The Spanish Journal of Psychology
1999, Vol. 2, No. 1, 74-94

1138-7416

Fitting Logistic IRT Models: Small Wonder

Miguel A. García-Pérez
Complutense University of Madrid

State-of-the-art item response theory (IRT) models use logistic functions exclusively as their item response functions (IRFs). Logistic functions meet the requirements that their range is the unit interval and that they are monotonically increasing, but they impose a parameter space whose dimensions can only be assigned a metaphorical interpretation in the context of testing. Applications of IRT models require obtaining the set of values for logistic function parameters that best fit an empirical data set. However, success in obtaining such set of values does not guarantee that the constructs they represent actually exist, for the adequacy of a model is not sustained by the possibility of estimating parameters. This article illustrates how mechanical adoption of off-the-shelf logistic functions as IRFs for IRT models can result in off-the-shelf parameter estimates and fits to data. The results of a simulation study are presented, which show that logistic IRT models can fit a set of data generated by IRFs other than logistic functions just as well as they fit logistic data, even though the response processes and parameter spaces involved in each case are substantially different. An explanation of why logistic functions work as they do is offered, the theoretical and practical consequences of their behavior are discussed, and a testable alternative to logistic IRFs is commented upon.
Key words: goodness of fit, parameter estimation, item response theory, logistic models, finite state polynomic models, BILOG

La función de respuesta al ítem (FRI) asumida en los modelos al uso en teoría de respuesta al ítem (TRI) es, en la práctica, exclusivamente la función logística. Las funciones logísticas cumplen los requisitos de que su rango es el intervalo [0, 1] y son monótonamente crecientes, pero imponen un espacio paramétrico cuyas dimensiones sólo tienen una interpretación metafórica en el contexto de la evaluación mediante pruebas objetivas. La aplicación de modelos TRI requiere la estimación de los parámetros logísticos que mejor describen unos datos empíricos. Sin embargo, el éxito en la obtención de estos parámetros no garantiza que los constructos representados mediante ellos existan en realidad, puesto que la validez de un modelo no queda establecida sólo por la posibilidad de estimar sus parámetros. Este trabajo muestra que la adopción mecánica de funciones logísticas como FRI en modelos TRI produce estimaciones y ajustes estereotipados. Como prueba, se presentan resultados de un estudio de simulación en el que el modelo logístico produjo un patrón de estimaciones y ajustes de datos no logísticos que fue indistinguible del patrón obtenido para datos logísticos, a pesar de que los datos no logísticos se generaron de acuerdo con un modelo que implica un proceso de respuesta y un espacio paramétrico marcadamente diferentes del logístico. El trabajo termina con unas reflexiones acerca de las razones por las que los modelos logísticos se comportan así y de las consecuencias teóricas y prácticas de ese comportamiento, y también se describe una alternativa empíricamente falsable a las FRI logísticas.
Palabras clave: bondad de ajuste, estimación de parámetros, teoría de respuesta al ítem, modelos logísticos, modelos polinómicos de estados finitos, BILOG

Correspondence concerning this article should be addressed to Dr. Miguel A. García-Pérez, Departamento de Metodología. Facultad de Psicología. Universidad Complutense. Campus de Somosaguas. 28223 Madrid (Spain).
Phone: (+34) 91 394 3061. Fax: (+34) 91 394 3189. E-mail: miguel@psi.ucm.es

I sometimes have a nightmare about Kepler. Suppose a few of us were transported back in time to the year 1600, and were invited by the Emperor Rudolph II to set up an Imperial Department of Statistics in the court at Prague. Despairing of those circular orbits, Kepler enrolls in our department. We teach him the general linear model, least squares, dummy variables, everything. He goes back to work, fits the best circular orbit for Mars by least squares, puts in a dummy variable for the exceptional observation–and publishes. And that's the end, right there in Prague at the beginning of the 17th century.

(Freedman, 1985, p. 359)

A major concern in the application of item response theory (IRT) is the estimation of item and examinee parameters. The interest arises because only when this is done can the theoretical advantages of IRT be obtained. The availability of computer programs such as LOGIST (Wingersky, Barton, & Lord, 1982) or BILOG (Mislevy & Bock, 1984, 1986), which estimate IRT parameters under the one-, two-, or three-parameter logistic models (1PL, 2PL, or 3PL models) has provided test practitioners with a powerful tool to harvest these benefits.

Numerous simulation studies have assessed the efficiency and accuracy with which these and other programs attain their goal in a variety of circumstances, including tests of different lengths, examinee samples of different sizes and/or different distributions of true parameters (e.g., Hambleton & Cook, 1983; Harwell & Janosky, 1991; Hulin, Lissak, & Drasgow, 1982; Ree, 1979; Seong, 1990; Skaggs & Stevenson, 1989; Swaminathan & Gifford, 1983; Vale & Gialluca, 1988). Also, some papers have compared LOGIST and BILOG as to the algorithms they implement, their computational cost, and the characteristics of the estimates they provide (Mislevy & Stocking, 1989; Yen, 1987). The results of all these studies provide a positive outlook of the performance of the programs, as the generating parameters could successfully be recovered in the vast majority of cases.

The capability of logistic models to fit artificial data that violate the assumptions of local independence and unidimensionality has also been explored (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Forsyth, Saisangjan, & Gilmer, 1981; Harrison, 1986; McKinley & Mills, 1985; Reckase, 1979; Yen, 1984). Although these studies showed that logistic functions can fit this type of data and provide parameter estimates, attention to the issue of the extent to which the model fitted the data was only paid by McKinley and Mills.

The extent to which logistic models can fit data sets generated by a different model has also occasionally been assessed. Some studies showed that data generated by logistic models of various numbers of parameters can be fitted to logistic models of fewer parameters (e.g., Dinero & Haertel, 1977; Yen, 1981), and Wood (1978) showed that the 1PL model can fit data generated by a coin-toss process, thus "recovering" fictitious parameters. Mislevy and Verhelst (1987) also showed that the 1PL model can fit a mixture of data generated by a random-guessing process and by the 1PL model itself. All these studies authenticate Bejar's (1983, p. 3) concern that "unfortunately, the programs that [estimate logistic parameters] are capable of returning reasonable-looking estimates even when the data are totally inappropriate for the model assumed by the program. That is, succeeding to estimate the parameters of the model does not insure that we have successfully fitted the model."

The research mentioned in the foregoing paragraphs has systematically failed to acknowledge properly that the choice of the logistic function as the item response function (IRF) for IRT models is an assumption of the theory (Hambleton & Swaminathan, 1985, pp. 9-10; Lord, 1980, p. 30; Weiss & Yoes, 1991, p. 74), one whose adequacy must be checked. Yet, on describing how to address the determination of model/data fit, Hambleton and Swaminathan (1985, chap. 8; see also Hambleton & Murray, 1983) did not list the IRF as an assumption to be checked. It seems that the adequacy of logistic IRFs has been taken for granted, and it is noteworthy that, paying little heed to Bejar's (1983) cautionary comment, practitioners are content with estimating logistic parameters without considering the issue of model/data fit any further than whether the 1PL, 2PL, or 3PL model should be chosen. Indeed, Mislevy and Stocking (1989, p. 57) state that obtaining the advantages of IRT "requires access to flexible and economical computer programs to estimate IRT parameters for items, examinees, and populations of examinees," neglecting to mention that, to begin with, the model should fit the data. The belief seems to have been established that logistic IRFs exist in the real world for computer programs to hunt for the parameters that characterize each conceivable item, and that only some know-how is needed to adjust the options of these programs in order to arrive at the solution that was there to be found (see Mislevy & Stocking, 1989, p. 68). Thus, work on IRT has almost exclusively focused on the development and comparison of parameter estimation techniques and the study of the effects of characteristics of the data sets (sample size, test length, and distribution of the true parameters) and violations of model assumptions (excluding the mathematical form of the IRF) on the capability of available algorithms to recover the generating parameters (Baker, 1987a, 1987b, 1991, 1998; De Ayala, 1992; Gifford & Swaminathan, 1990; Jannarone, Yu, & Laughlin, 1990; Kim & Nicewander, 1993; Kim, Cohen, Baker, Subkoviak, & Leonard, 1994; Lord, 1986; Mislevy, 1987; Swaminathan & Gifford, 1986; Tsutakawa, 1992; Tsutakawa & Johnson, 1990; Tsutakawa & Lin, 1986; Tsutakawa & Soltys, 1988; Wainer & Thissen, 1987; Wang & Vispoel, 1998; Warm, 1989; Weitzman, 1996; Zeng, 1997). No one seems to have questioned whether, in the real world, logistic item and examinee parameters are actually there to be recovered or, in other words, whether the mathematical form of the IRF can be derived from a

psychological theory of performance in objective tests as opposed to adopting a convenient function that the data are forced to fit by fiat.

It is understandable that there has been little discussion in the literature as to what mathematical form and what parameter space the IRF should hold, as there are no competing IRFs that are substantively different from logistic functions. By "substantively different," we mean functions that have theoretical underpinnings and whose parameters have theoretically sound interpretations. This characteristic prevents splines (Ramsay & Abrahamowicz, 1989), chosen with the only criterion of being more flexible than logistic functions, from being eligible as plausible replacements for them. As Goldstein and Wood (1989; see also Blinkhorn, 1997) pointed out, IRT has developed with a stunning disregard for psychological theory which might provide theoretically sound IRFs as replacements for logistic functions. In the past few years we have proposed, developed, and tested a model of performance in objective tests (García-Pérez, 1985, 1987, 1989b, 1990, 1993; see also García-Pérez & Frary, 1989, 1991a) that is central to the work described here.

This paper aims at investigating the capability of logistic functions to fit artificial data generated from IRFs that differ in their resemblance to logistic functions in several aspects. Thus, the paper is similar to Wood's (1978) in its goal, but it differs in three major respects. First, more realistic generating models, involving different and interpretable parameter spaces, are used. Second, the 1PL, 2PL, and 3PL models are all fitted to the generated data. Third, some practical recommendations are given, and an alternative way to define and test (as opposed to merely fit) IRFs for use with IRT models is described. PC-BILOG (Mislevy & Bock, 1986) was used to obtain the logistic parameterizations. There is reason to believe that other programs would have performed just as PC-BILOG did in the situations that will be described below (see Mislevy & Stocking, 1989; Yen, 1987).

The paper is divided into two studies. In the first, a data set was generated by the 3PL model and another set was generated by a slight modification of the 3PL model. The analysis of the 3PL parameterization of 3PL data serves as a baseline with which the rest of the results are compared. The analysis of the data generated by modifying the 3PL model indicates how minor differences between generating and fitted models affect the fit and the recovery of the (still logistic) true parameters. For the second study, data were generated by two different finite state polynomic models (García-Pérez & Frary, 1991a). These models represent major departures from the assumption of underlying logistic IRFs and their associated parameter spaces and, therefore, provide for a more stringent test of the capability of logistic functions to "recover" fictitious parameters.

General Procedure

Responses of 500 examinees to a four-option 50-item test were simulated using four different generating models which will be described in detail below.[1] Random numbers required at several points in the programs were obtained as described by Wichmann and Hill (1982). The programs created data files which were subsequently input to PC-BILOG to obtain examinee and item parameter estimates as well as measures of fit for the test and the individual items.

Each of the four data sets was subject to three PC-BILOG runs in order to obtain their 1PL, 2PL, and 3PL parameterizations. Default options for PC-BILOG were used throughout, and the metric of the logistic function was chosen. Default options do not always guarantee best fit, but using the same options in all cases serves the more relevant goal of making results comparable across data sets. By default, PC-BILOG considers omissions as wrong responses when the 3PL model is fitted. Since two of the data sets included omissions, a fourth PC-BILOG run on each of them sought to obtain their 3PL parameterizations when omissions are treated as fractionally correct responses. The 3PL model fitted with this choice for the treatment of omissions will be called 3PL-C.

