AFHRL-TP-83-12

AIR FORCE

HUMAN RESOURCES

LABORATORY

ITEM RESPONSE THEORY: SOME STANDARD ERRORS

By

David Thissen

Department of Psychology
University of Kansas
Lawrence, Kansas 66045

Howard Wainer

T-254
Educational Testing Service
Princeton, New Jersey 08541

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

May 1983

Final Technical Paper

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

JANOS B. KOPLYAY
Contract Monitor

NANCY GUINN, Technical Director
Manpower and Personnel Division

J.P. AMOR, Lt Col, USAF
Chief, Manpower and Personnel Division

AFHRL Technical Paper 83-12 May 1983

ITEM RESPONSE THEORY: SOME STANDARD ERRORS

By

David Thissen

Department of Psychology
University of Kansas
Lawrence, Kansas 66045

Howard Wainer

T-254
Educational Testing Service
Princeton, New Jersey 08541

Reviewed and submitted for publication by

Janos B. Koplyay
Chief, Manpower and Force Management Systems Branch
Manpower and Personnel Division

This publication is primarily a working paper.
It is published solely to document work performed.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TP-83-12

2. GOVT ACCESSION NO.
AD-A129804

3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*
ITEM RESPONSE THEORY:
SOME STANDARD ERRORS

5. TYPE OF REPORT & PERIOD COVERED
Final

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR*(s)*
David Thissen
Howard Wainer

8. CONTRACT OR GRANT NUMBER*(s)*
F41689-81-C-0012

9. PERFORMING ORGANIZATION NAME AND ADDRESS
McFann-Gray & Associates, Inc.
5825 Callaghan Road, Suite 225
San Antonio, Texas 78228

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F
2313T137

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
May 1983

13. NUMBER OF PAGES
42

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*
Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

15. SECURITY CLASS *(of this report)*
Unclassified

15.*a.* DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of this abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*
computer adaptive testing
estimation
item parameter
item response theory
latent trait theory
maximum likelihood
standard errors
tailored testing
test construction
tests and measurement

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The mathematics required to calculate the asymptotic standard errors of the parameters of three commonly used logistic item response models is described and used to generate values for some common situations. It is shown that the maximum likelihood estimation of a lower asymptote reduces the accuracy of estimation of a location parameter. If one requires accurate estimates of location parameters (e.g., for purposes of test linking/equating or for computerized adaptive testing), the sample sizes required for acceptable accuracy may be so large as to make maximum likelihood estimation infeasible in most applications. It is suggested that other estimation methods be used if the three-parameter model is applied in these situations.

Preface

This research was completed under the Manpower and Force Management thrust and the Force Acquisition and Distribution subthrust. This is part of a continuing effort to improve assessment of personnel qualifications.

The authors wish to express their appreciation to Dr. Benjamin Fairbank, Jr., McFann-Gray & Associates, Inc., and to Dr. Janos Koplyay and To Dr. Malcolm Ree of the Manpower and Personnel Division of the Air Force Human Resources Laboratory for stimulating and encouraging their interest in this research. A special acknowledgement is due the Research Statistics Project of Educational Testing Service, Inc. for additional support of this research.

Table of Contents

List of Figures

Abstract

The mathematics required to calculate the asymptotic standard errors of the parameters of three commonly used logistic item response models is described and used to generate values for some common situations. It is shown that the maximum likelihood estimation of a lower asymptote reduces the accuracy of estimation of a location parameter. If one requires accurate estimates of location parameters (e.g., for purposes of test linking/equating or for computerized adaptive testing) the sample sizes required for acceptable accuracy may be so large as to make maximum likelihood estimation infeasible in most applications. It is suggested that other estimation methods be used if the three-parameter model is applied in these situations.

Introduction

The literature of Item Response Theory (IRT) (e.g., Lord, 1980) has not given extensive treatment to the standard errors of estimated item parameters for the commonly used models. A standard error is an indication of the precision with which a parameter is estimated. A useful rule is that the range of values from two standard errors below an estimate to two standard errors above the same estimate has associated with it a probability of .95 that the true population value is in the interval. Thus the smaller the standard errors are, the more precisely the parameters may be estimated. Few statistical estimation procedures yield exact results; the estimation of intervals is more common than point estimation. Even when point estimation is used, the point is usually understood to represent the middle of a distribution of possible values. An interval is often constructed by placing a confidence interval, based on standard errors, around a point (see, for example, Bradley, 1976.)

There have been few attempts at determining confidence intervals around the parameters associated with item response theory. That may be because the closed form formulae for those standard errors as a function of sample size and parameters are complicated and difficult to apply.

Item Response Theory (Lord, 1980, for a thorough introduction; Hambledon & Cook, 1977, for a survey) covers a wide range of aspects of test theory. The key facet for the purpose of this report is that each item in a multi-item multiple choice ability test can be characterized by specifying the probability that an examinee of any given ability will answer the item correctly. Furthermore, the equation which relates the probability of a correct response to the ability of the examinee is assumed to be logistic. The following equation is widely used to express the probability of a correct response to an individual item

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

where θ = ability of the examinee, a = the discriminating power of the item (higher values of a indicate that an item is more effective in discriminating between more and less able examinees), b = the difficulty of the item (high values indicate more difficult items), and c = the probability that an examinee with no knowledge of the subject will answer correctly.

Two commonly used simplifications of the model exist; the first assumes that the c parameters are all known (usually zero), and the second assumes not only that the c parameters are known, but also that the a parameters (slopes) are all equal. The three variations of the model are usually referred to as the logistic item response models.

In the simplest case, the computing of asymptotic standard errors of the parameters of logistic item response models estimated through maximum likelihood requires numerical integration of a complicated composite function. Nevertheless, those standard errors may be directly, if tediously, computed for any set of parameters and sample size. No data are required. It is assumed that the models are appropriate for the data, and the data fall within the ranges the model specifies. Since neither assumption is likely to be exactly true in practice, the standard errors obtained by this method represent lower limits for actual standard errors. Nevertheless, it is useful to consider the minimum values obtainable for the standard errors for the parameters of various models. If a test developer requires certain precision in parameter estimates, the methods and tables of this report may be used to select sample sizes large enough to allow such precision.

The purpose of this report is to demonstrate methods for calculating the standard errors, to provide representative values of the errors and to consider the implications which the magnitudes of the errors may have for operational testing situations.

Estimation and Standard Errors

Consider the development of an item pool for computer adaptive or paper-and-pencil testing. The simplest situation in which there is a problem of item parameter estimation arises when one item is to be added to the pool. For instance, a sample of examinees may be tested with an established test, plus one new item. The parameters of the items in the established test are known. In order to determine the parameters of the new item, one may use the old items to estimate the abilities of the examinees, and then estimate the parameters of the new item using the ability estimates of the examinees.

