1PL) subtests, but the 48 items taken together certainly do not constitute a Rasch test.

Modeling Response Alternatives. Figures 11 through 14 illustrate polynomial logistic function fits of correct and incorrect responses to four real pretest items based on responses from 1,000 real examinees. Ability estimates for the examinees were derived from their operational sections using BILOG. Figure 11 displays a quite easy item with all distractors functioning at about the chance level. Figure 12 illustrates another easy item, but with one preferred distractor, especially around the middle of the ability range. Figure 13 illustrates a fairly hard item and, again, one strong distractor. Finally, Figure 14 displays another hard item, but with a more even pull among distractors. These types of graphs can be very helpful to test developers who are concerned with the influence of distractors on item performance. Note that in all these four cases the correct response ICCs are monotonically increasing, the alternatives are not. So, while BILOG could model the correct response of these four pretest items, it could not model the alternatives. Within a pretest item analysis, for example, if one of these four items was miss-keyed, BILOG would simply skip calibrating it.

Small-Sample Results. Because the use of polynomial logistic regression models have been seen to reduce the numbers of estimated parameters, one natural application of such models is to the estimation of item parameters with small sample sizes. As a preliminary investigation of the effects of small samples on polynomial logistic function estimation, the original dataset containing real item responses was reduced from 2,000 examinees to samples of 1000, 750, 500 and 250. These samples were chosen to be inclusive in ascending order. That is, the 250 examinees in the smallest sample were also included among the 500 examinees in the second smallest sample, who were all included in the third smallest sample, and so on. (Note that the ability estimates used as fixed independent regression variables were calculated from the full sample, as before.) This design might be appropriate for CAT environments in which statistics for new items could be checked periodically while they are being pretested. As such, obviously unstable items might then be removed from pretesting after relatively few responses (i.e., 250 or 500) are collected to release a pretest position that might be used to investigate a potentially better item.

Table 2 contains selected polynomial logistic function frequencies for the different sample sizes, including the original sample (i.e., the first row of Table 2 corresponds to the last row of Table 1). As expected, the number of 2PL fits increases as the sample size decreases, because with smaller sample sizes additional parameters cannot be estimated as well. Interestingly, the largest increase in frequency of 2PL selection is from the 2,000 to 1,000 sample sizes. This may suggest that the rule-of-thumb indicating sample sizes of 1,000 are minimally sufficient for 3PL calibrations might be too low for some cases.

Tables 3 and 4 provide more specific information on the comparative fits of the polynomial logistic function, based on various sample sizes, versus BILOG estimated Birnbaum 3PL ICCs for the 100 real items. Note that the BILOG estimates were based on the full 2,000 examinee sample. Table 3 contains bias statistics for seven ability levels. As expected, the least amount of bias is found in the middle ability range. Interestingly, for the smaller sample sizes the increase in bias below the zero ability level appears less than the corresponding increase in bias above the zero ability level. The root mean square errors (RMSEs) in Table 4 are more symmetrical about the zero ability level. Tables 3 and 4 suggest that the polynomial logistic function can do a relatively good job of identifying the shape of ICCs for good quality items, even with small sample sizes, at least in comparison to large-sample BILOG calibrations.

Figures 15 through 19 provide examples of the fit of polynomial logistic regressions to real item responses of various sample sizes. These residuals are derived with regard to the BILOG estimates obtained by calibrating all original real responses to 100 items by 2,000 examinees. The items displayed in Figures 15 through 19 correspond to those shown in Figures 5 through 9, respectively. Note that while Figures 5 through 9 are based on calibrations of simulated data, the curves designated as "True 3PL" are in fact the BILOG 3PL estimates obtained by calibrating all original real responses to 100 items by 2,000 examinees, which were then used as generating values.

To make Figures 15 through 19 comparable, the residual scale was kept constant. As a result, some of the residual curves are cropped in the tails. However, within the range of ability containing most of the examinees, the residual curves typically lie well within the -.05 to .05 range for these five items, and for most of the remaining 95 items not shown. Some examples to the contrary do exist, however. For instance, the 2PL model selected for the sample size of 250 in Figure 19 departs markedly from the BILOG 3PL fit below the -1 ability level. Note that the BILOG 3PL estimate might not always be the best fitting curve for all items. However, given that residuals typically decrease with large sample sizes, as shown in Figures 15 through 19 (and in Table 3), there is some indication that the BILOG 3PL model is reasonable for this data.

Summary

Currently, LSAT items, like those of many other large-scale testing programs, are pretested, then pre-equated, and finally given operationally. Hopefully, poor performing items are weeded out at the pretest stage. These poor quality items can elicit erratic responses from examinees and are not easy to model at times, especially with a monotonically increasing Birnbaum 3PL model. This study investigated polynomial logistic functions as a possible adjunct to the Birnbaum 3PL model when ability estimates are available and more model flexibility is required, such as in the case of pretest item analyses.

Results indicate that the polynomial logistic function can fit Birnbaum 3PL items very well. This model can also be used to evaluate the choice of the Birnbaum 3PL model or reduced models (i.e., 2PL or 1PL) for a set of data. This was accomplished in this paper by including an estimated Birnbaum logit term in the polynomial logistic regression specifications and then seeing whether that term was eliminated or not in a sequential analysis (i.e., by employing a stepwise elimination procedure).

As the polynomial logistic function is not necessarily monotonically increasing, it can be used to fit incorrect as well as correct responses. These alternative response curves can be very informative to test developers who might wish to edit poorly performing items. The precision of polynomial logistic curve estimation can easily be obtained with analytically derived simultaneous confidence bands. Finally, polynomial logistic calibrations can be performed with a variety of popular statistical packages, such as SAS and SPSS. Standard packaged routines usually provide users with the option of reducing the model specification by selecting out noncontributing variables.

In computerized testing programs that provide on-demand assessments, item exposure is more of a concern than in regularly scheduled paper-and-pencil administrations. To minimize item exposure and reduce testing time within a computerized testing environment, responses to pretest items should be collected from as few examinees as possible while still producing reliable item statistics. Preliminary results given in this paper indicate that a polynomial logistic regression approach, that includes a variable elimination routine, can provide relatively stable estimates within small sample conditions.

The findings discussed in this paper suggest that the polynomial logistic regression model might be a useful tool, especially for the purposes of pretest item analyses under large and small sample conditions. However, practioners will need to evaluate the efficacy of this model within their particular assessment environment.

References

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11*, 111-141.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Bulletin 81-20.* Princeton, NJ: Educational Testing Service.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Hambleton, R, Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. *The American Statistician*, *37*, 158-160.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Lord, F. M., & Pashley, P. J. (1988). Confidence bands for the three-parameter logistic item response curve (*Research Report RR-88-67*). Princeton, NJ: Educational Testing Service.

McDonald, R. P. Linear versus non-linear models in item response theory. *Applied Psychological Measurement*, *6*, 379-396.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-195.

Mislevy, R. J., & Bock, R. D. (1984). *BILOG: Item analysis and test scoring with binary logistic models* [Computer Program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Maximum likelihood item analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Qualls, A. L., & Ansley, T. N. (1985, April). *A comparison of item and ability parameter* *estimates derived from LOGIST and BILOG.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Ramsay, J. O. (1993). *TESTGRAF: A Program for the Graphical Analysis of Multiple Choice Test Data* [Computer Program]. Montreal: McGill University.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. [Expanded edition University of Chicago Press, 1980.]

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397-412.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.). *Applications of Item Response Theory* (pp. 45-56). British Columbia: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

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

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

Table 1

Number of Polynomial Logistic Function Types Selected While Modeling Simulated and Real Responses to 100 Items

	Polynomial Logistic Function Selected		
Dataset	Linear (2PL)	Quadratic	Cubic*
Simulated	43	39	12
Original	48	33	19

*All third-order models selected were classified as cubic in this table, including some that lacked a linear and/or a quadratic term.