From each PC-BILOG run, a number of statistics and estimates were obtained for further analysis. These included the measure of overall fit given by the marginal log-likelihood statistic ($-2 \log L$), the approximate chi-square index of fit for every item, the estimated item parameters, and the estimated examinee abilities. On analyzing these measures, the following issues were specifically addressed in each of the four simulations:

1. Variations in overall fit, as given by the marginal log-likelihood statistic, as a function of the correspondence (or lack thereof) between the generating and the fitted model.

2. Variations in the distribution of the fit of individual items, as given by the approximate chi-square indexes within each fitted model.

3. The relationships between the various estimated item parameters within each fitted model and across the various fitted models.

4. The relationship between true and estimated item parameters within each fitted model.

[1] Sample size and test length were chosen to be large enough (yet reasonably small) to minimize estimation errors caused simply by scarcity of data. It should be noted that Baker (1998) concluded from simulation studies that data sets of 500 examinees and 50 items yield excellent item parameter recovery by BILOG.

5. The relationships between estimated examinee abilities across different fitted models.

6. The relationship between true and estimated examinee abilities within each fitted model.

It should be kept in mind that our goal is to determine the extent to which logistic models can fit data generated using models which differ from the fitted models in several aspects. Therefore, the study does not address the possible effects of varying the true parameter distributions, the number of options per item, the length of the test, or the number of examinees. The choices that are made below about these characteristics are hence of no concern with respect to the outcomes of this study.

Study 1: Fitting Logistic Models to Logistic Data

Generating Models

The first data set for this study was generated from the conventional 3PL model equation

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\,[-a_i\,(\theta_j - b_i)]}, \qquad (1)$$

where b_i, a_i, and c_i are, respectively, the difficulty, discrimination, and pseudo-chance level parameters of item i, and $P_i(\theta_j)$ is the probability that examinee j, with ability θ_j, answers item i correctly. This generating model will be referred to as 3PL.

The second data set was generated by adding to the 3PL generating model the disturbance described by Mislevy and Bock (1986, p. 5-17) to allow a small number of omissions. Examinees responded according to Equation 1 whenever $P_i(\theta_j) \geq .3$, and otherwise they had a .5 probability of omitting. Examinees who did not omit when $P_i(\theta_j) < .3$ also responded according to Equation 1. The generating model resulting from this modification will be referred to as 3PL-O.

True Parameters

For the two simulations, 500 values to represent examinee true abilities were randomly generated to be distributed $N(0, 1)$. Similarly, item difficulties were generated to be uniformly distributed in $[-2.0, 2.0]$ and item discriminations to be uniformly distributed in $[0.6, 2.0]$, whereas pseudo-chance level parameters were kept constant at .25 for all items.[2] The observed θ distribution had a mean of .005 and a standard deviation of .958, with a minimum of –2.79 and a maximum of 2.80. Observed item difficulties ranged from –1.98 to 1.99, with a mean of 0.159 and a standard deviation of 1.143, and observed item discriminations ranged from 0.603 to 1.984 with a mean of 1.338 and a standard deviation of 0.430. The product-moment correlation between difficulty and discrimination was $r = .07$.

Results and Discussion

Table 1 gives the values of $-2 \log L$ obtained for every model fitted to each data set.[3] For 3PL data, fitting the 3PL model resulted in the best $-2 \log L$. For these data, the difference between the 3PL and 1PL parameterizations in terms of $-2 \log L$ was 371.42, and the difference between the 3PL and 2PL parameterizations was 113.66. Of course, the cost of these improvements was, respectively, to estimate 100 or 50 more parameters. The situation for 3PL-O data is about the same, but it is noteworthy that fitting the 3PL-C model resulted in a value of $-2 \log L$, which was much worse than that obtained when fitting the 1PL model. This is probably a consequence of the inappropriateness of default PC-BILOG options for data including omissions. No attempt was made to run PC-BILOG with different options to improve the fit, because obtaining the best possible fit was not a goal of this study.

It is interesting to note that the disturbance in the 3PL-O model did not have a strong effect in model/data fit, provided that omissions are treated as wrong responses when fitting the 3PL model. In fact, the pattern of $-2 \log L$ values across the fitted 1PL, 2PL, and 3PL models is similar for the data with and without omissions. On the other hand, large degrees of mismatch between the generating and the fitted model, as represented by the 2PL and 1PL parameterizations, resulted in a deterioration of the fit. In short, when the same PC-BILOG options are used, the fit seems to be best when the fitted model matches the generating model and, when the 3PL model is fitted to data

---

[2] These choices might be disdained as non-realistic. However, similar choices were made by Swaminathan and Gifford (1983), Hambleton (1983) or Baker (1998), and it should be kept in mind that "notions of what 'realistic' means are determined by what available programs provide, and programs do not necessarily provide the true parameters for any dataset of reasonable size" (Mislevy & Stocking, 1989, p. 73). In any case, this type of realism is not relevant to the goal of this paper and, in addition, the results to be presented below indicate that the possibility of a logistic parameterization of a test is not hampered by this choice for the distributions of the generating parameters.

[3] It should be remembered that likelihood is a function of the data and, then, only comparisons across models fitted to the same data are legitimate. Also, direct comparison of values of $-2 \log L$ gives only a crude indication of fit, but their chi-square approximation is suspect. Also note that, for any given data set, each fitted model lies in a boundary of the next higher-dimensional model, which further lessens the validity of $-2 \log L$ as true chi-squares (for a similar treatment of $-2 \log L$, see Mislevy & Verhelst, 1987).

with omissions, the best choice seems to be to treat omissions as wrong responses. Whether or not this is true in general will be left unexplored here because it is beyond the goal of this paper.

Table 1
Values of − 2 log L in Study 1

	Generating Model	
Fitted Model	3PL	3PL-O
1PL	28478.62	27835.87
2PL	28220.86	27503.89
3PL	28107.20	27482.93
3PL-C	—	28317.71

Table 2 summarizes the information provided by the approximate chi-square index of fit for each item.[4] Because degrees of freedom varied across items and across fitted models, the p-values of the approximate chi-square statistics are reported instead of their values. As can be seen, the 1PL parameterization results in poorer fits for the items, with an important number of them having p-values below .05. In contrast, the 2PL and 3PL parameterizations result in approximately equally good fits, with only a few misfitting items. It is noteworthy that when the 3PL-C model was fitted to 3PL-O data, the deterioration of fit indicated by the − 2 log L statistic in Table 1 does not show at the item level.

Using a modified version of Yen's (1981) criteria for choosing the most appropriate fitted model (amended to replace a comparison of the mean values of the item fit statistics with a comparison of their mean p-values), the

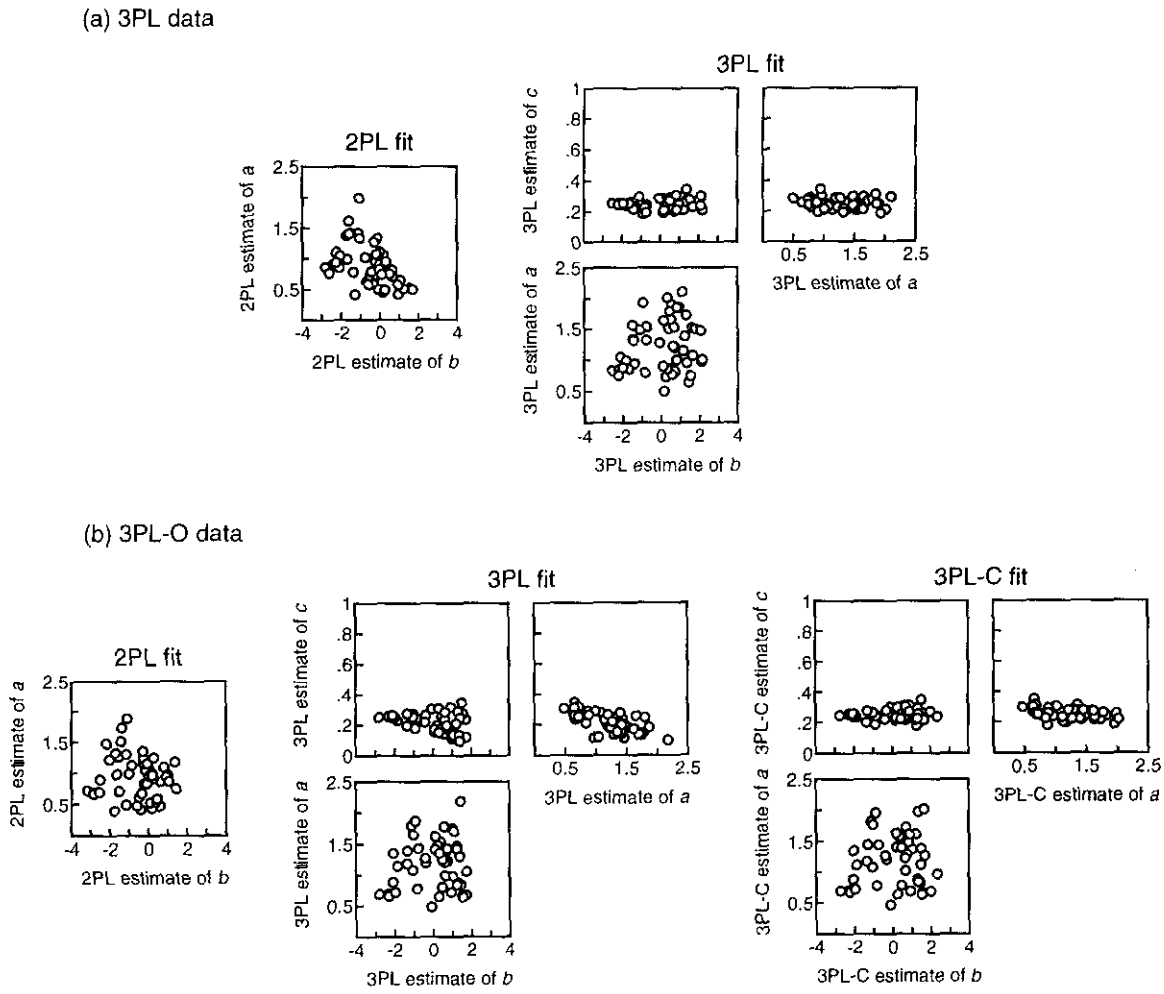

Figure 1. Relations between estimated logistic parameters within each fitted logistic model. (a) 3PL data. (b) 3PL-O data.

[4] These indexes of fit should be interpreted with caution, since they are not true chi-square statistics.

Table 2
Means and Standard Deviations of p-values, and Number of Misfitting Items ($p < .05$) in Study 1

Fitted Model	Generating Model: 3PL			3PL-O		
	M	*SD*	*N*	*M*	*SD*	*N*
1PL	0.36	0.29	13	0.42	0.33	8
2PL	0.55	0.28	1	0.60	0.30	3
3PL	0.55	0.27	1	0.60	0.28	2
3PL-C	—	—	—	0.58	0.33	1

1PL model should be rejected for these data, but a decision about the appropriateness of the 2PL vs 3PL models cannot be made from the values of the statistics reported so far. The differences between these two latter parameterizations are further examined next.

Figure 1 shows the relationships between the various item parameters estimated within each fitted model. For 3PL data (Figure 1a), fitting the 2PL model results in a negative relationship between estimated difficulty and discrimination ($r = -.45$), whereas this relationship almost vanishes when the appropriate 3PL model is fitted ($r = .19$). It should be remembered that true difficulty and discrimination did not bear any relation ($r = .07$). In addition, there is no evidence of any relationship between pseudo-chance level estimates and the two other 3PL parameter estimates ($|r| < .14$).

For 3PL-O data (Figure 1b), the situation is not very different for the 2PL and 3PL models when omissions are treated as fractionally correct responses (2PL and 3PL-C fit, respectively). Yet, when omissions are treated as wrong responses (3PL fit), pseudo-chance level estimates are negatively related to estimated discrimination ($r = -.65$), and their spread around the true value of .25 increases with increasing estimated difficulty.