The ability of the examinees can be estimated to any degree of accuracy required, by adding old items to the test. The parameters of the single new item may be estimated by nonlinear regression of the right/wrong responses to the new item on the ability of the examinees. Ability must always b estimated in practice, but since it may be estimated with any precisi required, given a large enough item pool, it will hereafter be assumed th the abilities are known fixed values, thus making the calibration of a sing item a nonlinear regression problem.

A nonlinear regression problem such as that described in the previous paragraph is easily solved by maximum likelihood methods. If the item response model is a function of θ_i (the ability of person i) and set of item parameters $\underline{\xi}$, which gives the probability of a correct response to the new item, r = 1,

$$P_i = P (r_i = 1 | \theta_i, \underline{\xi}) \tag{1}$$

then the likelihood of the observed responses for N independent examinees is (with the subscript i omitted leaving $P = P_i$ and $r = r_i$)

$$L = \prod_{i=1}^{N} P^{r} (1 - P)^{(1 - r)} \tag{2}$$

and the loglikelihood is

$$\ell = \sum_{i=1}^{N} r \log (P) + (1-r) \log (1-P) \tag{3}$$

and the maximum likelihood estimates of each parameter in the set $\underline{\xi}$ are located where the partial derivatives

$$\frac{\partial \ell}{\partial \underline{\xi}} = \sum_{i=1}^{N} \frac{r}{P} \frac{\partial P}{\partial \underline{\xi}} - \frac{1-r}{(1-P)} \frac{\partial P}{\partial \underline{\xi}} \tag{4}$$

are zero.

The second partial derivatives of the loglikelihood have the form

$$\frac{\partial^2 \ell}{\partial \xi_s \partial \xi_t} = \sum_{i=1}^{N} \left\{ \frac{r}{P} \frac{\partial^2 P}{\partial \xi_s \partial \xi_t} + \frac{\partial P}{\partial \xi_s} \frac{\partial P}{\partial \xi_t} \frac{-r}{P^2} \right.$$

$$\left. - \frac{(1-r)}{(1-P)} \frac{\partial^2 P}{\partial \xi_s \partial \xi_t} - \frac{\partial P}{\partial \xi_s} \frac{\partial P}{\partial \xi_t} \frac{(1-r)}{(1-P)^2} \right\} \tag{5}$$

for any parameter ξ_s and ξ_t. In general, the inverse of the negative expected value of the matrix of second derivatives of a loglikelihood is the asymptotic variance-covariance matrix of the estimates (Kendall and Stuart, 1960, pp. 54-55). The density of θ is taken to be $\phi(\theta)$, and substituting P for r results in the expectation of (5). Then, through simplification and integration, the equation becomes

$$-E \frac{\partial^2 \ell}{\partial \xi_s \partial \xi_t} = N \int_{-\infty}^{\infty} \left\{ \frac{1}{P} \frac{\partial P}{\partial \xi_s} \frac{\partial P}{\partial \xi_t} + \frac{1}{(1-P)} \frac{\partial P}{\partial \xi_s} \frac{\partial P}{\partial \xi_t} \right\} \phi(\theta)\, d\theta \tag{6}$$

Equations similar to (6) for the three-parameter model in the finite sample case are given by Lord (1980, p. 191). Evaluation of (6) requires only the derivatives of P with respect to its parameters, and the specification of $\phi(\theta)$. Hereafter, $\phi(\theta)$ will be taken to be Gaussian with mean zero and variance one.

The derivatives of P with respect to its parameters are extremely simple for commonly used logistic item response models. These models are based on the logistic function, which, in simplest form, is

$$P^* = 1/(1 + \exp -a(\theta - b)) \qquad (7)$$

$$Q^* = 1 - P^*. \qquad (8)$$

The parameter a is a function of the slope of the resulting item characteristic curve, and b is its location. For the so-called three-parameter logistic model,

$$P = P(r = 1|\theta) = c + (1 - c)P^* \qquad (9)$$

which is equivalent to the equation given in the introduction, except that the earlier equation contained the factor of 1.7 which changes the units to approximate those of a normal distribution with mean zero and standard deviation of 1.

The third parameter, c, is a (possibly) nonzero lower asymptote. In this model,

$$\begin{aligned} \frac{\partial P}{\partial a} &= (1-c)\ P^*Q^*\ (\theta-b) \\ \frac{\partial P}{\partial b} &= (1-c)\ P^*Q^*\ (-a) \\ \frac{\partial P}{\partial c} &= 1-P^* \end{aligned} \qquad (10)$$

For the two parameter model, (11)

$$P = P(r = 1|\theta) = P^*$$

and

$$\frac{\partial P}{\partial a} = P^*Q^*\ (\theta-b) \qquad (12)$$

and

$$\frac{\partial P}{\partial b} = P^{*} Q^{*} (-a) \qquad (13)$$

The one-parameter model is the same as the two-parameter model with the value of a fixed, so the only derivative required is that for b, the location parameter, and it is identical to that for the two-parameter version.

For any of these models, specified values of the parameters, and a given sample size, these derivatives may be substituted in (6) to give a k x k (for the k-parameter model) information matrix. That matrix may then be inverted to give the variance-covariance matrix of the parameters. The square roots of the diagonal elements of the inverse are the asymptotic standard errors of the parameters. In Appendix II are tables of those standard errors for the various models as functions of the parameters. The parameter values are chosen to represent the range encountered in testing situations.

Approach

The equations for the maximum likelihood estimates of the parameters (Lord, 1980), were used as a basis to derive equations for the matrix of second partial derivatives. The matrices were inverted to obtain standard errors under plausible assumptions (e.g., normality) and for a range of commonly encountered parameter values. The errors were studied to derive inferences about the sample sizes needed for various levels of accuracy.

Results

The results obtained through the use of equation (6) are shown for various situations in Tables 1-15 (Appendix II). These tables are presented for reference purposes, since they span most situations found in practice. One can interpolate between the values for intermediate ones. The reader is invited to peruse the tables and draw inferences; certain salient aspects of the tables are discussed below.

Before going into the details of Tables 1-15, a few remarks are in order. First, all standard errors considered here are proportional to $1 / \sqrt{N}$. Thus, to obtain the asymptotic standard error of a parameter for any particular test administration, one must divide the number in the appropriate table by $\sqrt{N}$. For example, if there are 100 examinees, divide the tabled numbers by 10; 2,500 examinees, divide by 50, etc. The numbers in the tables are asymptotic, but samples of 100 or more should be reasonably well represented by these asymptotic values.