Table 2

Number of Polynomial Logistic Function Types Selected While Modeling 100 Items with Varying Examinee Sample Sizes

	Polynomial Logistic Function Selected			
Sample Size	Constant	Linear (2PL)	Quadratic	Cubic*
2,000	0	48	33	19
1,000	0	64	21	15
750	0	64	24	12
500	2	71	19	8
250	3	77	9	11

*All third-order models selected were classified as cubic in this table, including some that lacked a linear and/or a quadratic term.

Table 3

Polynomial Logistic Function Real Data Calibration Bias[*] Statistics
for Various Ability Levels and Sample Sizes

	Ability Levels						
Sample Sizes	-3	-2	-1	0	1	2	3
2,000	-0.014	-0.005	0.002	0.001	-0.003	-0.003	-0.003
1,000	-0.022	-0.011	0.001	0.004	-0.003	-0.011	-0.023
750	-0.021	-0.010	0.001	0.003	-0.004	-0.008	-0.016
500	0.001	-0.011	0.002	0.004	-0.004	-0.012	-0.035
250	-0.017	-0.020	0.001	0.006	-0.008	-0.016	-0.054

[*] Bias was calculated as the mean residual from the 100 BILOG estimated ICCs using all 2,000 examinees.

Table 4

Polynomial Logistic Function Real Data Calibration RMSE[*] Statistics
for Various Ability Levels and Sample Sizes

	Ability Levels						
Sample Sizes	-3	-2	-1	0	1	2	3
2,000	0.029	0.017	0.008	0.006	0.006	0.016	0.031
1,000	0.050	0.030	0.021	0.015	0.015	0.033	0.062
750	0.061	0.037	0.025	0.019	0.018	0.036	0.066
500	0.166	0.071	0.036	0.024	0.029	0.063	0.113
250	0.197	0.094	0.047	0.035	0.046	0.081	0.171

[*] RMSE was calculated as the standard deviation of the residuals from the 100 BILOG estimated ICCs using all 2,000 examinees.

Figure Captions

Figure 1: A Birnbaum 3PL with item parameters $a = 1$, $b = 0$, and $c = .2$, in the (i) probability correct space, and (ii) logit space. Upper and lower asymptotes are also given in the logit space.

Figure 2: Best fitting polynomial logistic functions of varying orders to a Birnbaum 3PL response curve in the (i) probability correct space, and (ii) as residuals (i.e., minus the true curve).

Figure 3: Estimated Birnbaum 3PL item response curves for 100 LSAT five-choice items.

Figure 4: Normal plot of 2,000 examinee ability estimates.

Figure 5: Example of a well fitting polynomial logistic function.

Figure 6: Example of a polynomial logistic function that does not model the true Birnbaum 3PL item response curve in the upper and lower tails as well as a BILOG estimated curve.

Figure 7: Example of a polynomial logistic function that models a Birnbaum 3PL item response curve in the upper and lower tails better than a BILOG estimated curve.

Figure 8: Example of an item that required a cubic logistic function approximation.

Figure 9: Example of a non-monotonically increasing polynomial logistic function estimate.

Figure 10: Slope versus intercept estimates for 48 items selected as 2PL functions.

Figure 11: Example of polynomial logistic item alternative response curves for an easy pretest item. (The key for this item is alternative "C.")

Figure 12: Example of polynomial logistic item alternative response curves for an easy pretest item with one distinct distractor. (The key for this item is alternative "A.")

Figure 13: Example of polynomial logistic item alternative response curves for a hard pretest item with one distinct distractor. (The key for this item is alternative "C.")

Figure 14: Example of polynomial logistic item alternative response curves for a hard pretest item with three active distractors. (The key for this item is alternative "D.")

Figure 15: Estimated polynomial logistic function minus a BILOG estimate (designated as "True 3PL" in Figure 5) residual curves for various examinee sample sizes.

Figure 16: Estimated polynomial logistic function minus a BILOG estimate (designated as "True 3PL" in Figure 6) residual curves for various examinee sample sizes.

Figure 17: Estimated polynomial logistic function minus a BILOG estimate (designated as "True 3PL" in Figure 7) residual curves for various examinee sample sizes.

Figure 18: Estimated polynomial logistic function minus a BILOG estimate (designated as "True 3PL" in Figure 8) residual curves for various examinee sample sizes.

Figure 19: Estimated polynomial logistic function minus a BILOG estimate (designated as "True 3PL" in Figure 9) residual curves for various examinee sample sizes.

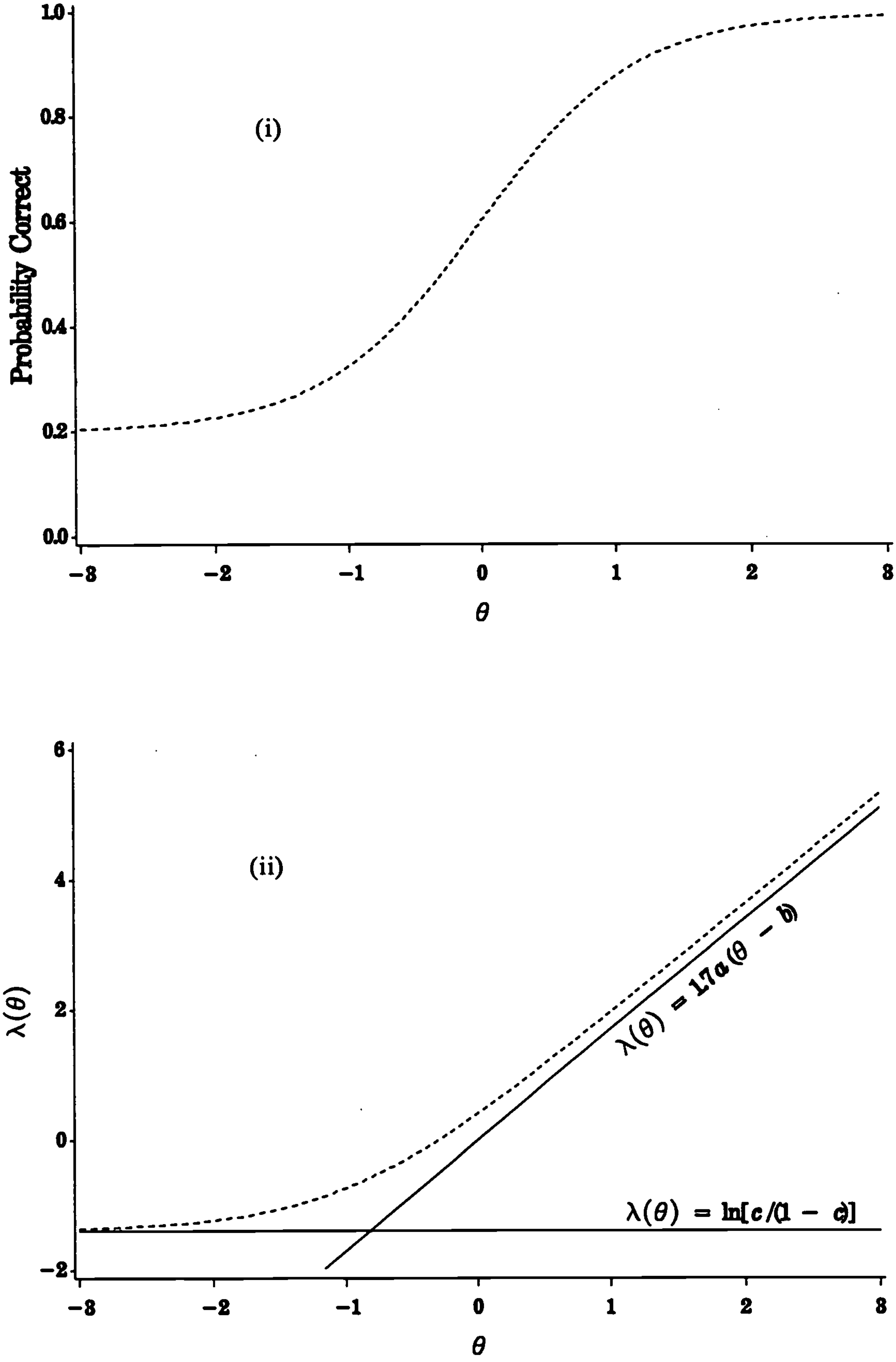

(i)
Probability Correct
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
θ
(ii)
λ(θ)
6
4
2
0
−2
λ(θ) = 1.7a(θ − b)
λ(θ) = ln[c/(1 − c)]
−3
−2
−1
0
1
2
3
θ