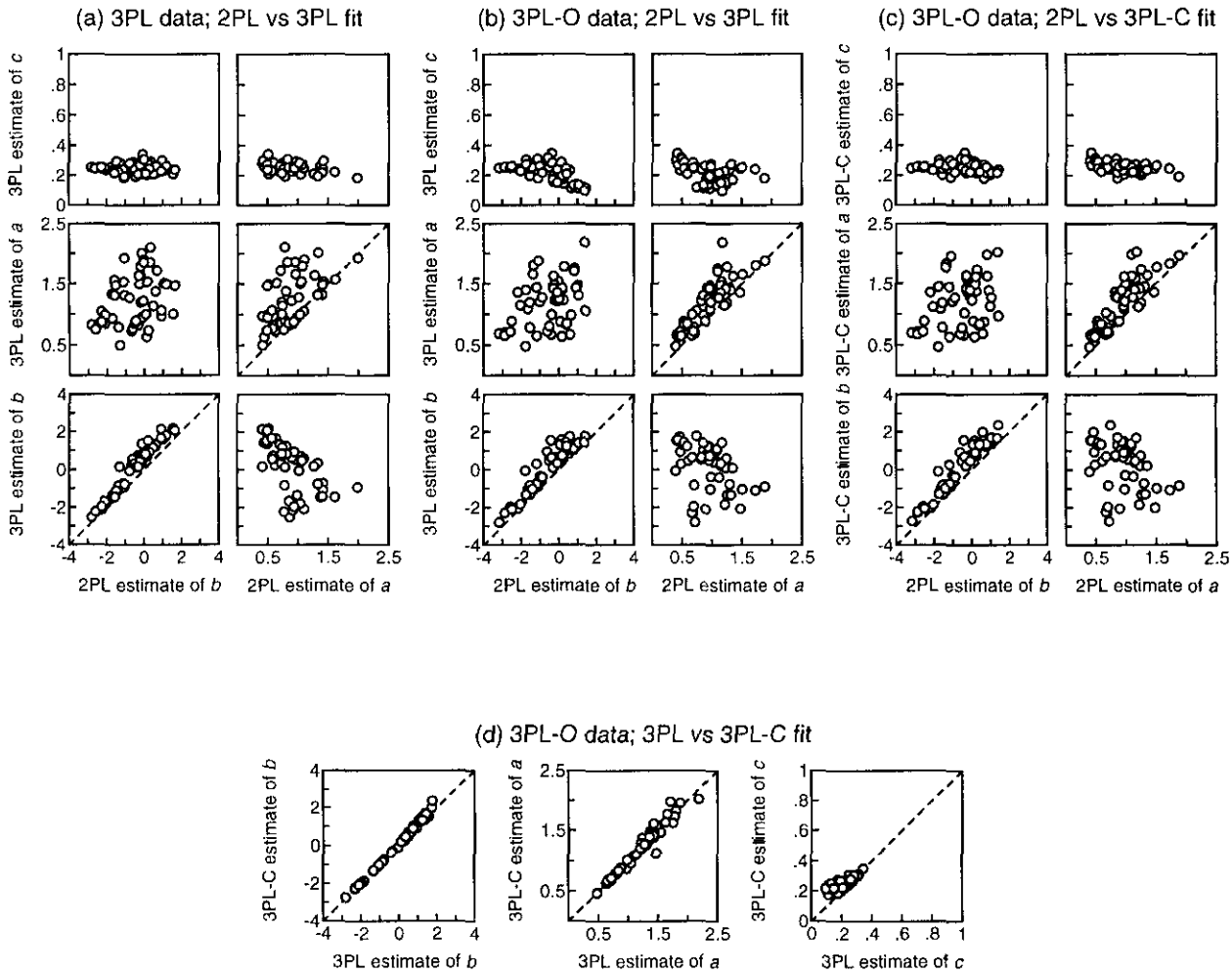

Figure 2. Relations between estimated logistic parameters across different fitted logistic models. (a) 2PL vs 3PL estimates for 3PL data. (b) 2PL vs 3PL estimates for 3PL-O data. (c) 2PL vs 3PL-C estimates for 3PL-O data. (d) 3PL vs 3PL-C estimates for 3PL-O data. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

Figure 2 shows how the estimates vary across fitted models. For 3PL data (Figure 2a), 2PL and 3PL difficulty estimates are similar and highly linearly related (r = .97), but 2PL and 3PL discrimination estimates are not so closely related (r = .53), with values for the 3PL parameterization being generally larger than those resulting from the 2PL fit. The presence of this type of relationship between 2PL and 3PL discrimination estimates was regarded by Yen (1981) as evidence that the 2PL model is inappropriate for the data. Finally, items for which 3PL pseudo-chance level estimates deviate the most from the true value of .25 tend to obtain average 2PL difficulty estimates.

In contrast, for 3PL-O data (Figure 2b) the presence of omissions breaks up slightly the relationship between 2PL and 3PL difficulty estimates (r = .93), and brings the 2PL and 3PL discrimination estimates closer than they were in the absence of omitted responses (r = .84). On the other hand, 3PL pseudo-chance levels noticeably tend to be estimated well below their true value for items that simultaneously obtain high 2PL difficulty and average 2PL discrimination estimates. The relationships between 2PL and 3PL-C estimates (Figure 2c) are similar to those shown in Figure 2b, except that 3PL-C pseudo-chance level estimates bear with 2PL difficulty and discrimination estimates a similar relation as they did for data without omissions (compare with Figure 2a). Therefore, it would seem that, as far as a comparison between 2PL and 3PL or 3PL-C item estimates is concerned, the only meaningful difference between the two options for the treatment of omissions shows in the estimation of pseudo-chance levels. This is best seen in Figure 2d: the different treatments of omissions do not affect item difficulty estimates, slightly affect discrimination estimates for items of high estimated discrimination, and substantially affect pseudo-chance level estimates.

To determine whether these characteristics affect items within specific ranges of true parameters, the relationships between true and estimated item parameters were explored within each fitted model. Figure 3a shows the relationships between true difficulty and discrimination and their 2PL estimates for 3PL data. Figure 3b does the same for 3PL-O data. In both cases, difficulty seems to be slightly underestimated by the 2PL model, but the estimates are highly correlated with true difficulty (r = .97 and r = .96 in Figures 3a and 3b, respectively) and virtually uncorrelated with true discrimination (r = .23 and r = .27). Conversely, 2PL discrimination estimates are less strongly related to true discrimination (r = .61 and r = .83) and they are negatively related to true difficulty (r = –.56 and r = –.28). Consistent with Yen's (1981) interpretation of the way in which the absence of a third parameter is made up for when the 2PL model is fitted to 3PL data, Figures 3a and 3b show that the underestimation of 2PL discrimination affects items of high true difficulty and discrimination.

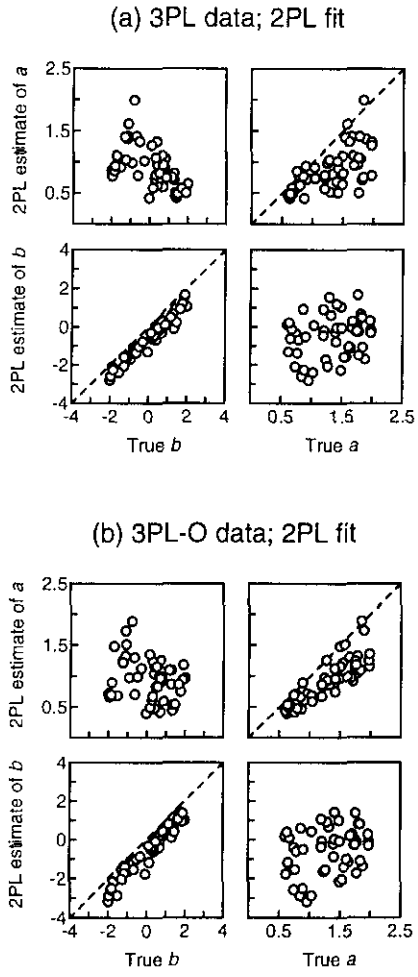

Figure 3. Relations between true a and b and 2PL estimates. (a) 3PL data. (b) 3PL-O data. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

The relationships between true parameters and their 3PL estimates (for 3PL and 3PL-O data) or their 3PL-C estimates (for 3PL-O data) are shown in Figures 4a, 4b, and 4c, respectively. In all three cases, difficulty estimates are very close and linearly related to true difficulty (r > .97), and they bear no relation to true discrimination ($|r|$ < .06). Estimated discriminations are only slightly less related to the true values (.86 < r < .90), but they do not bear any relation to true difficulty ($|r|$ < .19). On the other hand, pseudo-chance level estimates are unrelated to true difficulty and discrimination, except when omissions are treated as wrong responses (Figure 4b), where they show signs of a negative relation to true difficulty (r = –.49) and discrimination (r = –.62).

As for examinee ability, the relations among the estimates obtained from the various models fitted to each data set are shown in Figure 5, revealing that ability estimates are almost

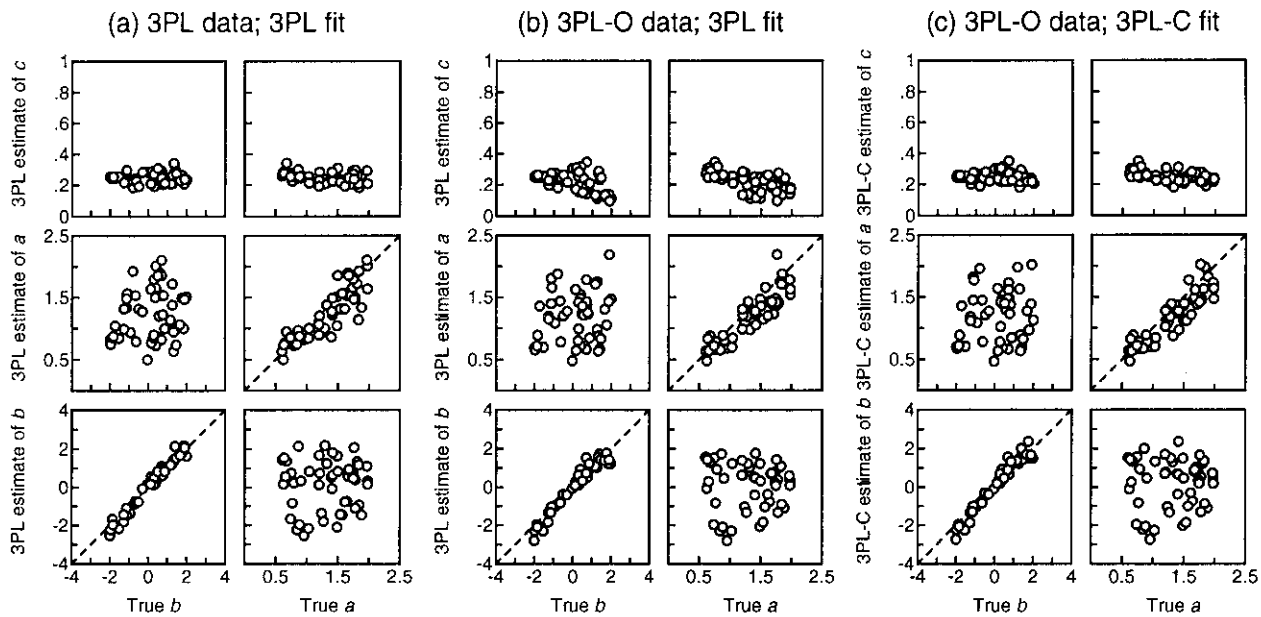

Figure 4. Relations between true a and b and 3PL estimates. (a) 3PL fit to 3PL data. (b) 3PL fit to 3PL-O data. (c) 3PL-C fit to 3PL-O data. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

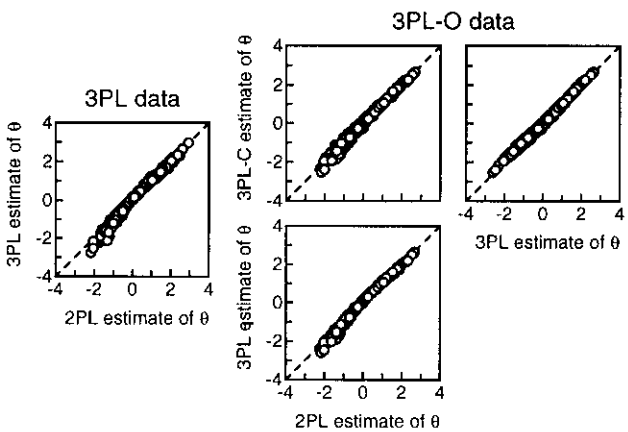

Figure 5. Relations between estimated logistic abilities across models fitted to each data set. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

the same throughout ($r > .99$). Figure 6 shows that all ability estimates are also about equally related to true ability ($.92 < r < .94$) no matter which model was used to generate the data or how omissions were treated. It is also clear from a comparison of Figures 5 and 6 that the different ability estimates are much closer to each other than any of them is to true ability.