A second point of note is that the slope parameter (a) used here is for a logistic function. If the user is interested in transforming this to be comparable to a normal ogive, divide the expressed values of the slope by 1.7. The transformation is necessary because the logistic curve approximates the normal curve's shape, but not the size of its standard deviation. The factor of 1.7 makes both the shape and size comparable. Thus the slope values given in the tables correspond as shown below:

Logistic Slope	Normal Ogive Slope
.25	.15
.50	.29
.75	.44
1.00	.59
1.50	.88
2.00	1.18
3.00	1.76

The tables shown can be used to aid in the determination of the sample size required to yield desired accuracy. For example, suppose one is equating tests and needs accurate estimates of item location (difficulty), and the decision is made that one decimal place of accuracy is sufficient. This implies that the standard error of location should be of the order of .05. Good ability test items have slopes in the 1 to 1.5 range, and so using Table 1 one can see that for items whose location is in the range -2 to +2 one will need sample sizes of 2500 (calculated as $(2.5 / .05)^2$) for a worst case situation. This is for the one-parameter model. For the two-parameter model a sample of 7500 is required (from Table 2: $(4.33 / .05)^2$), and for the three parameter model one needs 67,000 (from Table 4: $(12.94 / .05)^2$).

Such calculations are easily carried out and point toward sample sizes required a priori for minimally acceptable accuracy.

To better understand these results, compare the standard errors for the three models for a common sample (N = 2500). Shown in Figure 1 are the standard errors for the three models when slope is 1.5 (a representative value for most serious testing applications). Further, assume that the lower asymptote is zero.

The overwhelming first impression given by examination of such figures is that the use of an unrestricted maximum likelihood estimation for the three-parameter model either yields results too inexact to be of much practical use, or requires samples of such enormous size as to make them prohibitively expensive. This problem arises for items that are easier than average. This effect is a result of the huge covariance (computable from equation 2) between location and lower asymptote. When an item is relatively easy (b = - 1), there are few observations available to estimate the lower asymptote thus making its standard error very large. The large covariance between lower asymptote and location then causes this uncertainty to be shared with the estimate of location. With more difficult items the effect is lessened somewhat. The two-parameter model has problems as well, but they are far less severe.

If plots similar to that in Figure 1 were constructed but the slope and the lower asymptote varied:

1) the same general structure would continue to hold,
2) as slopes became more gradual the size of the standard errors would get larger, and
3) as the lower asymptote rose, the standard errors of the three-parameter model would rise apace.

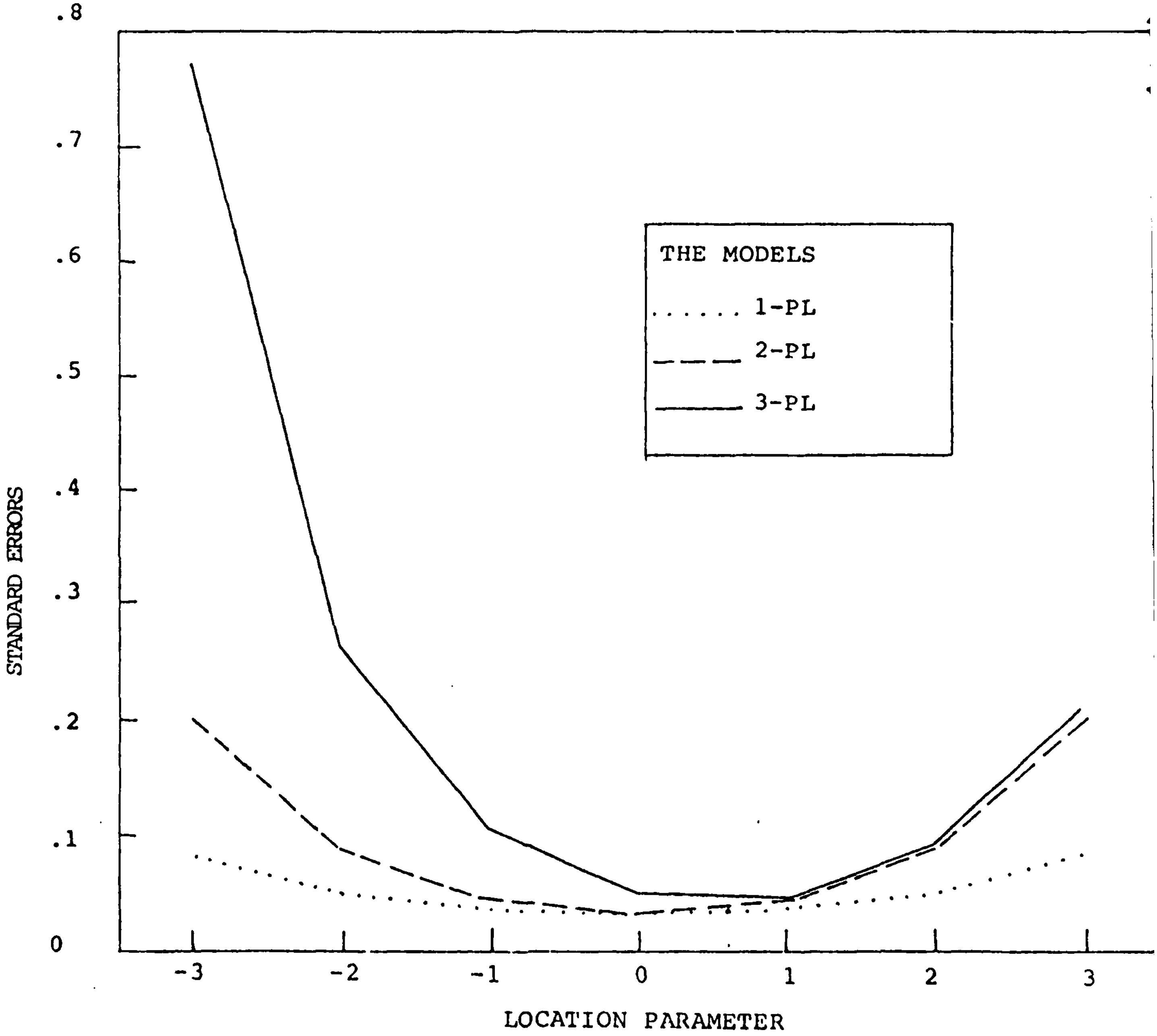

Figure 1. Standard errors of the location parameters for three logistic models, slope = 1.5, lower asymptote = 0, n = 2500.

To better understand the relationship that the slope and the location parameter for the three models have with standard error, examine some three dimensional plots. Shown in Figures 2, 3, and 4 are the natural logs of the standard errors (taken from Tables 1, 2, and 4) of location for the three models. To facilitate comparisons among the models, the figures illustrate the case where the lower asymptote is zero. When it is not, the one- and two-parameter models do not apply, and the standard errors for the three-parameter model, already large, get larger.