Figure 1

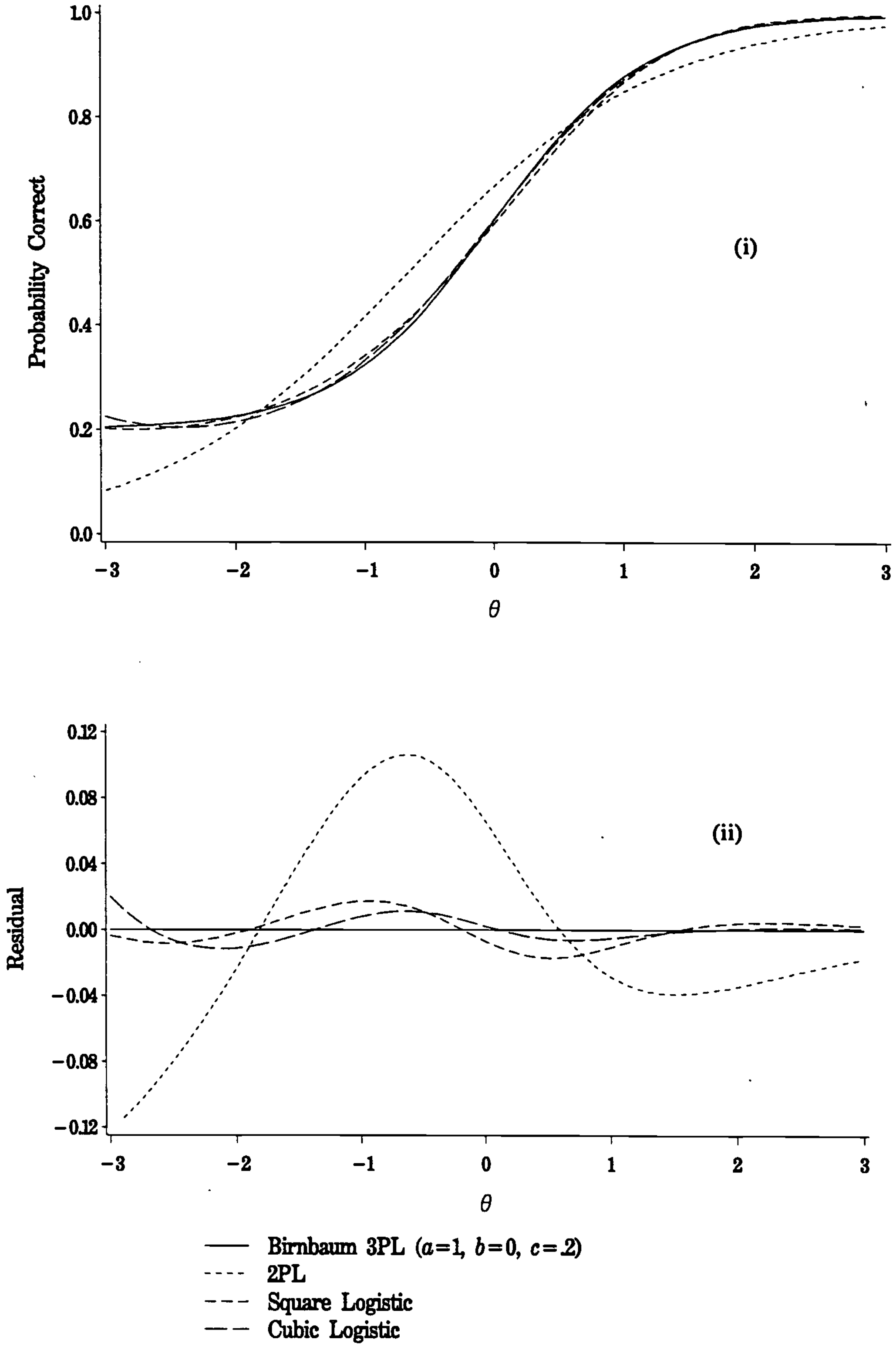

Probability Correct
1.0
0.8
0.6
0.4
0.2
0.0
(i)
−3
−2
−1
0
1
2
3
θ
Residual
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
(ii)
−3
−2
−1
0
1
2
3
θ
Birnbaum 3PL (a=1, b=0, c=.2)
2PL
Square Logistic
Cubic Logistic