Conclusion

The main goal of this first study was to obtain a 3PL parameterization of 3PL data that could set a standard of comparison for the parameterizations obtained for data differing from 3PL data in several aspects. In the first simulation, where data were generated to match both the parameter space and the response process that the fitted model assumes, the 3PL parameterization recovered the true

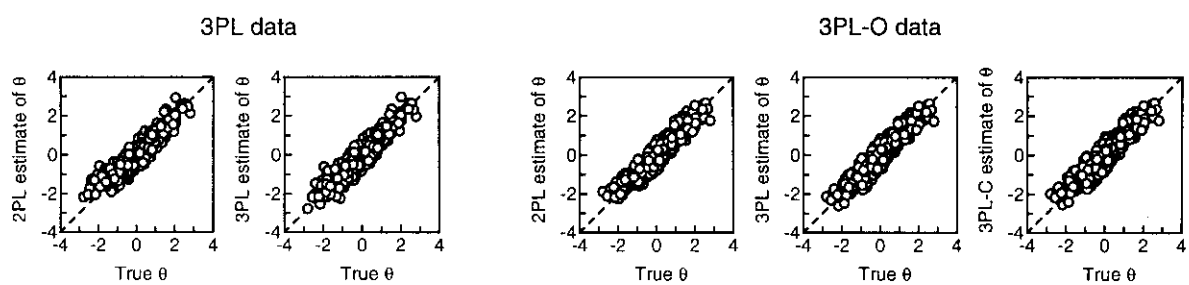

Figure 6. Relations between true and estimated logistic abilities across models fitted to each data set. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

parameters reasonably well. In the second simulation, where generating and fitted models shared the parameter space but implied slightly different response processes, the 3PL parameterization could also recover the true parameters. In both simulations, the recovery was far from perfect, with noticeable errors in the item discrimination parameters and, to a lesser extent, in the item difficulty and examinee ability parameters. In agreement with results of Yen (1981) and McKinley and Mills (1985), differences between 2PL and 3PL parameterizations were minimal, as were those between 3PL parameterizations obtained with different options for the treatment of omissions.

These analyses have focused on the recovery of parameters as well as on the relationships among estimated parameters across and within fitted models. The issue of the recovery of IRFs (i.e., whether the shape of the estimated IRF approximates that of the true IRF regardless of differences between estimated and true parameters) has not been addressed because the recovery of parameters was indeed sufficiently accurate. These results characterize the performance of logistic model parameter estimation techniques when generating and fitted models match or almost match. The next study explores how these techniques behave when the differences between generating and fitted models are more dramatic.

Study 2: Fitting Logistic Models to Finite State Polynomic Data

Generating Models

Each of the two data sets on which this study is based was generated using a different finite state polynomic (FSP) model. A thorough description of these models can be found in García-Pérez and Frary (1991a), but a brief introduction follows. FSP models were developed in the context of measurement of educational achievement, and they include two examinee parameters and an item parameter. The main examinee parameter, λ ($0 \leq \lambda \leq 1$), represents ability or level of knowledge and bears no relation to its logistic counterpart θ. In finite state theory, λ stands for the proportion of statements about a subject matter whose truth value the examinee knows. Thus, unlike θ in logistic models, λ is directly interpretable as the proportion of knowledge that the examinee has. (Out of the context of domain-referenced testing, or simply to avoid sampling considerations, a less stringent definition of λ is that it is the proportion of statements on a test that the examinee knows.) On a test, a statement is represented by the completion of an item stem with any one of its response options. Thus, the probability of an examinee's identifying a randomly drawn option in a randomly drawn item as the correct answer or a distractor is, before considering item characteristics, simply λ.

The FSP item difficulty parameter δ ($0 < \delta < 1$) modifies this probability. Like λ, δ bears no relation to its logistic counterpart *b*. It characterizes items from the difficult ($\delta \rightarrow 0$) to the easy ($\delta \rightarrow 1$), and interacts with λ to determine the probability of an examinee's identifying an option in a certain item as its correct answer or a distractor. García-Pérez and Frary (1991a) proposed a mathematical form for the interaction between examinee ability and item difficulty, but that equation was later amended by García-Pérez (1994). Here we will adopt the latter form so that the probability that an examinee of ability λ_j *knows the truth value of an option* in an item of difficulty δ_i is

$$p_{ij} = \lambda_j^{-\log_2 \delta_i} \qquad (2)$$

This equation has a more natural interpretation than the original proposal because the effects of λ and δ are symmetric. Specifically, for a statement of average difficulty ($\delta = .5$), the probability that an examinee of ability λ knows its truth value becomes just λ; similarly, for an examinee of average ability ($\lambda = .5$) the probability that he/she knows the truth value of a statement of difficulty δ becomes just δ. In other words, an examinee's λ is interpretable as the probability of his/her knowing the truth value of statements of average difficulty, whereas a statement's δ is interpretable as the probability that its truth value be known by examinees of average ability.

On a multiple-choice test, the probability that an examinee *responds correctly* to an item is definitely a function of this basic probability as applied to each of the options in the item, but it also depends on a number of other factors, among which guessing is included. This calls for the second examinee parameter, γ ($0 \leq \gamma \leq 1$), which represents the examinee's willingness to guess when unsure of the answer, regardless of how many distractors they had the knowledge to identify. This parameter has no parallel in logistic models and represents the probability that an examinee will guess at random among the unclassified options (as opposed to omitting) when unsure of the correct answer to an item.

García-Pérez and Frary (1991a) discussed some other assumptions about examinee behavior and test and item characteristics reflecting features of the testing situation that can be taken into account to derive a matching FSP model. These assumptions cover the guessing strategy of the examinees, the number of options per item, the relative identifiability of correct answers vs distractors, the format of administration of the test, and other item characteristics such as use of "none of the above" (see also García-Pérez, 1993; García-Pérez & Frary, 1991b). They also exemplified the procedure for developing FSP models and provided a number of models that incorporate different sets of assumptions. Two of these were selected for the simulations carried out here.

Data for the *second* simulation were generated using Equations 3a-3c of García-Pérez and Frary (1991a). Those equations embody the model for a four-option test administered with the conventional format (i.e., asking examinees to mark the correct option), in which distractors are as easily classifiable as correct answers, and where examinees guess without following any consistent strategy. This type of examinee behavior was referred to as *random omission* (RO) by García-Pérez and Frary (1989). The corresponding generating model, which we will refer to as FSP-RO, is given by

$$c_{ij} = p_{ij}^4 + 4p_{ij}^3(1-p_{ij}) + 3p_{ij}^2(1-p_{ij})^2 + \frac{3}{2}p_{ij}^2(1-p_{ij})^2\gamma_j + p_{ij}(1-p_{ij})^3 + p_{ij}(1-p_{ij})^3\gamma_j + \frac{1}{4}(1-p_{ij})^4\gamma_j, \quad (3a)$$

$$w_{ij} = \frac{3}{2}p_{ij}^2(1-p_{ij})^2\gamma_j + 2p_{ij}(1-p_{ij})^3\gamma_j + \frac{3}{4}(1-p_{ij})^4\gamma_j, \quad (3b)$$

$$u_{ij} = 3p_{ij}^2(1-p_{ij})^2(1-\gamma_j) + 3p_{ij}(1-p_{ij})^3(1-\gamma_j) + (1-p_{ij})^4(1-\gamma_j) \quad (3c)$$

with p_{ij} as in Equation 2. Equations 3a-3c, respectively, represent the probabilities that an examinee of ability λ_j and willingness to guess γ_j responds correctly, wrongly, or leaves unanswered an item of difficulty δ_i. Figure 7 plots the IRF of Equation 3a for various γ and δ. Note that, in addition to providing an equation for the probability of a correct response to an item under the conditions assumed for the test, FSP models also supply equations for the probabilities of a wrong response and an omission and, in general, for every outcome that may occur under any format of administration.

Data for the *first* simulation were generated by a variant of this model which differed only in that examinees were assumed to attempt all the items by guessing whenever necessary, a behavior that was referred to as *number correct* (NC) by García-Pérez and Frary (1989). This behavior is often encouraged on evidence of the major effects of differential guessing strategies on test scores (Albanese, 1988; Bliss, 1980; Cross & Frary, 1977; Rowley & Traub, 1977; Slakter, 1968). The equations for this generating model, which will be referred to as FSP-NC, are straightforwardly obtained from Equations 3a-3c by noting that NC behavior makes $\gamma_j = 1$ for all examinees. When this substitution is made, the right-hand side of Equation 3c reduces to 0 and, not unexpectedly, the model reflects that omissions do not occur under this guessing strategy.

Note that, as a consequence of this choice of models, this study also used a data set in which there were no omitted responses and a data set in which there were omissions.

True Parameters

For both simulations, 500 values to represent examinee abilities were randomly drawn to be uniformly distributed in (.1, .9). Another set of 500 values to represent examinee willingness to guess was generated to be uniformly distributed in [0, 1]. Item difficulties were generated to be uniformly distributed in (.1, .9). The observed λ distribution had a mean of .520 and a standard deviation of .227, with a minimum of .103 and a maximum of .899; the observed γ distribution ranged from .001 to .999, with a mean of .501 and a standard deviation of .288; λ and γ correlated $r = .04$. Finally, observed δs ranged from .104 to .886 with a mean of .529 and a standard deviation of .234.

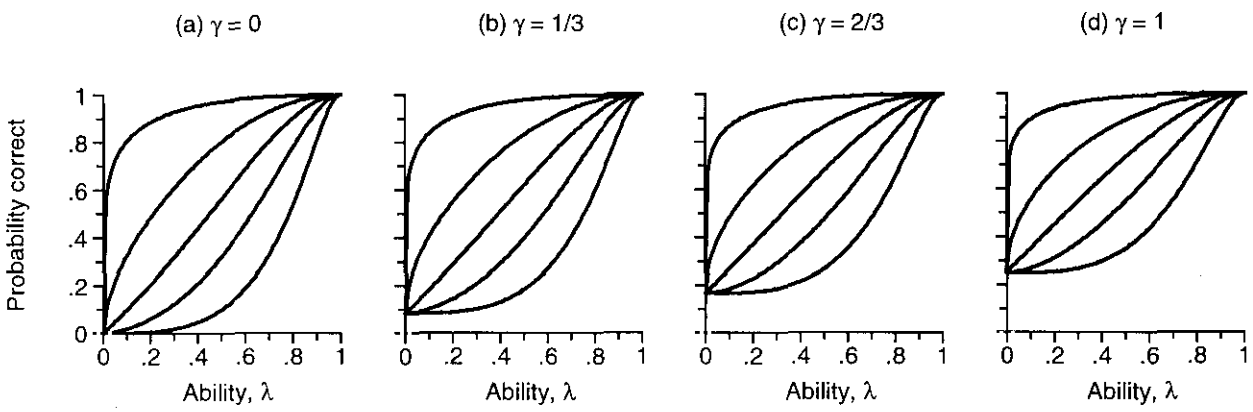

Figure 7. Finite state polynomic IRFs given by Equation 3a for various γs. In each panel, curves represent, from top to bottom, IRFs for items with δs of .9, .7, .5, .3, and .1.