Consider again the question of how large a sample is needed in order to achieve one decimal place of accuracy for each of the three models as computed by maximum likelihood estimation. While the answer to this question is contained in Tables 1, 2, and 4, it is seen more easily in a graph. Consider Figure 5, in which are shown the standard errors for a rather good item (slope = 1.5) which, in this instance, is assumed to have no guessing, or, equivalently, to have a probability of zero associated with a correct answer when attempted by an examinee of very low ability. Shown are the *standard errors for a sample of 10,000.*

This is about as large a sample as is plausible with existing software for the three-parameter model. The one-parameter model provides adequate accuracy of item parameter estimation throughout the range of the test from -3 θ to +3 θ. The two-parameter model is adequate in the middle, but slips at the extremes to standard errors of the order of 0.1. The three-parameter model is dominated overall by the other two models, and is adequate only in the middle of the test. The estimates are hopeless for easy items. This is a benign finding, for it is on items such as this that the guessing parameter is not required. This provides hope that a hybrid model that computes a lower asymptote for difficult items and does not for easy ones may be useful when the assumption of no guessing seems implausible. It is disquieting to find inadequate estimates with even this size sample. Note that even with 100,000 observations all parameters are still not estimated to within the broad levels of acceptability proposed. Further, note the size samples that would be required for any model to provide estimates of difficulty acceptable to the two to four decimal places often reported.

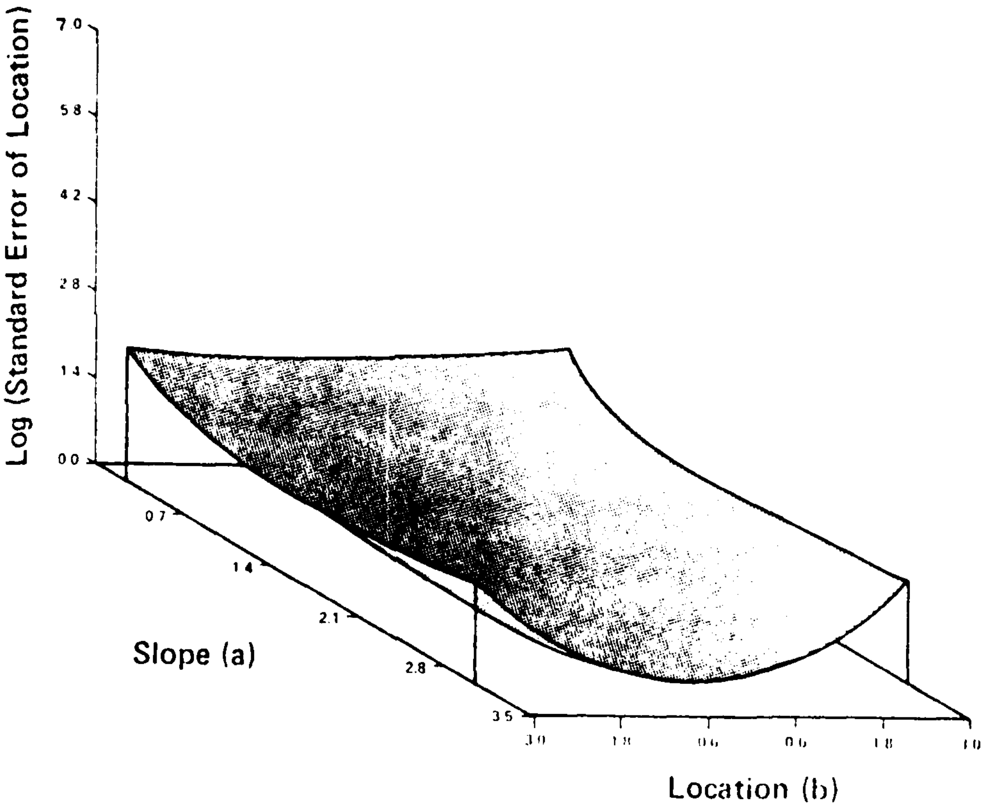

Figure 2. Natural log of standard errors for one-parameter logistic model.

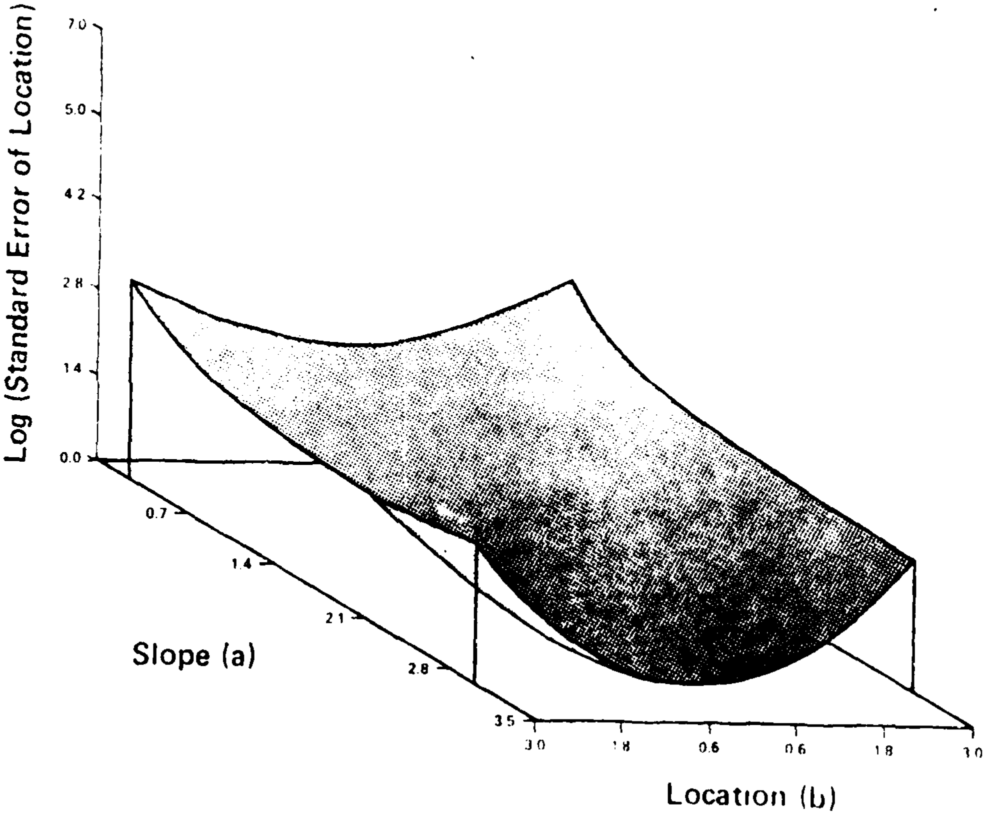

Figure 3. Natural log of standard errors for two-parameter logistic model.