Figure 2

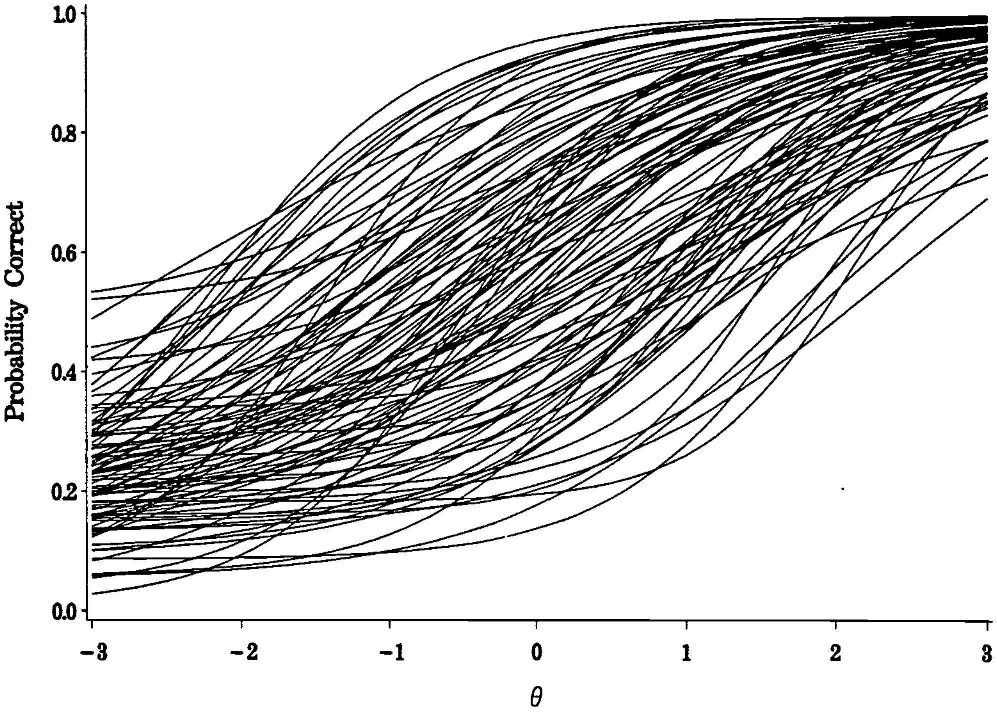


Figure 3

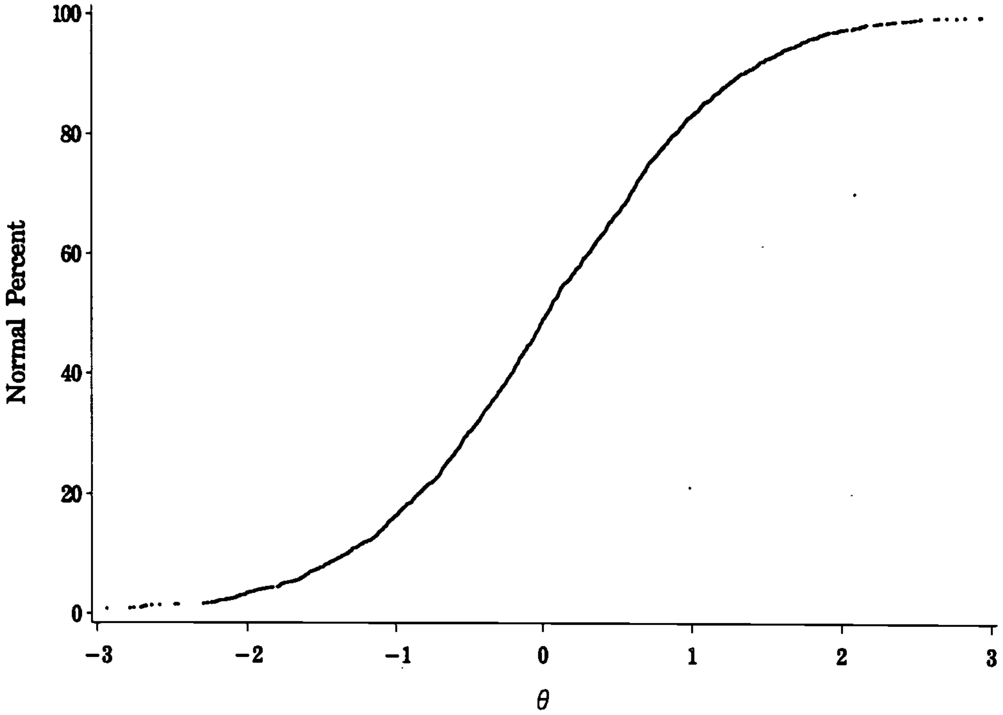


Figure 4

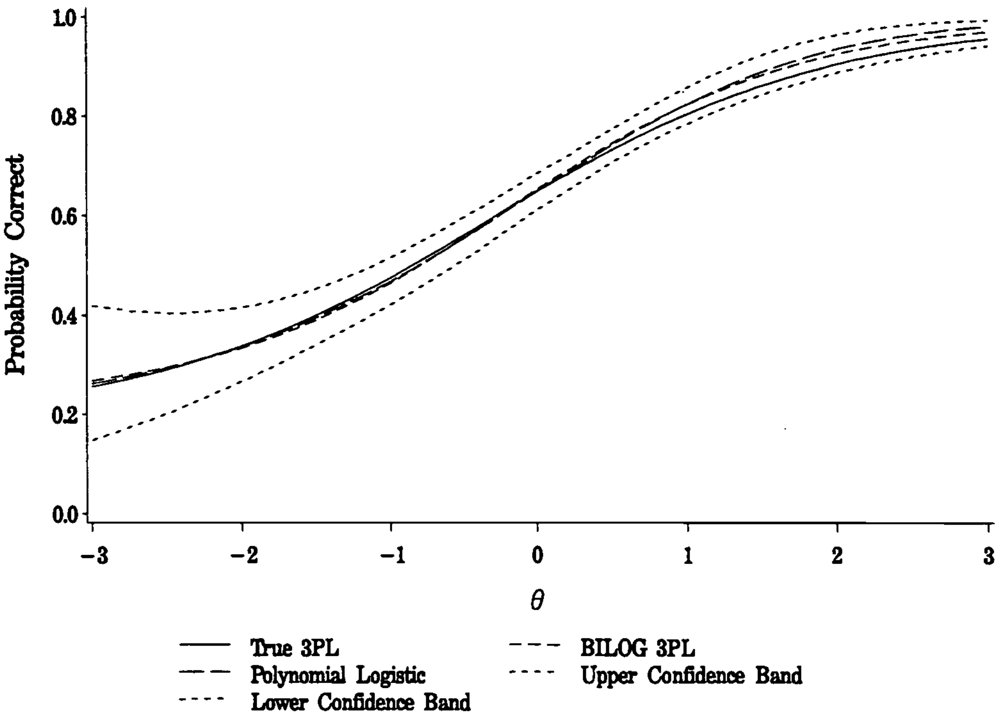

Probability Correct
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
θ
True 3PL
Polynomial Logistic
Lower Confidence Band
BILOG 3PL
Upper Confidence Band

Figure 5

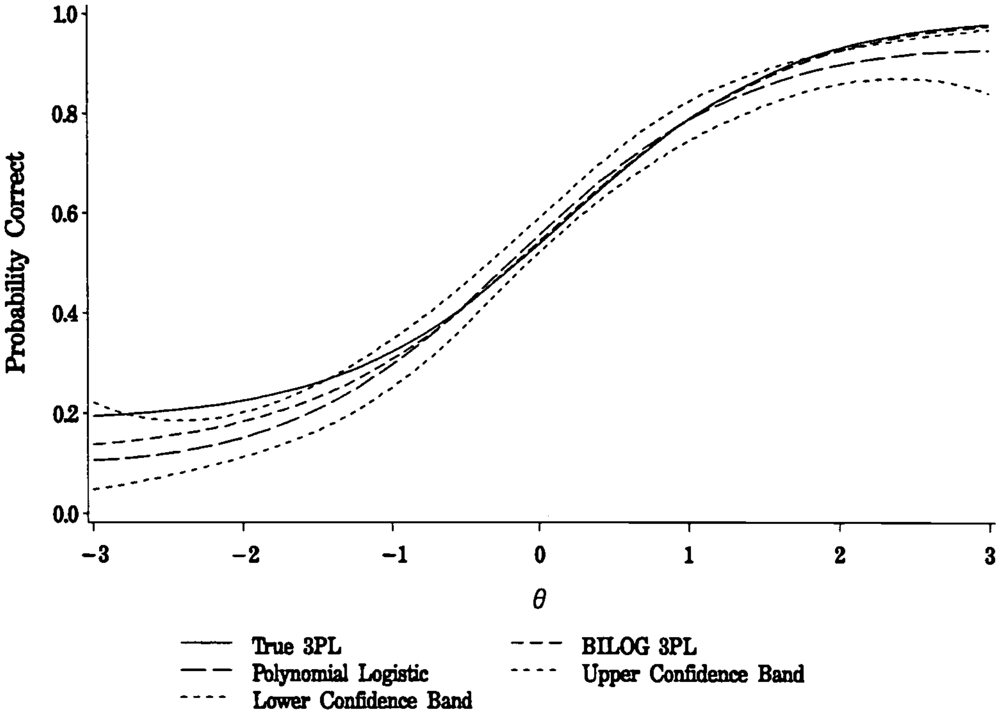

Probability Correct
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
θ
True 3PL
Polynomial Logistic
Lower Confidence Band
BILOG 3PL
Upper Confidence Band