Table 3
Values of – 2 log L in Study 2

Fitted Model	Generating Model: FSP-NC	Generating Model: FSP-RO
1PL	22484.12	23231.31
2PL	22406.22	23172.29
3PL	22334.31	23182.51
3PL-C	—	23263.97

Table 4
Means and Standard Deviations of p-values, and Number of Misfitting Items ($p < .05$) in Study 2

Fitted Model	Generating Model: FSP-NC			FSP-RO		
	M	*SD*	*N*	*M*	*SD*	*N*
1PL	0.44	0.32	3	0.56	0.28	0
2PL	0.55	0.28	1	0.65	0.26	0
3PL	0.64	0.27	0	0.59	0.24	0
3PL-C	—	—	—	0.71	0.25	0

Results and Discussion

Values of – 2 log *L* for all the models fitted in each simulation are shown in Table 3. Within each data set, the values of – 2 log *L* resulting from the different fitted models are much closer together than they were in the previous study. For FSP-NC data, the value of – 2 log *L* improves by 77.90 when the 2PL model is fitted instead of the 1PL model, and by 71.91 when the 3PL model is fitted instead of the 2PL model. For FSP-RO data, the situation is much the same, including the worsening of the fit when the 3PL model is fitted by considering omissions as fractionally correct responses.

Approximate item chi-square statistics are provided in Table 4. As in the previous study, for both data sets the 1PL model tended to fit worse than the 2PL or 3PL models, both of which did about equally well.

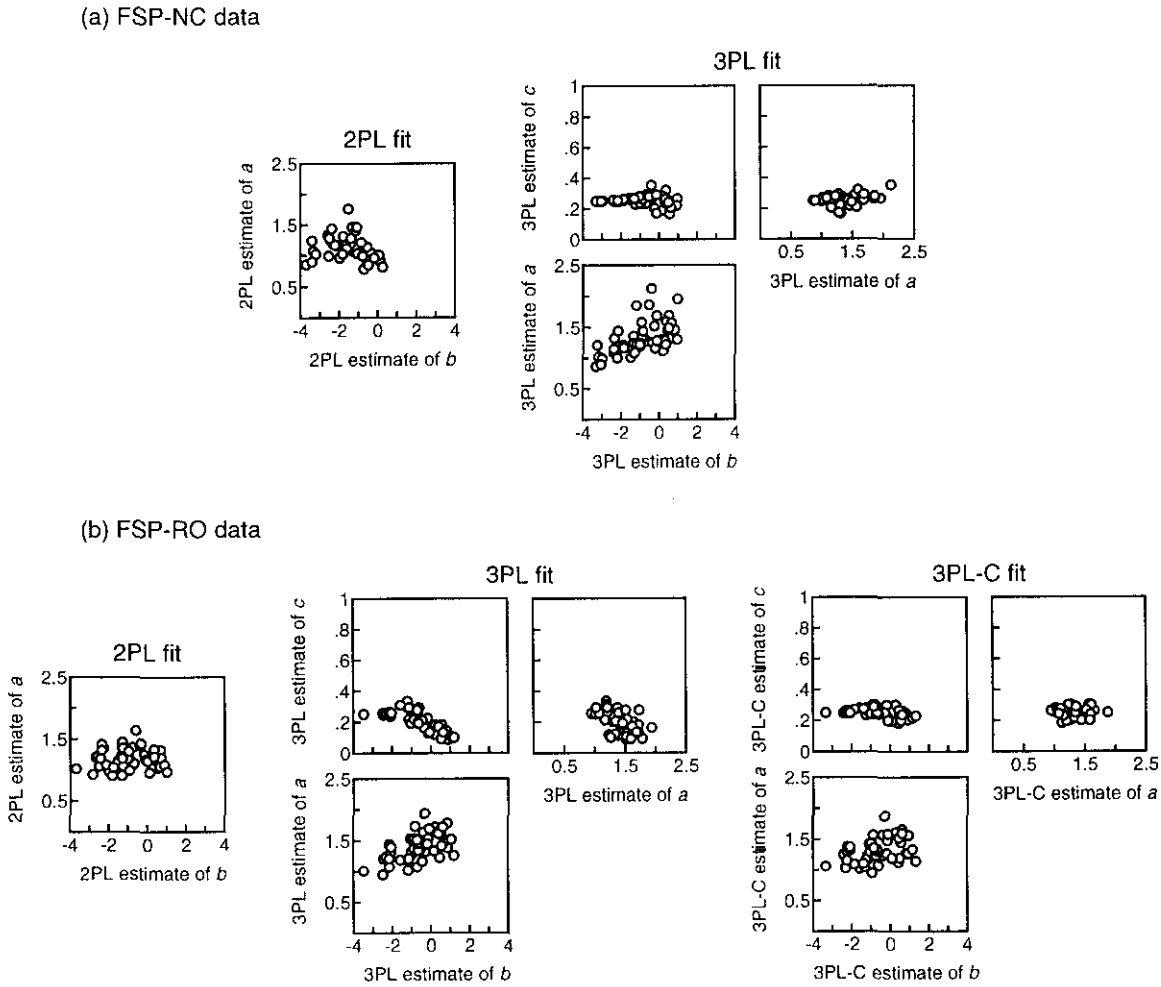

Figure 8. Relations between estimated logistic parameters within each fitted logistic model. (a) FSP-NC data. (b) FSP-RO data.

Comparing the patterns in Tables 3 and 4 with those in Tables 1 and 2, one would be tempted to say that logistic models fit FSP data better than they fit logistic data! However, the different data and the different parameter spaces involved in each generating model make any such comparison unfair. Other FSP models or other distributions for the true parameters could possibly be found that would reverse these results. In any case, the conclusion that can reasonably be drawn from these results is that the form of the generating IRF and its associated parameter space goes unnoticed to the logistic function fitting procedure: logistic models can fit FSP data at least no worse than they fit logistic data.

Applying again the above-mentioned version of Yen's (1981) first step to determining which logistic model is more appropriate, from Table 4, one would discard the 1PL model on the basis of its comparatively lower mean item p-value and the similarity of the means for the 2PL and 3PL models. Again, an analysis of the relationships between estimated 2PL and 3PL parameters will be necessary to decide between the 2PL and 3PL models, an analysis that, in this case, will also reveal how logistic function fitting procedures dress the FSP parameter space in logistic costume.

The relationships between estimated item parameters within each fitted model are shown in Figure 8a for FSP-NC data and in Figure 8b for FSP-RO data. In comparison with analogous plots for logistic data in Figures 1a and 1b, the only difference seems to be that the range of the estimated 2PL and 3PL discriminations is somewhat narrower, and the estimated 3PL discriminations are slightly shifted towards lower values than was found for 3PL data. Taking this into consideration, all plots in Figure 8 look very much like those in Figure 1 after a shrinkage of the item discrimination range.

The relationships between estimated item parameters across fitted models are shown in Figure 9. Results for FSP-NC data (Figure 9a) are very much like results for 3PL data in Figure 2a if the shrunken item discrimination range is taken into account. Estimated 2PL and 3PL difficulties are more similar to one another than estimated 2PL and 3PL discriminations are. Also, items for which the estimated 3PL pseudo-chance parameters deviate the most from the maximal chance level[5] of .25 tend to be assigned high 2PL difficulty estimates and low 2PL discrimination estimates.

For FSP-RO data, Figure 9b shows estimated 2PL parameters against estimated 3PL parameters obtained when omissions are regarded as wrong responses. Presence of omissions broadens the range of estimated difficulties for both fitted models and further shrinks the range of estimated 2PL and 3PL discriminations. In comparison with Figure 9a, estimated 3PL and 2PL difficulties remain close to one another, and estimated 3PL and 2PL discriminations are more tightly packed than they were in the absence of omissions. On the other hand, and paralleling what was discussed about Figure 2b in the previous study, 3PL pseudo-chance level estimates are below the theoretical chance level of .25 for items with high estimated 2PL difficulty.

If omissions are treated as fractionally correct responses (Figure 9c), the range of estimated 3PL-C discriminations shrinks, and estimated 3PL-C pseudo-chance levels shift back to the same relationship with 2PL difficulty estimates as they bore in the absence of omissions.

Figure 9d shows the relationships between estimated 3PL and 3PL-C item parameters when there are omissions. Just as was pointed out for 3PL-O data in Figure 2d, difficulty estimates are very similar across both fitted models (r = .99), discrimination estimates are only slightly less similar across models (r = .94), and pseudo-chance level estimates is where both options for the treatment of omissions differ the most (r = .75).

For these generating FSP models, it is more interesting to see how the various item parameter estimates resulting from fitting logistic models relate to the single true item parameter δ. Figure 10a shows these relationships for FSP-NC data. Estimated 2PL difficulty is highly negatively[6] related to true δ (r = –.96), and estimated 2PL discrimination is positively related to true δ (r = .50). When the 3PL model is fitted to FSP-NC data, the relationship between true δ and estimated difficulty remains the same (r = –.96), estimated discrimination becomes slightly negatively related to true δ (r = –.44), and estimated pseudo-chance level fails to hold any meaningful relation to true δ (despite r = .27, possibly because the spread of the estimates around .25 is broader when δ is below .5).

The relations between δ and estimated logistic parameters from FSP-RO data are shown in Figure 10b. The relation between true δ and estimated 2PL difficulty is similar to what it was in the absence of omissions (r = –.98) but the relation with estimated 2PL discrimination vanishes (r = .02). When the 3PL model is fitted treating omissions as wrong responses (3PL fit), estimated discrimination is negatively related to true δ (r = –.56), and estimated pseudo-chance level is highly positively related to true δ (r = .87). When the 3PL-C model is fitted, the relationships between true δ and estimated logistic parameters are about the same as those found in the absence of omitted responses.

---

[5] Bear in mind that, in FSP models, δ is the only item parameter. In addition, as Figure 7 shows, the theoretical chance level will in general be different for different examinees because it equals γ_j/n, n being the number of options in the item.

[6] Bear in mind that δ in FSP models decreases with increasing difficulty, whereas b in logistic models increases with increasing difficulty.

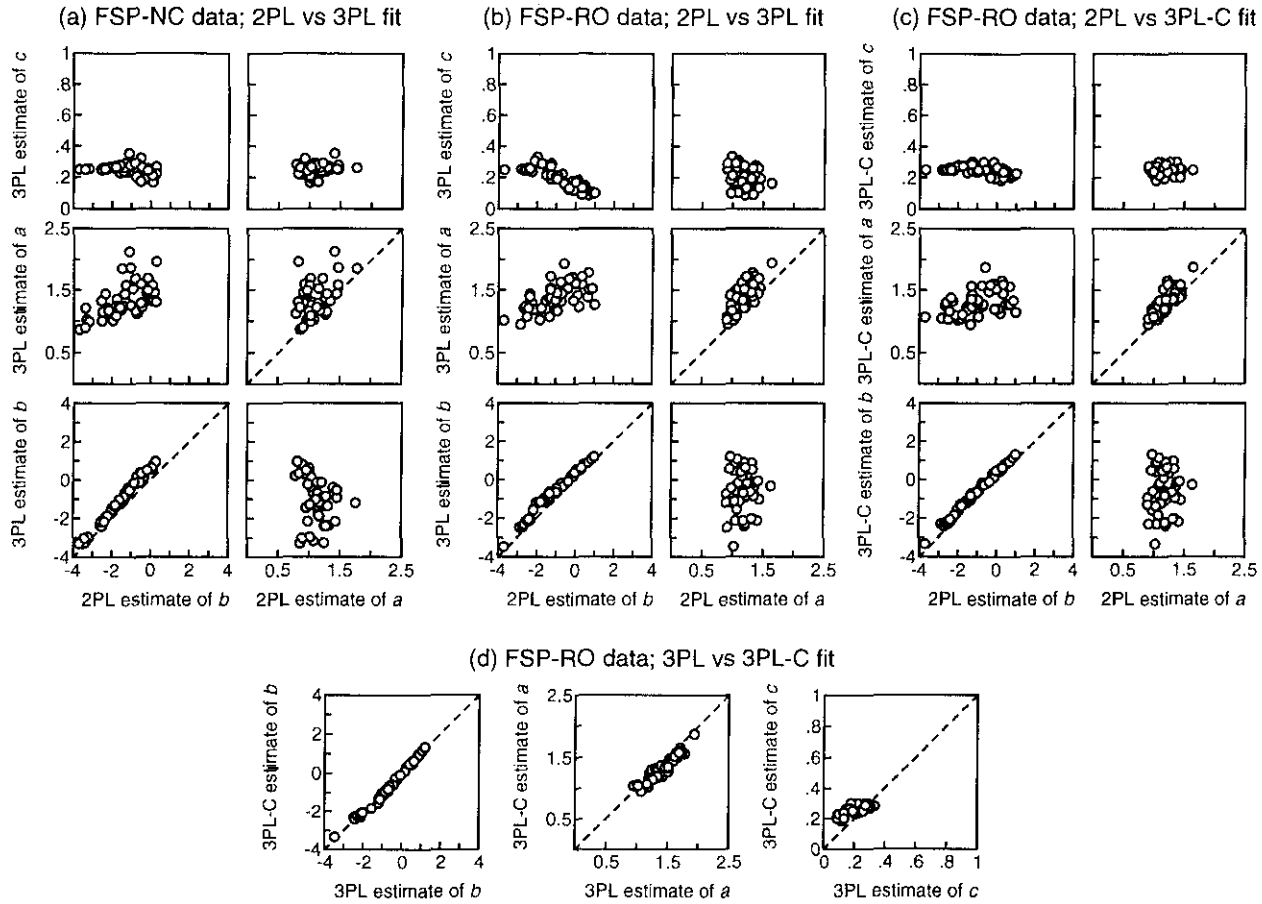

Figure 9. Relations between estimated logistic parameters across different fitted logistic models. (a) 2PL vs 3PL estimates for FSP-NC data. (b) 2PL vs 3PL estimates for FSP-RO data. (c) 2PL vs 3PL-C estimates for FSP-RO data. (d) 3PL vs 3PL-C estimates for FSP-RO data. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