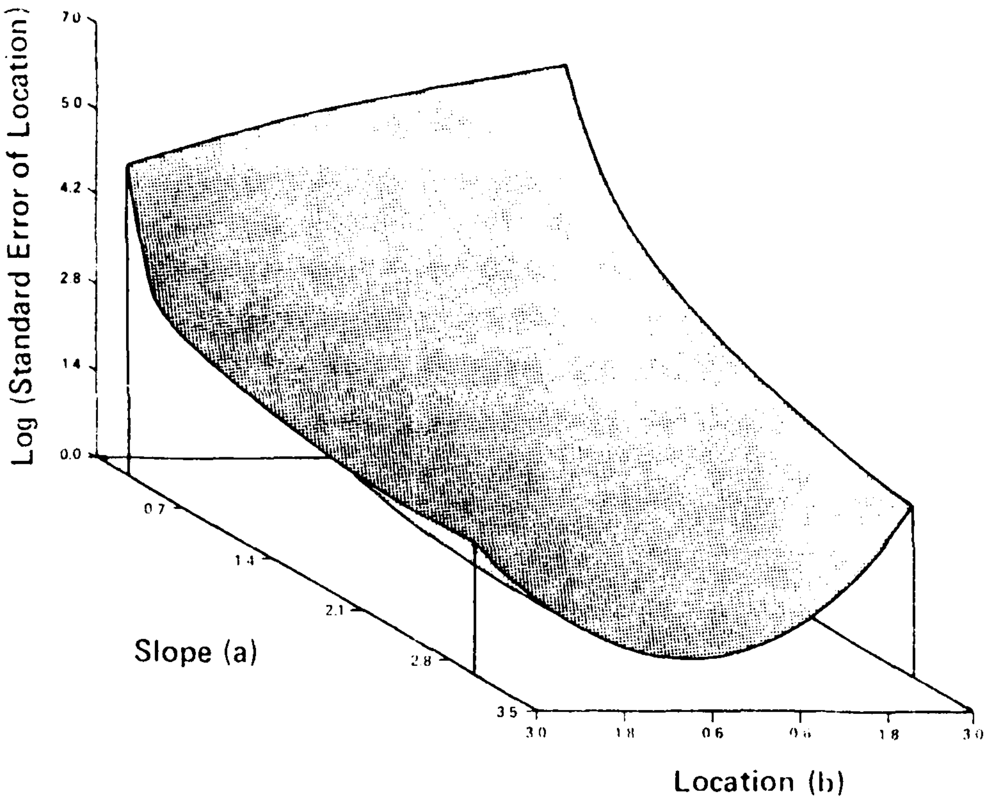

Figure 4. Natural log of standard errors for three-parameter logistic model.

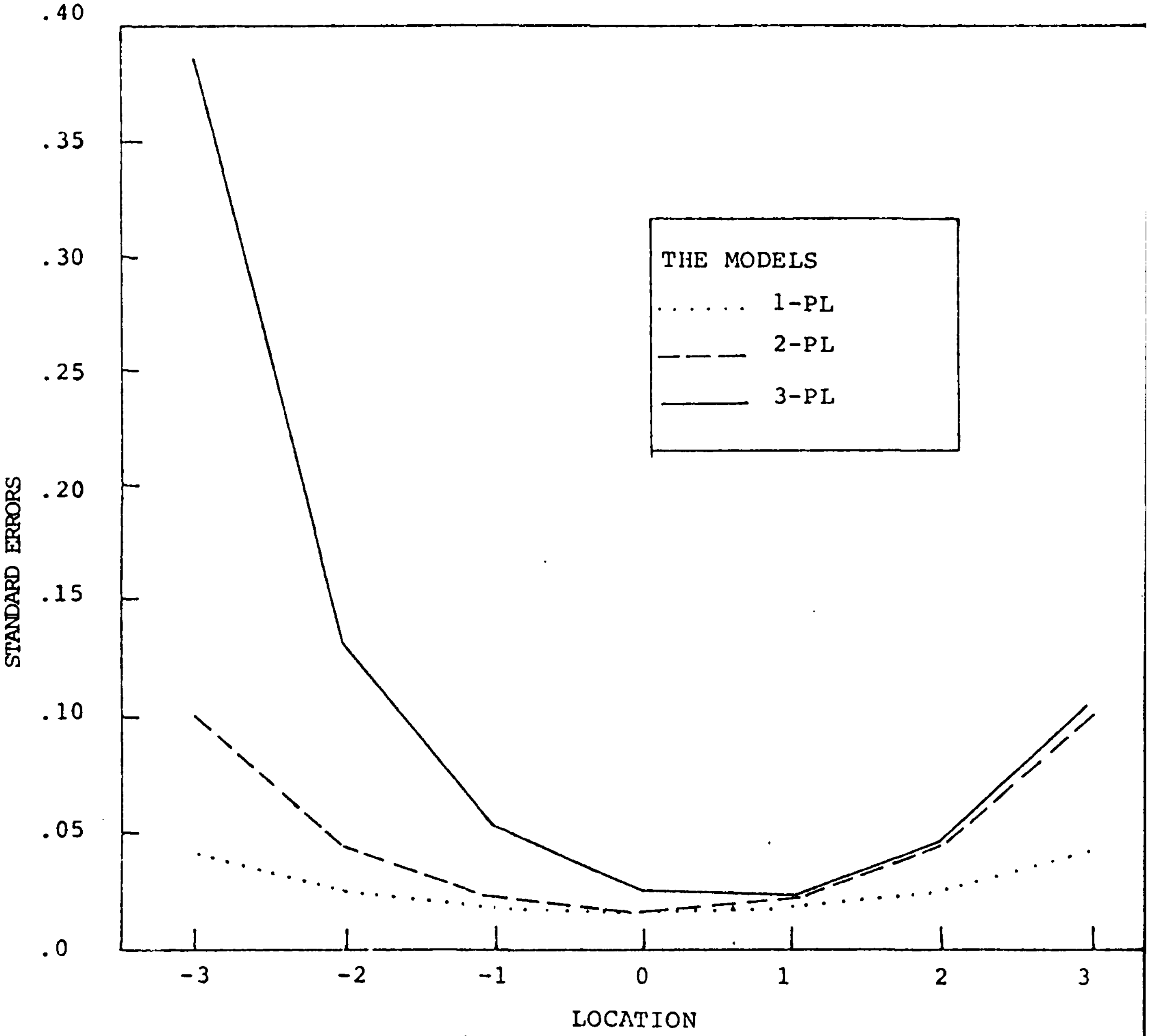

Figure 5. Standard errors of the location parameters for three logistic models, slope 1.5, lower asymptote = 0, n = 10,000.

Discussion

The lesson to be learned from this is that an investigator should try to fit the simpler models first, and only if they are found to be inadequate to move on (with the caution made appropriate by consideration of errors) to the more complex ones. If the more complex models are required, it would seem that a method of parameter estimation other than unrestricted maximum likelihood ought to be used.

Why does the three-parameter model do so badly at the lower (easy) end? The explanation is straightforward intuitively. The standard error of a parameter is a function of the square root of the number of observations available for the parameter that is to be estimated. There are very few observations low enough to provide information at the lower tail of an easy item. This means that the lower asymptote will be poorly estimated. The error of estimation of the lower asymptote is strongly related to the error of estimation of the location parameter (as the asymptote rises, the difficulty shifts to the right), thus uncertainty in the asymptote reflects itself in the uncertainty in the location estimate. The slope estimate is also affected, but not as severely.