Figure 6

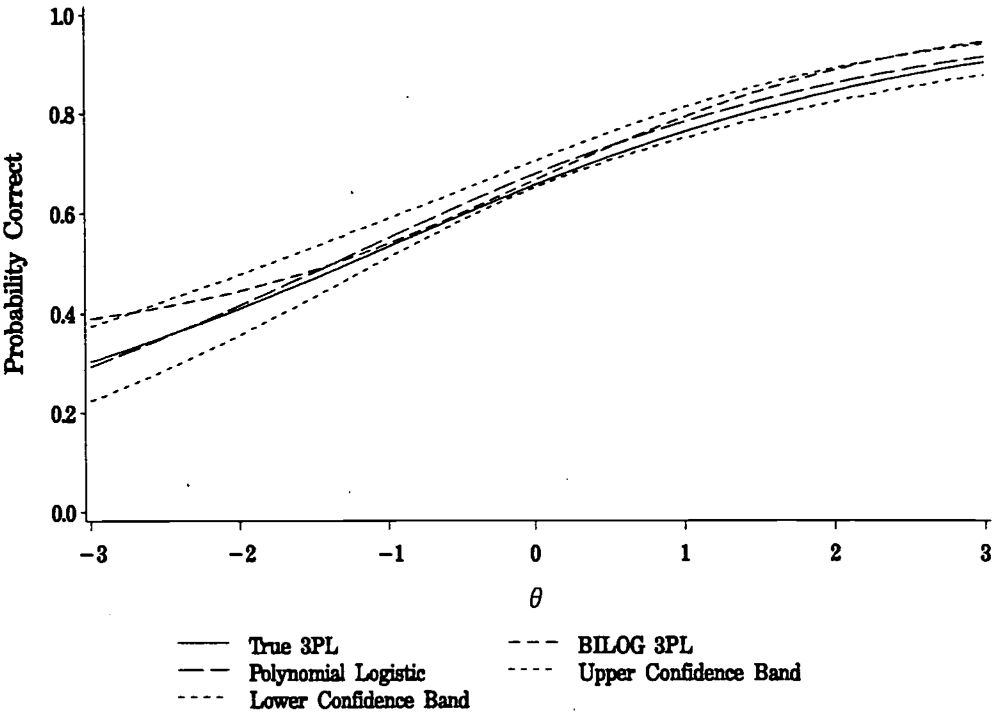

1.0
0.8
0.6
0.4
0.2
0.0
Probability Correct
−3
−2
−1
0
1
2
3
θ
True 3PL
Polynomial Logistic
Lower Confidence Band
BILOG 3PL
Upper Confidence Band

Figure 7

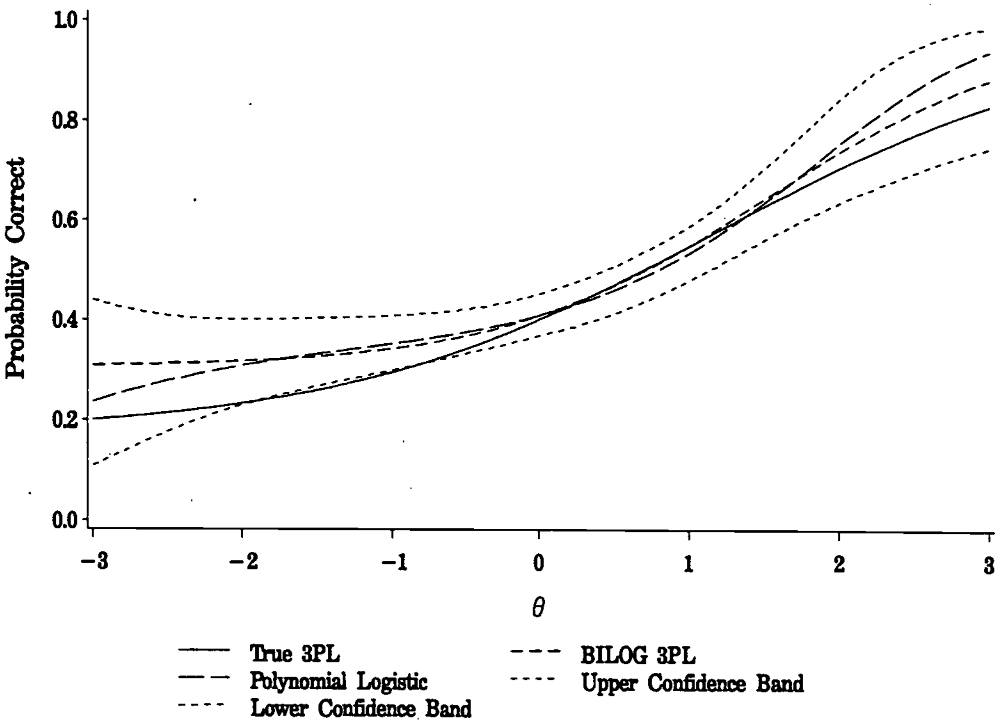
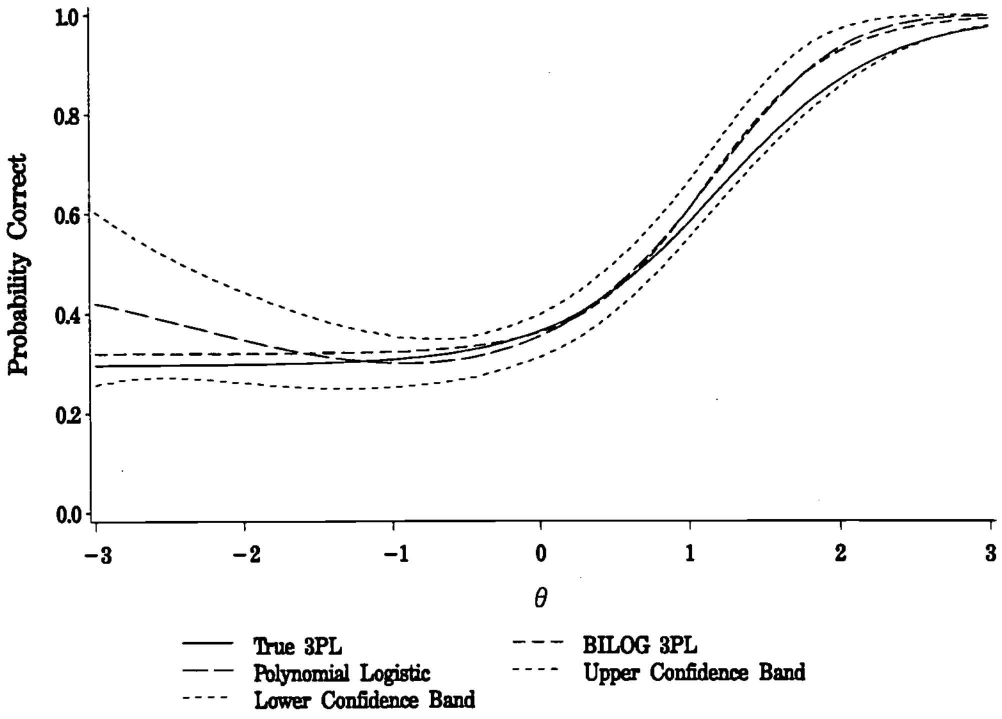

1.0
0.8
0.6
0.4
0.2
0.0
Probability Correct
−3
−2
−1
0
1
2
3
θ
True 3PL
BILOG 3PL
Polynomial Logistic
Upper Confidence Band
Lower Confidence Band