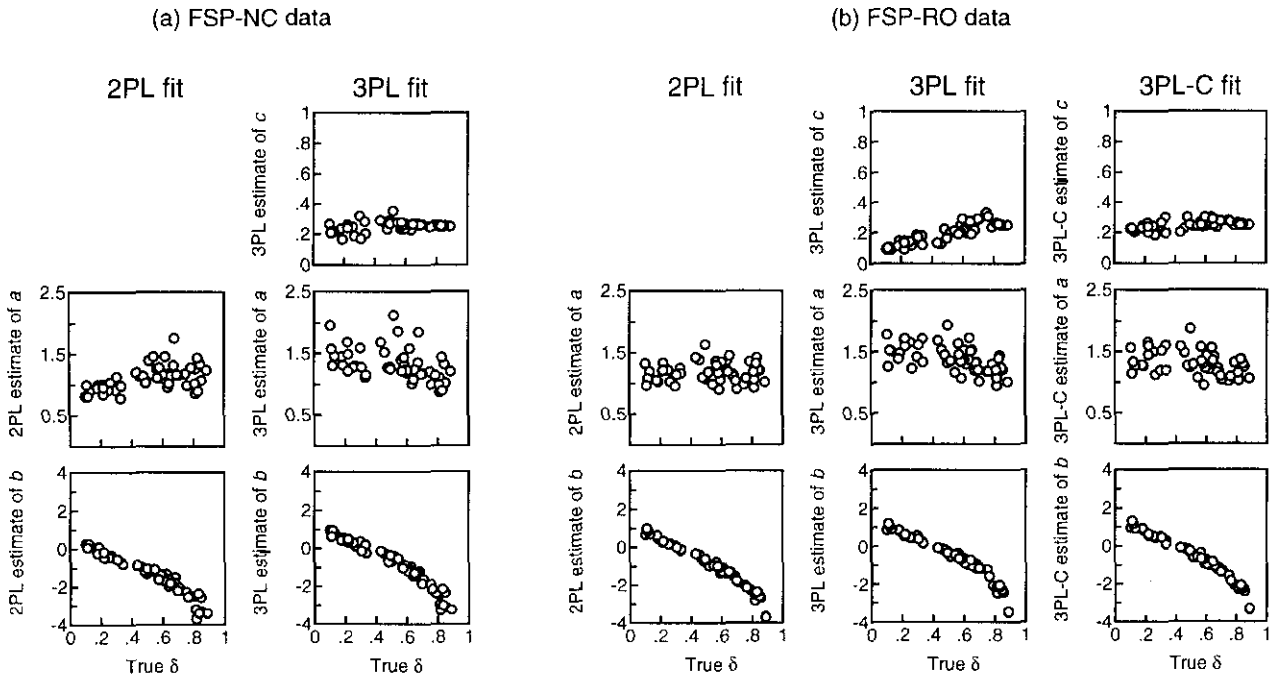

Figure 10. Relations between true δ and 2PL and 3PL estimates. (a) FSP-NC data. (b) FSP-RO data.

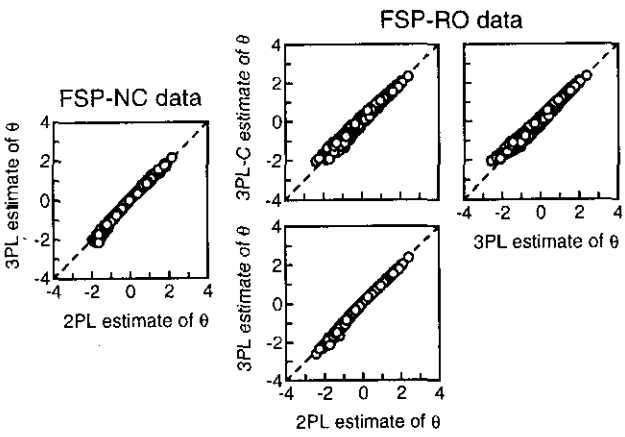

Figure 11. Relations between estimated logistic abilities across models fitted to each data set. Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

Figure 11 shows that the various θs estimated by fitting logistic models to FSP-NC and FSP-RO data are as linearly related as they were for actual 3PL data in Figure 5 ($r > .98$). Figure 12 shows, however, that none of them is as tightly related to actual λ ($.93 < r < .96$). It is interesting to note that estimated θs were unrelated to true γ (an irrelevant factor for the purpose of ability estimation) for FSP-NC data ($|r| < .03$), but for FSP-RO data they were positively related to true γ ($r \approx .23$). In other words, when omissions are allowed, examinees with higher propensities to guess will spuriously obtain higher logistic ability estimates.

Conclusion

In all relevant respects, the results of Study 2 are equivalent to those of Study 1. Tables 3 and 4, and Figures 8, 9, and 11 for FSP data show statistics and relationships that do not differ from what PC-BILOG produces for actual 3PL data (compare with Tables 1 and 2, and Figures 1, 2, and 5). If Yen's (1981) criteria were used, one would conclude that the 3PL model fits all data from Study 2, and the 3PL parameterization would then be used. The trouble, however, is that its referent constructs do not exist in the reality that generated the data.

Of course, Figures 10 and 12 show that estimated logistic difficulty is highly related to true δ, and that estimated logistic ability is also highly related to true λ. Then, one might think that there must be transformations–which there are not–that will translate logistic *b*s into FSP δs and logistic θs into FSP λs, and viceversa. Indeed, there is no way to go from one parameter space to the other because the logistic parameterization returns two extra "estimates" of fictitious parameters (item discrimination and pseudo-chance level) and fails to estimate one of the actual parameters[7] (γ). Note also that the recovery of IRFs cannot even be studied when logistic models are fitted to FSP data, because the domains of true and estimated IRFs are incommensurable.

General Discussion

It is perhaps surprising that the attempt to fit logistic functions where they do not belong can succeed with such flying colors. It seems that in the absence of an external clue as to what type of IRF and parameter space generated the data, trying to fit logistic functions to them is not going to provide one: BILOG produces estimates that bear the same relationships among each other when the input are either logistic or FSP data. In other words, application of an off-the-shelf computer program for logistic model parameter estimation produces off-the-shelf results regardless of how similar or different the generating model was from the would-be logistic model.

Routine and pragmatism guide the adoption of logistic IRFs, and most users are content fitting these functions to their data. Actually, Lord (1980, p. 31) explicitly expressed

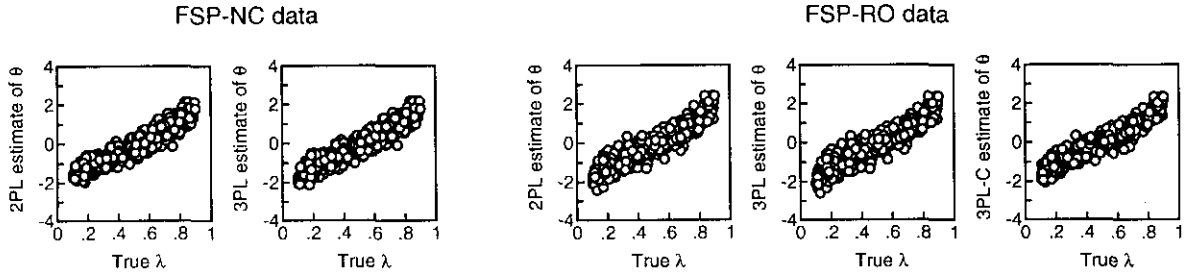

Figure 12. Relations between true λ and estimated logistic ability across models fitted to each data set.

[7] If there are no omissions, this parameter does not need to be estimated.

this pragmatic orientation towards a preference for logistic IRFs when he claimed that "justification of their use is to be sought in the results achieved, not in further rationalizations." Also, Hambleton and Swaminathan (1985, p. 162) pointed out that "item response models are often chosen as the mode of analysis in order to obtain the advantages [of IRT]." Some authors even regard the capability of fitting data as a manifestation of the maturity of IRT (Thissen & Steinberg, 1984, p. 518). Under this "good-fit-above-all-else" philosophy, justification for the use of logistic functions could be sought in their "success" at parameterizing the FSP data in Study 2. Yet, it is not clear what kind of success this is. An investigation into the behavior of algorithms for fitting logistic functions is undertaken next. Also, some theoretical and practical consequences of blind adherence to logistic IRFs are commented upon, and the characteristics of FSP IRFs, as alternatives to logistic IRFs, are briefly discussed.

Will Logistic Functions Always Fit Test Data?

This question needs some qualification, since it is clear that logistic models will not fit data that are not approximately Guttman scalable, and also that the 1PL and 2PL models are sometimes rejected in favor of the 3PL model. However, this is not the issue. For one thing, given the nested hierarchy of logistic models, no one should be surprised that the 3PL model fits better than the 2PL model which, in turn, fits better than the 1PL model. The issue is whether all three of them can be found to be "not-the-models-that-generated-the-data." In view of the results of Study 2 above, it is hard to imagine how this could happen.

Also, it is fallacious that logistic functions will fit everything. It is well known that better fits are sometimes obtained when a few items that seem to depart from the fitted model are dropped, a practice that has been criticized, apparently with little impact, by Goldstein (1979) and Traub (1983). The results of the first simulation in Study 1 show that one should expect to find a few apparently misfitting items even when the data are generated by the model to be fitted. Therefore, their eventual occurrence in the fitting of logistic models to real test data reveals standard performance of the algorithm and gives added emphasis to the point that estimation algorithms merely try to maximize overall fit, even at the expense of having to tag some items as misfitting.