The standard errors at location 0 are the same for both the one- and two-parameter models. This seems odd at first since the smaller parameterization ought to leave more observations to estimate the location. After some thought the reason for this becomes clear — the slope is irrelevant to the location at this point since no matter what the slope, it still must go through the point 0. A more rigorous proof of this is found Appendix I.

These findings suggest two lines of consideration. The first concerns the conditions under which the standard errors of the parameters are important; and the second is, under those conditions, what strategy should be followed to keep them as small as possible.

The enormous size of the standard errors for the three-parameter model when estimated by unrestricted maximum likelihood is not always a problem; it all depends upon whether or not the individual parameter estimates are to be used. One reason for the high standard errors is the covariation among the parameters. It is possible that such poorly estimated parameters will still yield an item characteristic curve that fits the observed data rather well (i.e., the Hessian is nearly singular). This is a well-known problem in other contexts. For example, suppose we have a set of data that are well described by the function,

$$y = a + bx + cx^2 + dx^3 + \ldots = F(x)$$

and we are trying to estimate the parameters a, b, c, d, ... Further, these same data are also well described by a Fourier series (with a different set of parameters), and a set of spline functions. Each of these formulations describes the data, although some of them yield parameter estimates that are highly unstable (i.e., that may have multicolinearities that yield tradeoff effects). Nevertheless, if the data are well described by the function, certain characteristics of the function (e.g., the value of a derivative of F at some point x_0) can be estimated well. Consequently, it is not necessarily true that because the parameters of a fitted function are not well estimated that certain other characteristics of the function are unstable. In the case of the three-parameter logistic model it seems that the parameters by themselves may not be useful, but certain characteristics (such as the information function or estimates of ability) may still be relatively stable if the model fits reasonably well.

An examination of the tabled results suggests that one cannot use unrestricted maximum likelihood in the estimation of the three-parameter model with samples of less than gigantic proportions if one is using the parameters of the model separately for some purpose. If, however, one is using those parameters in conjunction with one another to consider the item characteristic curve (ICC) or an item information curve, as for test construction, the model can be useful. The rest of this discussion deals with the situation where the parameters are needed separately - e.g., test equating/linking or computerized adaptive testing.

This paper has been concerned solely with the problems of the standard errors of the parameters of three logistic item response models when the model fits, but has not considered bias when it does not. If the situation is such that the data have zero as a lower a.ymptote, uniform slopes, and the ICCs differ only in location, a one-parameter model is the most suitable. If the items are reasonably discriminating (a = 1), maximum likelihood methods can estimate difficulties rather well (standard error less than 0.05) with 2,500 observations. If the user can tolerate standard errors of 0.1 or so, a sample of 500 is sufficient. With samples of less than 500, even with the one-parameter model, the inferences drawn about item difficulties based upon anything more than their integer value may be misleading due to the size of the standard errors. Certainly the practice of presenting item difficulties to more than one decimal place seems only very rarely to be justified. If the items are less discriminating (a = 0.5), one requires about 1,000 observations to obtain standard errors of the order of 0.1. This is under the best of circumstances. If slopes are not homogeneous, the standard errors get worse, although not dramatically so. As proved in Appendix I, the standard errors of the location parameter for items near the center of the ability distribution for the two-parameter model are not very much worse than for the one-parameter model. As the items get more extreme, the standard errors increase. Note (Table 3) that the standard errors of the slopes are not too bad, usually well within the bounds of acceptability when the sample size is sufficient to give acceptable location estimates.

If the lower asymptotes cannot be thought of as homogeneous (i.e., largely invariant from item to item), serious problems arise. Then acceptably small standard errors of the location parameter are unobtainable unless the samples are very large indeed.

As shown in Figure 5, 10,000 observations were barely adequate, and even with that many observations, estimates of the rather easy items are poor (if

$b = -2$, standard error $= .3$). If accuracy at that end is required, the sample must be near 100,000.

At the current state of program development there is no computer program available that will fit the three-parameter model by the method of unrestricted maximum likelihood with 100,000 examinees.

From the above, a reasonable strategy for fitting item response models to data (when accurate individual parameters are required) is as follows, assuming the use of unrestricted maximum likelihood:

(1) Try the one-parameter logistic model first. If it fits, stop. If it does not fit, examine those items that fit poorly with appropriate diagnostic statistics and plots to understand why the lack of fit occurred.

(2) If only a few items (a small proportion) do not fit, and they do not form a coherent grouping in terms of the subject matter of the test, consider omitting them from the test and continuing. If this is possible, s'op. An examinee pool of 500 to 1,000 will suffice to give one decimal accuracy to parameter estimates.

(3) If so many items do not fit that they can not be omitted, the diagnostics should indicate whether the problem is one of slopes or of lower asymptotes. If there is strong reason to believe that the lack of fit is caused by heterogeneous slopes, use the two-parameter formulation and increase the sample size. If it is possible to sample individuals at the extremes of the ability distribution more heavily, this will aid in the accuracy of the estimation.

(4) If the lack of fit is caused by both heterogeneous slopes and nonzero lower asymptotes, test to see if a uniform nonzero lower asymptote will correct the lack of fit. If so, specify it and apply procedure in (3).

(5) If it is necessary to use the three-parameter model with estimated rather than assigned asymptotes, consider changing the instructions to limit guessing, modifying the test, using a faster computer, and/or most importantly, using a different estimation scheme.

Conclusions

The problems associated with trying to get accurate unrestricted maximum likelihood parameter estimates with the three-parameter model seem to be enormous. They are so serious that almost any other strategy seems preferable. If one needs accurate estimates of location parameters, and variations of the one- or two-parameter logistic model do not fit successfully, the accurate estimation of location of easier-than-average items will require samples that are larger than are available under most circumstances.

Clearly, unrestricted maximum likelihood estimation in its most basic version is not a viable method for the estimation of item parameters in the three-parameter model. It appears that two solutions are viable. One is a Bayesian scheme in which highly restrictive prior probabilities constrain the variability of the parameter being estimated. Such methods are already in practice and seem to offer some hope. LOGIST (Wood, Wingersky, and Lord, 1976) does this in an informal way by using rectangular prior distributions to restrict the lower asymptote and the slope. When an item appears to be too easy to estimate its lower asymptote effectively, LOGIST groups it with others of the same type and assigns it a lower asymptote based upon the average of the asymptotes of a number of other items. Swaminathan and Gifford (1981) discussed the results of a more formal Bayesian approach which seems to accomplish the same ends more gracefully and more efficiently.

Another approach, suggested by Winsberg (1981) is quite different, and rests on a property of spline functions. She points out that the parameters of most continuous functions have a global effect. That is, the changing of a lower asymptote at one end of the function can have a profound effect on the estimation of another parameter at the other end. This need not be the case with spline functions. If the item characteristic curve is fit with a set of spline functions, the property of splines ensures that the effect of a change of a parameter is local. Thus a poorly estimated lower asymptote will leave the middle of the curve, that which contains the parameters of interest, rock steady. This more dramatic radical departure from traditional practice needs careful study before a judgment can be reached as to its suitability.