Figure 8



BEST COPY AVAILABLE

Figure 9

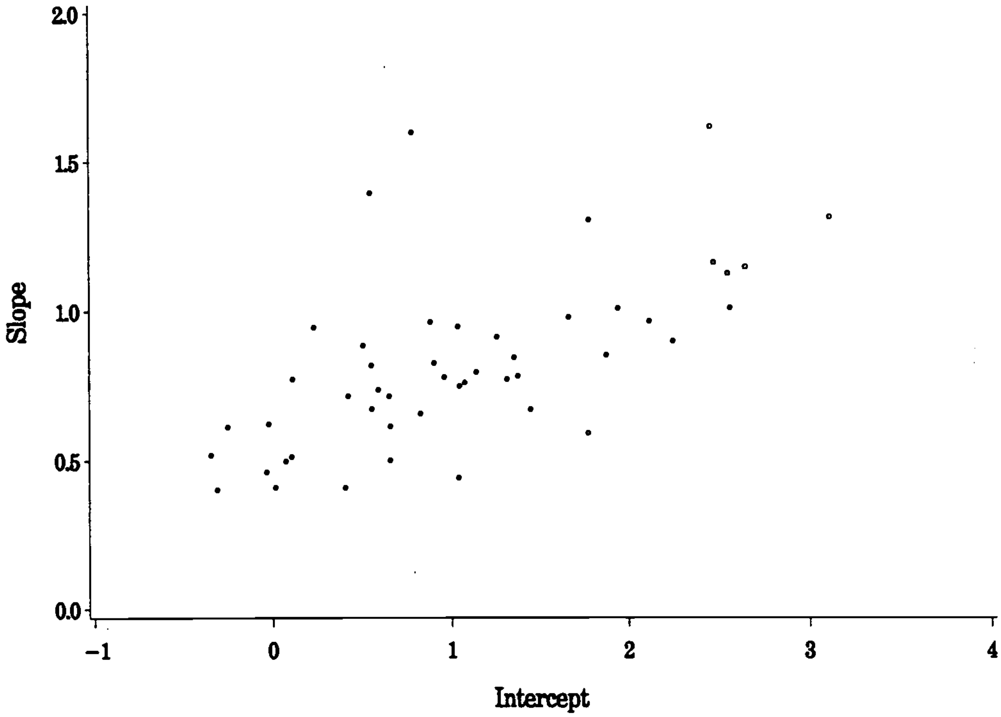


Figure 10

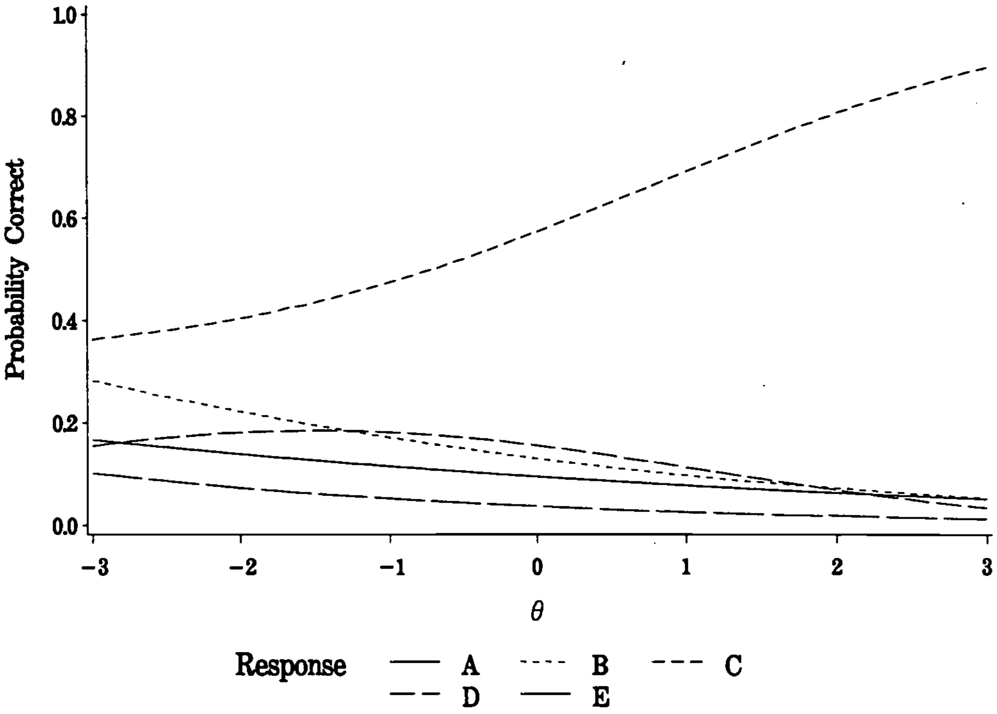

1.0
0.8
0.6
0.4
0.2
0.0
Probability Correct
−3
−2
−1
0
1
2
3
θ
Response
A
B
C
D
E

Figure 11

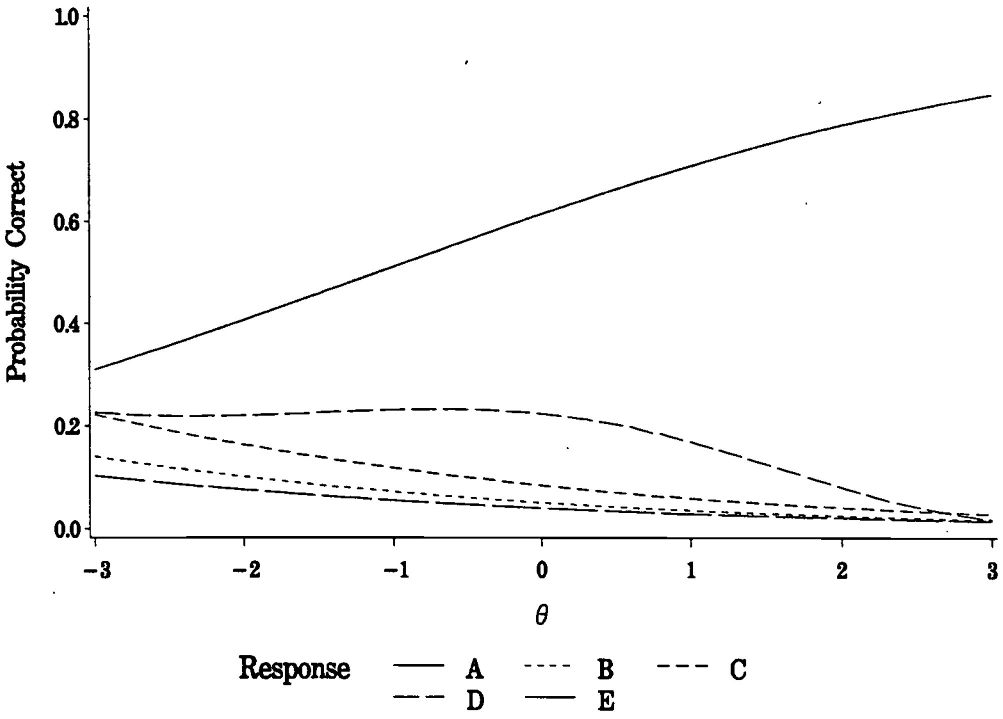

Probability Correct
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
θ
Response
A
B
C
D
E

Figure 12

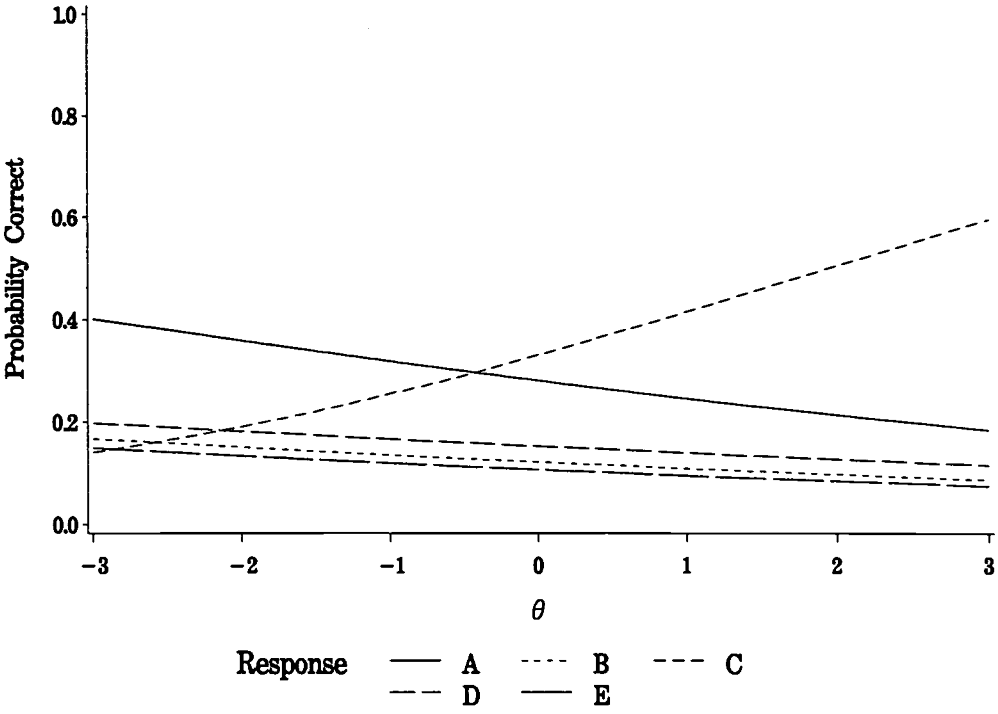

1.0
0.8
0.6
0.4
0.2
0.0
Probability Correct
−3
−2
−1
0
1
2
3
θ
Response
A
B
C
D
E