Moreover, the estimation algorithm seems indeed to focus on obtaining ability estimates with a certain distribution, obtaining along the way whatever item parameters are necessary to achieve this goal. This preference shows in that, for any given data set, estimated item parameters vary much more than examinee parameters across fitted models (see relations between 2PL and 3PL estimates of b and a in Figure 2, and compare with the much tighter relations between 2PL and 3PL estimates of θ in Figure 5; the same holds for analogous relations arising from FSP data in Figures 9 and 11). Figure 13 shows histograms of the logistic ability estimates obtained from each fitted model for each data set, clearly revealing that estimated logistic abilities wind up having roughly the same distribution. For logistic data (Figure 13a), the bell-shaped form of these distributions might be taken as a natural consequence of the fact that the data were generated from true θs that were

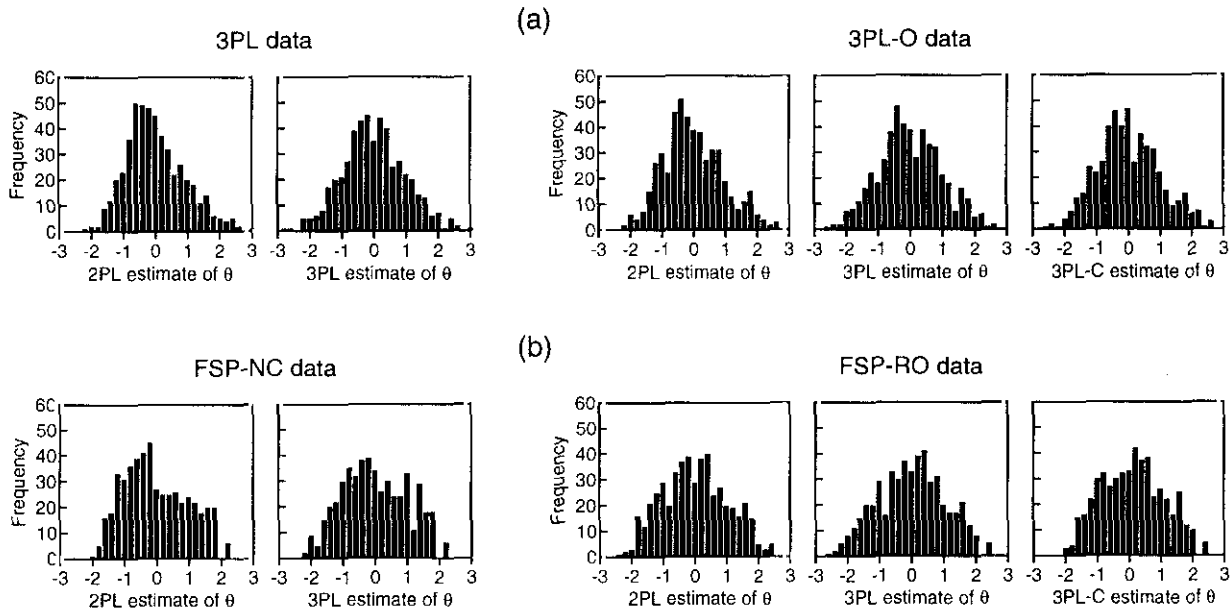

Figure 13. Histograms of the ability estimates obtained by the various models fitted to each data set. (a) Logistic data. (b) FSP data.

normally distributed. The inappropriateness of this conclusion shows in Figure 13b, which also reveals bell-shaped distributions of estimated logistic abilities despite the actual uniform distributions of the generating FSP λs!

Provided that some basic assumptions hold (see Rosenbaum, 1984), it is easy to understand how logistic models fit data so efficiently. IRT uses very powerful function-fitting procedures to find IRFs by searching for the best solution in the available parameter space. Logistic IRFs define a huge parameter space and do not incorporate any constraint (in the form of variables whose values are empirically measured independently). With so many degrees of freedom, no empirically-constrained independent variable, and a very powerful function-fitting algorithm, it is no wonder that solutions are found. Under these circumstances, *failure* to fit the data is what would be surprising.

This potential for fitting data derives from an observation by Lord and Novick (1968, p. 369), according to which "whenever any single [IRF] is a monotonic increasing function of θ, it is always possible and permissible to transform θ monotonically so that the characteristic curve becomes a normal ogive." Hence the power of normal ogive models (or logistic models, for that matter): the function-fitting algorithms implicitly apply the necessary transformation of the authentic ability scale in order to produce the θs that logistic models require. This process underlies the transformation of λs into θs when the logistic model is fitted to FSP data. The forfeit is that θ becomes only an ordinal measure bearing an unknown relation to the authentic ability.

Are Logistic Models Testable?

By fitting logistic models to test data, one merely determines whether parameters can be found that will make this general framework account for the data. The search for parameter estimates is made, using the procedures referred to in the previous section, in a way that maximizes a goal function which provides one of the possible sets of best-fitting parameter estimates.[8] Afterwards, the goodness of the fit is measured by making use of those estimates. This procedure involves a circularity (García-Pérez, 1994; García-Pérez & Frary, 1991a; Goldstein & Wood, 1989) which makes it unlikely that the goodness-of-fit test may result in a recommendation to reject the model.

Testing a model is different from fitting it (Marascuillo, 1988). It implies seeking independent empirical evidence supporting some model prediction. Fitting a model merely implies forcing the data into a theoretical scheme with the help of suitably chosen parameters. Fitting a model that has been successfully tested is legitimate, but fitting a model by fiat is not. In their present form, logistic models are not testable: they cannot make any testable prediction *before* parameter estimation. The models do not incorporate any observable independent variable, and only one observable dependent variable (a binary response to an item) is involved. Again, it is not surprising that presumed "tests" of the model using parameters that were estimated to fit the data confirm that the fit was actually obtained.

What Are the Consequences of Inappropriate Use of the 3PL Model?

Some misinterpretations that result from the failure to acknowledge the metaphoric nature of logistic models and their parameter space can easily be traced. For instance, when 3PL models are fitted to real test data, pseudo-chance level estimates are often found to have values far from the theoretical chance levels. This has been taken as evidence that examinees of low ability are attracted towards distractors in some items, thus performing below the guessing level, whereas, on other items, even examinees of low ability can eliminate some distractors, thus bringing the lower asymptotes above the guessing level (Lord, 1974). However, this interpretation sounds as appealing as it is unrealistic: a look at Figure 1 shows that this effect also occurs in a simulation in which there were no real items, distractors, or examinees to eliminate or be attracted to them, and where the true pseudo-chance levels were set at 0.25. Figure 8 shows the same pattern in a simulation involving a response process and a parameter space other than logistic. In view of this, the only sensible way to interpret these fluctuations is that they are an artifact of the fitting algorithm that does not reveal any underlying reality.[9]

Similarly unfounded conclusions were raised by Thissen and Steinberg (1984). Their model for multiple-choice items results in non-monotonic trace lines (IRFs) for the correct option in some items. This made them introduce the concept of "positive misinformation," which would permit low-ability examinees to give the right answer for the wrong reason. Again, the interpretation sounds reasonable and may imply a phenomenon that exists in reality but, instead of looking for direct empirical evidence of the actual phenomenon, the explanation was accepted at once without further ado in order to accommodate the outcomes of a model that the data was forced to fit by fiat.

Finally, Yen, Burket, and Sykes (1991) have shown that the likelihood equation for a non-trivial amount of real response vectors to multiple-choice items under the three-

[8] Alternative sets of estimates are obtained by changing the goal function in the parameter estimation method.

[9] The foregoing discussion does not deny the existence of partial information and misinformation. It simply claims that fluctuations of the estimates of c around the theoretical chance level have nothing to do with them.

parameter logistic model may have several local maxima of similar magnitude at widely different ability levels. The obvious consequence of this characteristic is a necessary introduction of uncertainty in ability estimation. Although only one of these will be a global maximum for any given response vector, they also showed that LOGIST did not always find this global maximum, and they were unable to determine the conditions under which the global maximum would be found. They acknowledged that this state of affairs is unsatisfactory, but they also explained away the problem by claiming that the occurrence of multiple maxima in the likelihood function for a particular response vector indicates that, rather than being inconsistent with the model, the response vector is consistent with the model in more than one way: that response vector might arise either from high ability or from low ability combined with successful guesses. Yet, no direct empirical evidence was provided to substantiate this interpretation.

The three examples just mentioned derive from an endemic feature of logistic models: the disregard for a representation of the response process. Partial information, misinformation, guessing, omitting, etc, are not properly represented in a model which simply states that the probability of a correct response to an item is given by a reasonable-looking function of ill-defined parameters. The situation is specially troublesome when the model is asked to fit data sets where guessing has been part of the response process, and where omissions may also have occurred. This situation led Baker (1987a, p. 135) to demand that "some effort should be devoted to developing a new model to cope with the issue of guessing." It is noteworthy how difficult it has proved to investigate and determine the effect of omissions or guessing on logistic fit. This is simply because there is no prescription as to how this should be done: there is no model-consistent way of introducing omissions or guessing in a data set. If this is done, it can be done in various ways (e.g., Lord, 1983; Mislevy & Bock, 1982; Wainer & Wright, 1980; Waller, 1989) each of which affects fit differently.

The simulations in Study 2 provide the basis for an assessment of the consequences of using the 3PL model when it does not hold (however well it fits). The losses associated with not using the adequate model are especially dramatic on theoretical grounds. According to Gulliksen (1961, p. 101), psychometric models should establish "the relation between the *ability* of the individual and his *observed score* on the test." But establishing that relationship implies an exercise in substantive theory and model building *before* any function is fitted to the data.

The first advantage of using a theoretically sound and empirically appropriate model is that claims to the effect that an examinee has an ability of, say, 0.7 would be meaningful. In logistic models, θ does not have units of measurement, nor is it related to any quantitative measure of knowledge or ability. Indeed, Lord (1975, p. 205) defined the ability scale as "the scale on which all item characteristic curves have some specified mathematical form, for example, logistic or normal ogive," thus expressing a clear disregard for the interpretability of θ. As a consequence, under the metric of logistic IRFs, ability estimates only reveal relative performance and, thus, that an examinee has an ability of 0.7 means, at the most, that his or her ability is greater than those of examinees obtaining lower ability estimates. But how much ability he or she has remains unknown.

Also, if the IRF embodied assumptions about the format of the test and the way it is administered, about examinee behavior, and about other characteristics of the testing situation, then these assumptions could be replaced to obtain IRFs applying to a variety of circumstances. As a result, Gulliksen's (1961, pp. 101-102) wish of being "able to say that, for certain specified tests constructed in this way, here is the relationship between the score and the ability measured, and this is the appropriate trace line to use" will be closer to becoming fulfilled. In addition, IRT methods could easily be used with items that are not binary scored. It would also be possible to determine theoretically what combination of these characteristics gives rise to more accurate ability estimates, thus providing a basis for advising in favor of or against certain testing practices. FSP models have indeed been successfully used for this purpose. For instance, García-Pérez (1989a) used FSP models to show that mastery decisions with any given practical degree of accuracy require dramatically different numbers of items depending on the format of administration of the test and the guessing behavior adopted by the examinees. Also, García-Pérez (1993) used FSP models to show that use of "none of the above" has the important advantage of reducing the size of the confidence intervals for maximum-likelihood estimation of λ, as compared to those of analogous conventional items with the same number of options. Of course, the extent to which these theoretical outcomes turn into advantages in empirical testing practice depends on the empirical validity of FSP models, an issue that will be commented on in the next section.

Is There an Alternative?

As discussed by García-Pérez and Frary (1991a), finite state theory produces IRFs that are free of the problems inherent to logistic functions. Potential benefits of using FSP IRFs in place of their logistic counterparts are also discussed there. For purposes of comparison with the picture of logistic IRFs that emerges from the simulations in Study 2 and the foregoing discussion, a few of the contrasting features of FSP IRFs will be mentioned here.

First, finite state models have mechanistic realism. They are built on parameters that are empirically meaningful, consider assumptions about how items are constructed, how tests are administered, and how examinees behave and, then, translate literally a description of test-taking behavior into mathematical terms. As a result, as many equations are

produced as there are response outcomes under the format of administration considered, each of which is interpretable on its own. For instance, in Equation 3a above, the probability of a correct response is expressed as the sum of the probabilities of all the situations that may lead the examinee to give the correct answer to the item of concern, from knowledge of the truth value of all options in the item (first addend in Equation 3a), through knowledge of three options (second addend), knowledge of two options, one of which is the correct answer (third addend), successful guess in case of knowledge of two options that are distractors (fourth addend), knowledge of only one option that turns out to be the correct answer (fifth addend), and successful guess in case of knowledge of only one distractor (sixth addend), to a successful guess under total ignorance (seventh addend). Equations 3b and 3c similarly embody the circumstances that may lead an examinee to mark a wrong option or omit the item. By doing so, FSP models incorporate realistically all the relevant concepts in test-taking behavior: total knowledge, partial knowledge, total ignorance, and guessing. García-Pérez and Frary (1991a) discuss how misinformation can be incorporated into finite state models, as well as how to use this framework for speeded tests.