As always in a complex situation, limited research cannot answer the question, "What is the best way?" But an answer to the "worst way" does seem to be clear. Unrestricted maximum likelihood estimation for the three-parameter model is not a technique that is likely to give useful results when it is important to have accurate estimation of individual parameters. This conclusion seems incontrovertible.

References

Bradley, J. V. Probability; decision; statistics. Englewood Cliffs, New Jersey: Prentice-Hall. 1976.

Hambledon, R. & Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Kendall, M.G. and Stuart, A. The advanced theory of statistics, Volume II. Third edition. London: Griffin & Co, 1960.

Lord, F. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Wood, R.L., Wingersky, M.S. and Lord, F.M. LOGIST: A computer program for estimating examinee ability on item characteristic curve parameters. Princeton, NJ: Educational Testing Service. Research Memorandum 76-6, 1976, (modified 1/78).

Reference Notes

Swaminathan, H. and Gifford, J.A. Bayesian estimation in item response models. A talk given at the annual meeting of the Psychometric Society, Chapel Hill, North Carolina, 1981.

Winsberg, S. (1981). Personal communication.

Appendix I

Equation (13) specifies

$$\frac{\partial P}{\partial b} = P^*Q^* \ (-a)$$

for both the one and two parameter models. Substituting this into equation (6) yields

$$-E\left(\frac{\partial^2 \ell}{\partial b^2}\right) = N \ a^2 \int_{-\infty}^{\infty} P^*Q^* \ \phi \ (\theta) \ d\theta = B \ . \qquad \textbf{(A1)}$$

This is the diagonal element of the expected Hessian. For the one-parameter model $1/\sqrt{B}$ is the standard error of b. For the two-parameter model there will be some effect of the covariance term. This term (from 6) is

$$-E\left(\frac{\partial^2 \ell}{\partial a \partial b}\right) = -a \ N \int_{-\infty}^{\infty} (\theta - b) \ P^*Q^* \ \ \phi(\theta) \ d\theta .$$

This can be rewritten as

$$-E\left(\frac{\partial^2 \ell}{\partial a \partial b}\right) = -a \ N\left[\int_{-\infty}^{\infty} \underbrace{\theta \ P^*Q^* \ \phi(\theta) \ d\theta}_{c_1} - \int_{-\infty}^{\infty} \underbrace{b \ P^*Q^* \ \phi(\theta) \ d\theta}_{c_2}\right] \qquad (A2)$$

Thus e_1 is the product of an odd function, θ, and an even function, $P^*Q^*\phi(\theta)$, for <u>any</u> symmetric ability distribution ϕ. This is then an odd function of θ which integrates to zero. Thus equation (A2) simplifies to:

$$-E\left(\frac{\partial^2 \ell}{\partial a \partial b}\right) = a\ b\ N \int_{-\infty}^{\infty} P^*Q^*\ \phi\ (\theta)\ d\theta \qquad \text{(A3)}$$

So when b = 0 this covariance vanishes making the standard errors of b for the one and two-parameter models identical. Through continuity arguments, as b approaches zero the standard errors of the two models draw closer. Similarly, as a gets smaller so does the covariance and again the difference between the standard errors diminishes.

Appendix II

Directory of Tables

Table 1. Minimal asymptotic standard errors for locations, one-parameter model, lower asymptote at 0.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	8.61	8.30	8.12	8.06	8.12	8.30	8.61
0.50	5.16	4.57	4.23	4.12	4.23	4.57	5.16
0.75	4.31	3.44	2.98	2.83	2.98	3.44	4.31
1.00	4.09	2.94	2.37	2.20	2.37	2.94	4.09
1.50	4.21	2.50	1.78	1.59	1.78	2.50	4.21
2.00	4.54	2.30	1.49	1.28	1.49	2.30	4.54
3.00	5.10	2.08	1.19	0.98	1.19	2.08	5.10

Table 2. Minimal asymptotic standard errors for locations, two-parameter model, lower asymptote at 0.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	26.85	18.34	11.40	8.06	11.40	18.34	26.85
0.50	15.45	9.75	5.81	4.12	5.81	9.75	15.45
0.75	12.20	7.00	3.98	2.83	3.98	7.00	12.20
1.00	10.84	5.67	3.08	2.20	3.08	5.67	10.83
1.50	9.77	4.33	2.20	1.59	2.20	4.33	9.77
2.00	9.32	3.63	1.76	1.28	1.76	3.63	9.32
3.00	8.65	2.86	1.32	0.98	1.32	2.86	8.65

Table 3. Minimal asymptotic standard errors for slopes, two-parameter model, lower asymptote at 0.

				Locations (θ)			
	-3	-2	-1	0	1	2	3
Slope							
0.25	2.18	2.11	2.06	2.05	2.06	2.11	2.18
0.50	2.68	2.39	2.23	2.18	2.23	2.39	2.68
0.75	3.45	2.82	2.48	2.38	2.48	2.82	3.45
1.00	4.47	3.35	2.80	2.63	2.80	3.35	4.47
1.50	7.30	4.72	3.60	3.28	3.60	4.72	7.30
2.00	11.31	6.47	4.57	4.06	4.57	6.47	11.31
3.00	23.43	11.11	7.01	6.00	7.01	11.11	23.43

Table 4. Minimal asymptotic standard errors for locations, three-parameter model, lower asymptote at 0.

				Locations (θ)			
	-3	-2	-1	0	1	2	3
Slope							
0.25	742.58	587.32	465.51	367.21	285.11	213.70	149.01
0.50	179.36	113.17	73.40	47.59	29.17	15.83	15.97
0.75	94.02	48.31	26.62	14.91	7.94	7.00	14.84
1.00	63.77	27.54	13.37	6.79	3.86	5.86	12.87
1.50	38.51	12.94	5.31	2.54	2.24	4.48	10.54
2.00	26.85	7.70	2.90	1.52	1.76	3.69	9.54
3.00	16.02	3.90	1.48	1.00	1.32	2.86	8.65

Table 5. Minimal asymptotic standard errors for slopes, three-parameter model, lower asymptote at 0.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	12.59	12.16	11.91	11.82	11.90	12.16	12.58
0.50	8.10	7.27	6.79	6.63	6.76	7.21	8.01
0.75	7.39	6.12	5.42	5.17	5.33	5.95	7.14
1.00	7.77	5.93	4.97	4.62	4.80	5.59	7.24
1.50	10.18	6.65	5.02	4.44	4.67	5.89	8.82
2.00	14.27	8.09	5.57	4.75	5.12	7.00	11.98
3.00	27.03	12.27	7.44	6.16	7.08	11.16	23.47