Figure 13

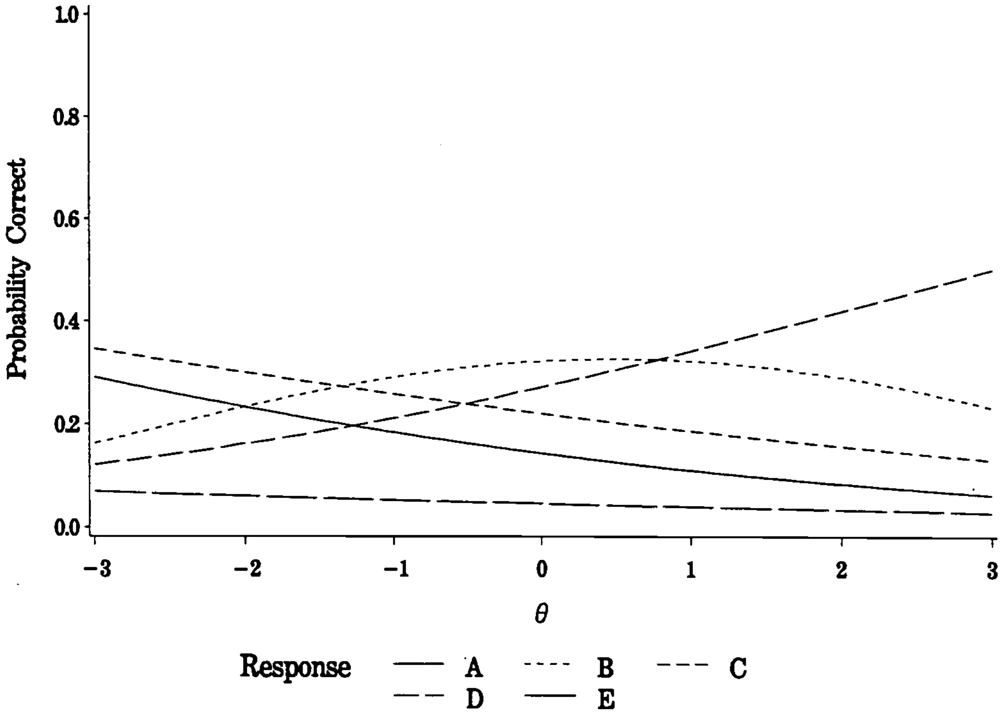
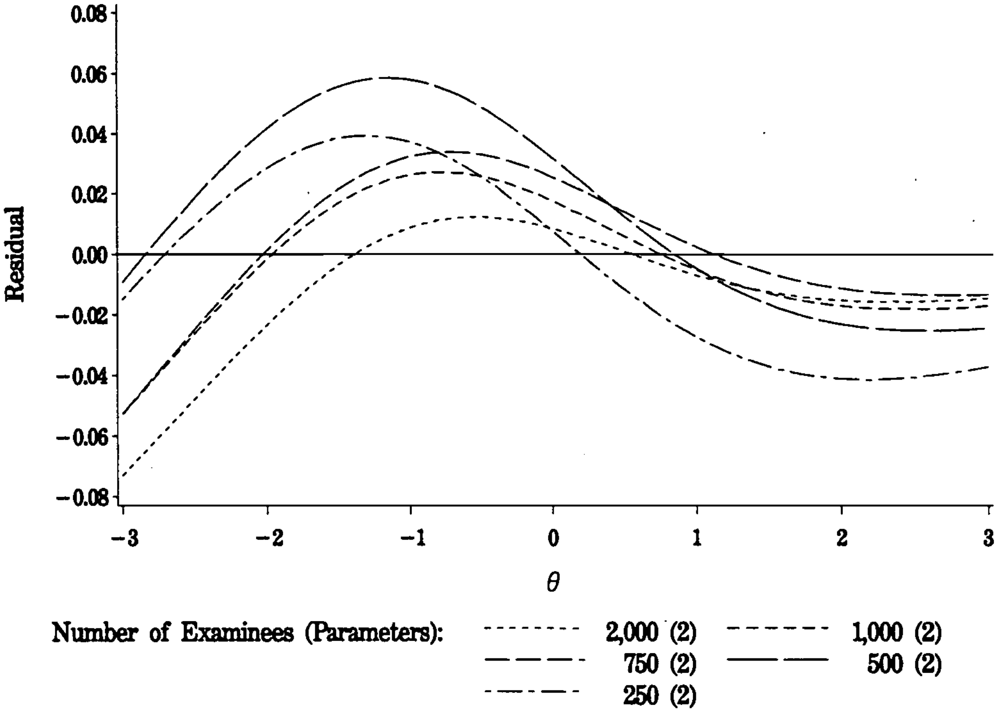

1.0
0.8
0.6
0.4
0.2
0.0
Probability Correct
−3
−2
−1
0
1
2
3
θ
Response
A
B
C
D
E