Second, FSP IRFs are testable. Although, like their logistic counterparts, FSP IRFs hypothesize the relationship between correct responses on a test and a number of unobservable parameters, FSP models include additional equations for the relationships of these unobservable parameters with the remaining response outcomes under a given format of administration of a test. It is all these explicit relationships that allow deriving model predictions that can be tested without estimating model parameters. This provides the grounds for testing (and, then, accepting or rejecting) the models before searching for parameters that will maximize model fit. Empirical examples of FSP model testing can be found in García-Pérez (1987, 1990; see also García-Pérez & Frary, 1991b; Zin, 1992).

Finally, FSP models incorporate an interpretable definition of ability. This point was sufficiently illustrated when the models were introduced at the beginning of Study 2, and will not be further expanded upon here. It should be pointed out that FSP models do not include an item discrimination parameter, but there is no a priori reason why a psychometric model would be incomplete if it did not have one. That logistic functions with them fit data better than logistic functions without them is a result that is local to logistic functions. The adequacy of an alternative IRF is to be measured by its accomplishments, and not by how it compares conceptually with the IRFs for which it is an alternative.

A final question about the qualifications of FSP models as an alternative to logistic models is whether they are ready for use, especially in what regards parameter estimation methods. García-Pérez (1985, 1987, 1989b) and García-Pérez and Frary (1989) described simple analytical methods for the estimation of λ that proceed by consideration that all items have identical, average difficulty. García-Pérez (1993, 1994) described and studied alternative methods for the estimation of λ that make use of the same assumption (items of identical, average difficulty), but relying on the optimization of goal functions derived from the minimum-distance measures of

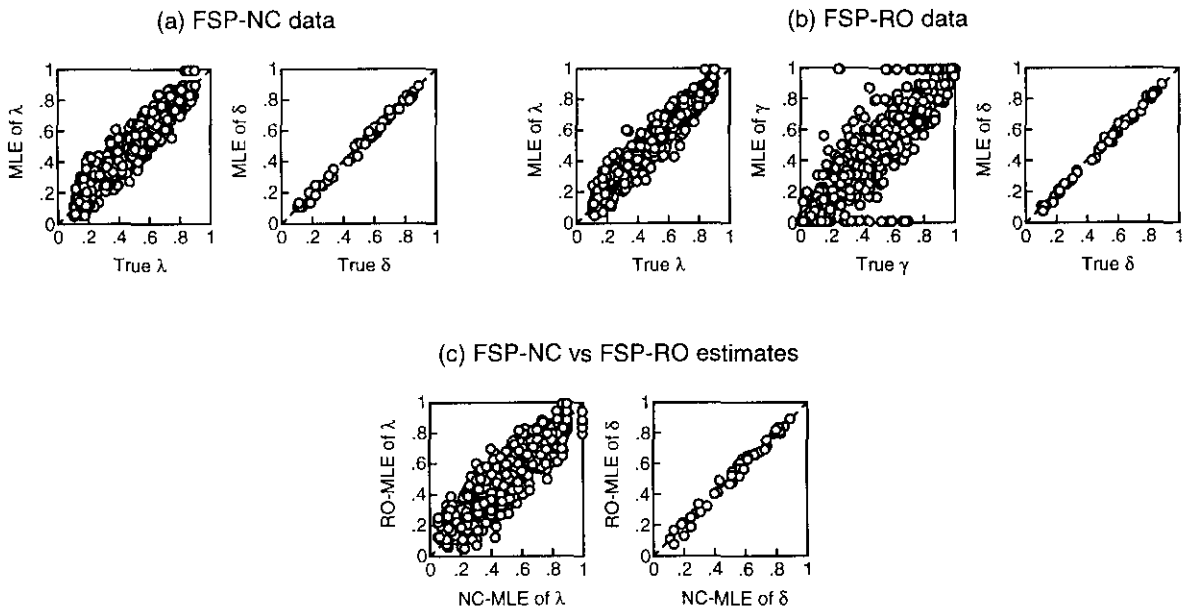

Figure 14. Relations between true and MLEs of FSP parameters in each data set [(a) for FSP-NC data; (b) for FSP-RO data], and between MLEs of λ and δ across data sets (c). Dashed diagonal lines indicate an expected identity relationship between the variables in the abscissa and the ordinate.

Cressie and Read (1984), which include the popular minimum chi-square and maximum likelihood methods. Application of these methods for the estimation of all relevant parameters (λ and δ from FSP-NC data and λ, γ, and δ from FSP-RO data) is straightforward, and Figures 14a and 14b show the relationships between true parameters and their maximum-likelihood estimates (MLEs) for the FSP-NC and FSP-RO data sets from the simulations in Study 2. Note that MLEs of δ are much more accurate than MLEs of λ (and γ, where applicable), but this is only a result of the fact that estimates of δ are each based on responses from 500 examinees, whereas estimates of λ are each based on responses to 50 items (i.e., a factor of ten fewer data). The estimation of γ is further hampered because opportunities to guess are scarce for medium- and high-ability examinees and, therefore, random variations dominate the data on which the estimation of γ is based. This noise does not affect MLEs of λ, which can be seen to be equally linearly related to true λ when there are omissions (Figure 14b; $r = .95$) and when all items are answered (Figure 14a; $r = .96$). Finally, Figure 14c shows that MLEs of λ from FSP-NC vs FSP-RO data are less related to one another ($r = .90$) than either of them is to true λ.

References

Albanese, M.A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement, 25*, 149-157.

Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

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

Baker, F.B. (1987b). Item parameter estimation via minimum logit chi-square. *British Journal of Mathematical and Statistical Psychology, 40*, 50-60.

Baker, F.B. (1991). Comparison of minimum logit chi-square and Bayesian item parameter estimation. *British Journal of Mathematical and Statistical Psychology, 44*, 299-313.

Baker, F.B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.

Bejar, I.I. (1983). Introduction to item response models and their assumptions. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 1-23). Vancouver, BC: Educational Research Institute of British Columbia.

Blinkhorn, S.F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50*, 175-185.

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement, 17*, 147-153.

Cressie, N., & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B, 46*, 440-464.

Cross, L.H., & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests. *Journal of Educational Measurement, 14*, 313-321.

De Ayala, R.J. (1992). The influence of dimensionality on CAT ability estimation. *Educational and Psychological Measurement, 52*, 513-528.

Dinero, T.E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 1*, 581-592.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5*, 175-186.

Freedman, D.A. (1985). Statistics and the scientific method. In W.M. Mason & S.E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem* (pp. 343-366). New York: Springer-Verlag.

García-Pérez, M.A. (1985). A finite state theory of performance in multiple-choice tests. In E. Terouanne (Ed.), *Proceedings of the 16th European mathematical psychology group meeting* (pp. 55-67). Montpellier: European Mathematical Psychology Group.

García-Pérez, M.A. (1987). A finite state theory of performance in multiple-choice tests. In E.E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology-1* (pp. 455-464). Amsterdam: Elsevier.

García-Pérez, M.A. (1989a). Item sampling, guessing, partial information and decision-making in achievement testing. In E.E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 249-265). Berlin: Springer-Verlag.

García-Pérez, M.A. (1989b). La corrección del azar en pruebas objetivas: un enfoque basado en una nueva teoría de estados finitos. *Investigaciones Psicológicas, 6*, 33-62.

García-Pérez, M.A. (1990). A comparison of two models of performance in objective tests: Finite states versus continuous distributions. *British Journal of Mathematical and Statistical Psychology, 43*, 73-91.

García-Pérez, M.A. (1993). In defence of 'none of the above.' *British Journal of Mathematical and Statistical Psychology, 46*, 213-229.

García-Pérez, M.A. (1994). Parameter estimation and goodness-of-fit testing in multinomial models. *British Journal of Mathematical and Statistical Psychology, 47*, 247-282.

García-Pérez, M.A., & Frary, R.B. (1989). Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study. *Applied Psychological Measurement, 13*, 403-417.

García-Pérez, M.A., & Frary, R.B. (1991a). Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology, 44*, 45-73.

García-Pérez, M.A., & Frary, R.B. (1991b). Testing finite state models of performance in objective tests using items with 'none of the above' as an option. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 273-291). New York: Springer-Verlag.

Gifford, J.A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*, 33-43.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5*, 211-220.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika, 26*, 93-107.

Hambleton, R.K. (1983). Application of item response models to criterion-referenced assessment. *Applied Psychological Measurement, 7*, 33-44.

Hambleton, R.K., & Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). New York: Academic Press.

Hambleton, R.K., & Murray, L.N. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 71-94). Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91-115.

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

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

Jannarone, R.J., Yu, K.F., & Laughlin, J.E. (1990). Easy Bayes estimation for Rasch-type models. *Psychometrika, 55*, 449-460.

Kim, J.K., & Nicewander, W.A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587-599.

Kim, S.-H., Cohen, A.S., Baker, F.B., Subkoviak, M.J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59*, 405-421.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

Lord, F.M. (1975). The 'ability' scale in item characteristic curve theory. *Psychometrika, 40*, 205-217.

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

Lord, F.M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48*, 477-482.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Marascuilo, L.A. (1988). Introduction to model building and rank tests. *Contemporary Psychology, 33*, 794-795.

McKinley, R.L., & Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Mislevy, R.J. (1987). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Mislevy, R.J., & Bock R.D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725-737.

Mislevy, R.J., & Bock, R.D. (1984). *BILOG Version 2.2: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Mislevy, R.J., & Bock, R.D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models. 1986 edition*. 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.

Mislevy, R.J., & Verhelst, N. (1987). *Modeling item responses when different subjects employ different solution strategies*. Research Report RR-87-47-ONR. Princeton, NJ: Educational Testing Service.

Ramsay, J.O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association, 84*, 906-915.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Rec, M.J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*, 371-385.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Rowley, G.L., & Traub, R.E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*, 15-22.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. *Applied Psychological Measurement, 13*, 391-402.

Slakter, M. (1968). The penalty for not guessing. *Journal of Educational Measurement, 5*, 141-144.

Swaminathan, H., & Gifford, J.A. (1983). Estimation of parameters in the three-parameter latent trait model. In D.J. Weiss (Ed.). *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13-30). New York: Academic Press.

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

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Tsutakawa, R.K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology, 45*, 51-74.

Tsutakawa, R.K., & Johnson, J.C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.

Tsutakawa, R.K., & Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251-267.

Tsutakawa, R.K., & Soltys, M.J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13*, 117-130.

Vale, C.D., & Gialluca, K.A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement, 12*, 53-67.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339-368.

Wainer, H., & Wright, B.D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45*, 373-391.

Waller, M.I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*, 233-243.

Wang, T., & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Warm, A.W. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika, 54*, 427-450.

Weiss, D.J., & Yoes, M.E. (1991). Item response theory. In R.K. Hambleton & J.N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 69-95). Boston, MA: Kluwer.

Weitzman, R.A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement, 56*, 779-790.

Wichmann, B.A., & Hill, I.D. (1982). Algorithm AS 183. An efficient and portable pseudo-random number generator. *Applied Statistics, 31*, 188-190.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST 5.0 version 1.0 users' guide*. Princeton, NJ: Educational Testing Service.

Wood, R. (1978). Fitting the Rasch model–A heady tale. *British Journal of Mathematical and Statistical Psychology, 31*, 27-32.

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

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

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

Yen, W.M., Burket, G.R., & Sykes, R.C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika, 56*, 39-54.

Zeng, L. (1997). Implementation of marginal Bayesian estimation with four-parameter beta prior distributions. *Applied Psychological Measurement, 21*, 143-156.

Zin, T.T. (1992). *Comparing 12 finite state models of examinee performance on multiple-choice tests*. Ph.D. Dissertation. Virginia Polytechnic Institute and State University.

Received December 21, 1998
Revision received January 20, 1999
Accepted March 3, 1999