Table 6. Minimal asymptotic standard errors for lower asymptotes, three-parameter model, lower asymptote at 0.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slcpe							
0.25	99.75	75.56	58.03	45.19	35.70	28.59	23.20
0.50	53.22	30.25	17.98	11.19	7.27	4.92	3.44
0.75	44.79	19.68	9.41	4.88	2.71	1.60	0.99
1.00	41.90	14.81	5.94	2.65	1.30	0.68	0.38
1.50	37.64	9.58	2.96	1.05	0.42	0.18	0.08
2.00	32.37	6.52	1.66	0.49	0.16	0.06	0.02
3.00	21.90	3.10	0.56	0.12	0.03	0.01	0.00

Table 7. Minimal asymptotic standard errors for locations, three-parameter model, lower asymptote at 0.1.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	806.23	642.56	514.33	410.88	324.20	248.10	177.90
0.50	194.13	124.27	82.40	55.29	35.82	20.91	19.00
0.75	102.12	53.70	30.79	18.46	10.88	8.52	18.50
1.00	69.79	31.23	16.14	9.12	5.54	6.83	17.34
1.50	43.21	15.59	7.17	3.96	2.96	5.30	15.47
2.00	31.22	10.06	4.40	2.50	2.20	4.39	14.09
3.00	20.37	5.96	2.52	1.54	1.59	3.33	11.77

Table 8. Minimal asymptotic standard errors for slopes, three-parameter model, lower asymptote at 0.1.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	13.64	13.27	13.10	13.15	13.42	13.94	14.72
0.50	8.72	7.91	7.51	7.51	7.93	8.86	10.49
0.75	7.94	6.68	6.07	6.03	6.61	8.03	10.83
1.00	8.35	6.51	5.67	5.59	6.34	8.33	12.69
1.50	10.99	7.44	5.97	5.79	6.90	10.27	19.16
2.00	15.51	9.29	6.93	6.58	8.10	13.20	28.75
3.00	30.06	14.83	9.84	8.96	11.44	20.94	56.09

Table 9. Minimal asymptotic standard errors for lower asymptotes, three-parameter model, lower asymptote at 0.1.

				Locations (θ)			
	-3	-2	-1	0	1	2	3
Slope							
0.25	97.53	74.45	57.73	45.52	36.50	29.77	24.70
0.50	51.95	29.98	18.23	11.72	7.98	5.75	4.37
0.75	43.94	19.82	9.90	5.49	3.37	2.29	1.70
1.00	41.52	15.32	6.61	3.32	1.93	1.28	0.96
1.50	38.55	10.81	3.91	1.78	1.00	0.67	0.52
2.00	34.97	8.45	2.82	1.24	0.70	0.49	0.39
3.00	28.12	6.05	1.92	0.85	0.50	0.37	0.32

Table 10. Minimal asymptotic standard errors for locations, three-parameter model, lower asymptote at 0.2.

				Locations (θ)			
	-3	-2	-1	0	1	2	3
Slope							
0.25	878.74	705.10	569.00	459.07	366.54	284.60	207.77
0.50	210.96	136.60	92.05	63.17	42.23	25.60	22.18
0.75	111.23	59.47	34.93	21.68	13.36	10.10	21.56
1.00	76.42	34.95	18.63	10.97	6.90	7.82	20.38
1.50	48.07	17.87	8.51	4.88	3.60	5.98	18.14
2.00	35.36	11.76	5.31	3.09	2.62	4.92	16.32
3.00	23.72	7.10	3.06	1.88	1.86	3.71	13.31

Table 11. Minimal asymptotic standard errors for slopes, three-parameter model, lower asymptote at 0.2.

				Locations (θ)			
	-3	-2	-1	0	1	2	3
Slope							
0.25	14.83	14.52	14.44	14.63	15.08	15.85	16.96
0.50	9.42	8.63	8.29	8.43	9.08	10.39	12.64
0.75	8.57	7.29	6.73	6.83	7.69	9.63	13.41
1.00	9.01	7.12	6.32	6.38	7.44	10.09	15.90
1.50	11.87	8.18	6.70	6.66	8.16	12.51	24.13
2.00	16.81	10.26	7.81	7.57	9.57	16.05	36.12
3.00	32.79	16.46	11.11	10.31	13.46	25.25	69.65

Table 12. Minimal asymptotic standard errors for lower asymptotes, three-parameter model, lower asymptote at 0.2.

				Locations (θ)			
	-3	-2	-1	0	1	2	3
Slope							
0.25	94.54	72.66	56.80	45.21	36.65	30.28	25.49
0.50	50.27	29.36	18.15	11.92	8.32	6.17	4.84
0.75	42.68	19.62	10.06	5.76	3.67	2.59	1.99
1.00	40.61	15.39	6.87	3.60	2.18	1.51	1.16
1.50	38.53	11.27	4.27	2.04	1.19	0.83	0.66
2.00	35.92	9.13	3.20	1.47	0.87	0.62	0.51
3.00	30.56	6.92	2.28	1.05	0.64	0.48	0.43

Table 13. Minimal asymptotic standard errors for locations, three-parameter model, lower asymptote at 0.3.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	963.82	777.77	631.94	513.96	414.20	325.12	240.41
0.50	230.61	150.78	102.90	71.79	49.02	30.48	25.70
0.75	121.78	65.97	39.43	25.04	15.88	11.82	24.79
1.00	84.01	39.04	21.23	12.83	8.26	8.92	23.52
1.50	53.45	20.24	9.84	5.78	4.26	6.73	20.85
2.00	39.76	13.46	6.18	3.66	3.06	5.51	18.59
3.00	27.10	8.22	3.58	2.21	2.15	4.15	14.92

Table 14. Minimal asymptotic standard errors for slopes, three-parameter model, lower asymptote at 0.3.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	16.24	15.98	15.99	16.31	16.96	17.98	19.44
0.50	10.25	9.46	9.18	9.44	10.32	12.01	14.86
0.75	9.30	7.98	7.46	7.69	8.80	11.23	15.96
1.00	9.77	7.81	7.02	7.21	8.55	11.82	19.01
1.50	12.89	8.99	7.46	7.53	9.39	14.67	28.88
2.00	18.28	11.29	8.70	8.56	11.00	18.77	43.10
3.00	35.78	18.15	12.38	11.65	15.41	29.36	82.36

Table 15. Minimal asymptotic standard errors for lower asymptotes, three-parameter model, lower asymptote at 0.3.

	Locations (θ)						
	-3	-2	-1	0	1	2	3
Slope							
0.25	90.78	70.16	55.22	44.29	36.22	30.21	25.70
0.50	48.16	28.42	17.79	11.86	8.42	6.36	5.08
0.75	41.00	19.13	9.99	5.84	3.81	2.74	2.15
1.00	39.22	15.16	6.93	3.71	2.31	1.63	1.28
1.50	37.79	11.35	4.42	2.16	1.29	0.92	0.74
2.00	35.87	9.39	3.37	1.59	0.95	0.69	0.58
3.00	31.52	7.33	2.47	1.16	0.71	0.55	0.49