Figure 14



BEST COPY AVAILABLE

Figure 15

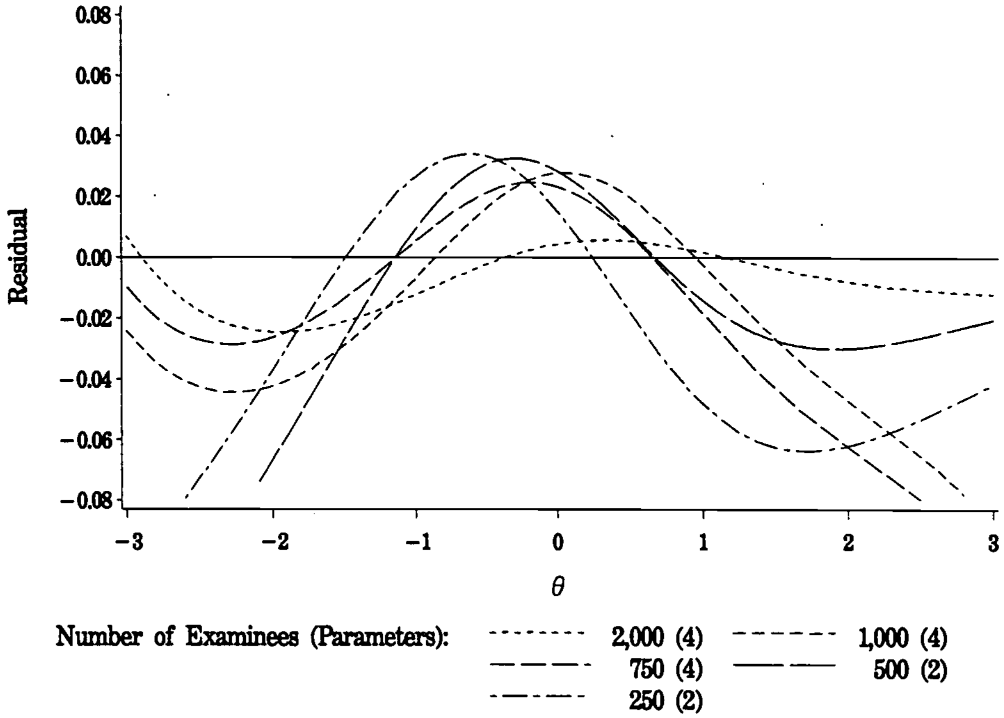


Figure 16

ERIC Full Text Provided by ERIC

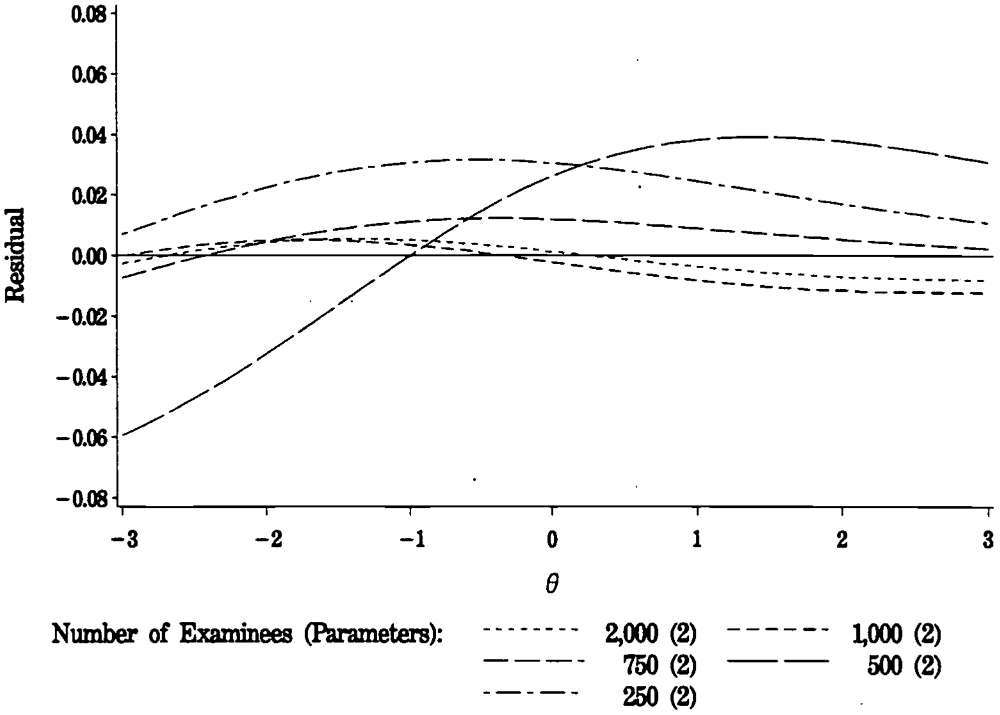

0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
Residual
−3
−2
−1
0
1
2
3
θ
Number of Examinees (Parameters):
2,000 (2)
1,000 (2)
750 (2)
500 (2)
250 (2)

Figure 17

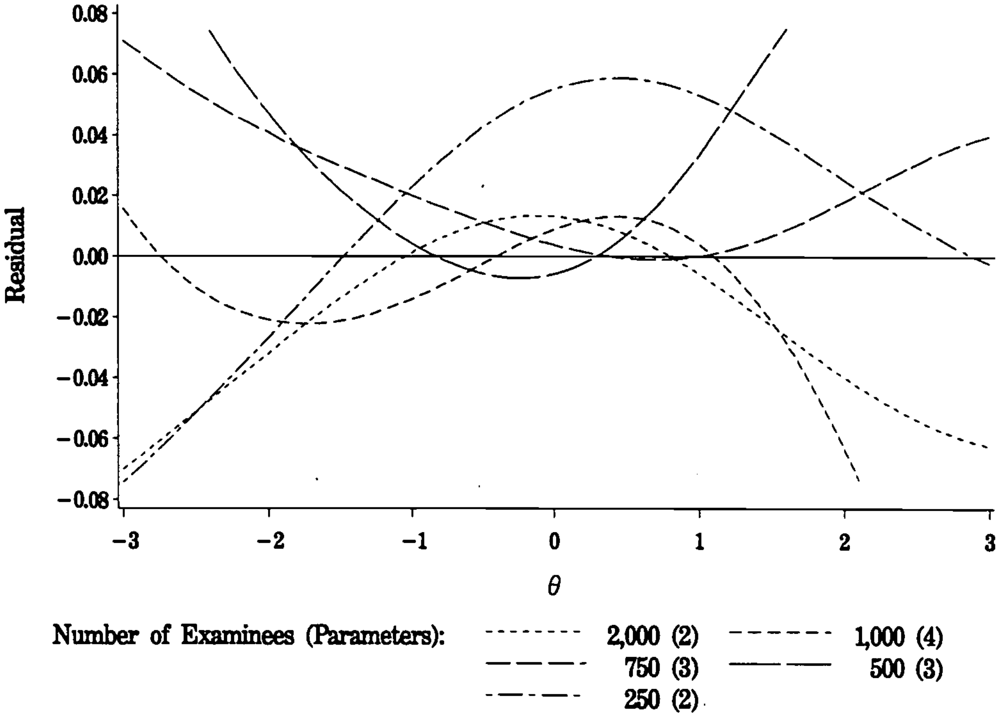

Residual
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−3
−2
−1
0
1
2
3
θ
Number of Examinees (Parameters):
2,000 (2)
1,000 (4)
750 (3)
500 (3)
250 (2)

Figure 18

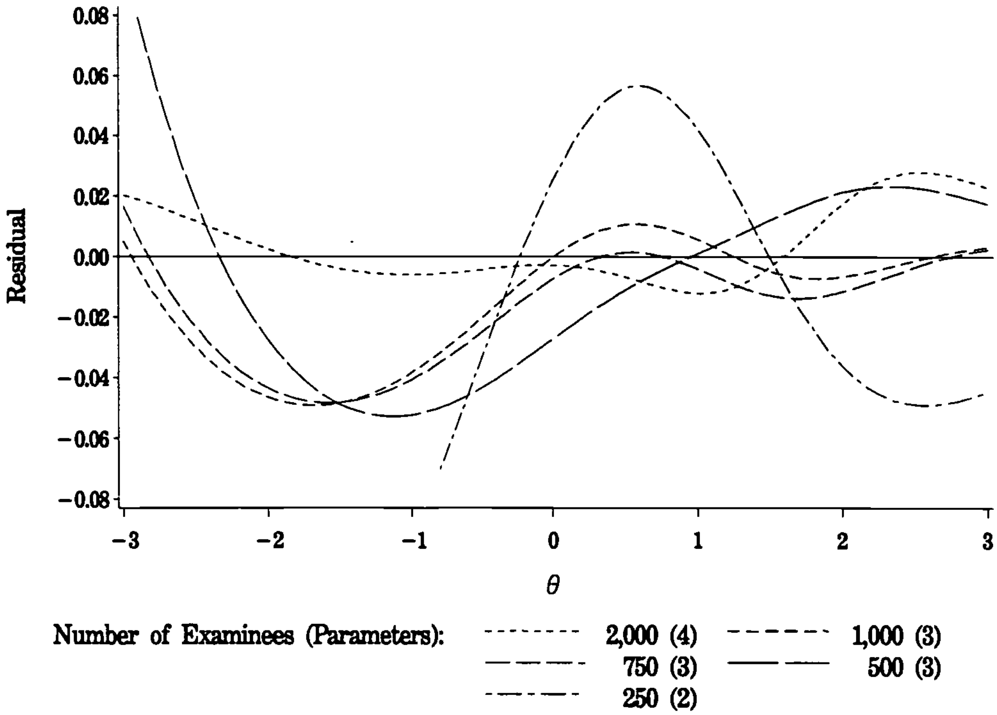

0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
Residual
−3
−2
−1
0
1
2
3
θ
Number of Examinees (Parameters):
2,000 (4)
1,000 (3)
750 (3)
500 (3)
250 (2)

Figure 19

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE
(Specific Document)

TMO27863

I. DOCUMENT IDENTIFICATION:

Title: Pretest Item Analyses Using Polynomial Logistic Regression: An Approach to Small Sample Calibration Problems Associated with Computerized Adaptive Testing	
Author(s): Liane N. Patsula and Peter J. Pashley	
Corporate Source: University of Massachusetts Law School Admission Council	Publication Date: April, 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/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. 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 the identified document, please CHECK ONE of the following options and sign the release below.

☒ ← **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Level 1

Sample sticker to be affixed to document → □

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical 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: Liane Patsula	Position: Graduate Student
Printed Name: Liane Patsula	Organization: University of Massachusetts
Address: 10 Pleasant Court Amherst, MA 01002	Telephone Number: (413)549-8644
	Date: October 27, 1997

ERIC
Full Text Provided by ERIC

OVER