Modeling Guessing Behavior: A Comparison of Two IRT Models

Michael I. Waller
University of Wisconsin—Milwaukee

This study compared the fit of the three-parameter model to that of the Ability Removing Random Guessing (ARRG) model (Waller, 1973) on data from a wide range of tests of cognitive ability in three representative samples. Although both models were designed to remove only the effects of random guessing, the results of this study indicated that the three-parameter model also makes an adjustment for partial-knowledge guessing. Fit of the three-parameter model with guessing parameters estimated at a constant value of 1 divided by the number of alternatives was compared to fit with individually estimated guessing parameters. The latter were found to produce fit far superior to those estimated at a constant value. A solution to the convergence problems often encountered with the three-parameter model is discussed.
Index terms: Ability Removing Random Guessing model, convergence in three-parameter estimation procedures, item response theory, maximum likelihood estimation, partial-knowledge guessing, random guessing.

The three-parameter model (Birnbaum, 1968) and the Ability Removing Random Guessing (ARRG) model (Waller, 1973) are two versions of the logistic form of the free-response, two-parameter item response theory (IRT) model that were designed to deal with the problem of random guessing. This study compared them empirically using goodness-of-fit tests that compared the fit of each model to a variety of cognitive data from a nationally known battery of achievement tests on a representative sample of elementary, middle school, and high school students.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 3, September 1989, pp. 233–243
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/030233-11$1.80

The Models

The free-response logistic IRT model is given by Equation 1, where P_{ij} represents the probability that person i with ability θ_i responds correctly to item j:

$$P_{ij} = \frac{1}{1+\exp[a_j(b_j-\theta_i)]} \quad , \tag{1}$$

where b_j and a_j are item difficulty and discrimination, respectively. The most commonly used modification is Birnbaum's (1968) familiar three-parameter model:

$$P_{ij} = c_j + \frac{1-c_j}{1+\exp[a_j(b_j-\theta_i)]} \quad , \tag{2}$$

where c_j represents the item guessing parameter or lower asymptote.

As can be seen from Equation 2, the three-parameter model adds to the free-response model a parameter associated with each item. Birnbaum designed this model to adjust the free-response model to handle the situation in which examinees either guess totally randomly or answer on the basis of their knowledge. That is, the model is viewed as ignoring the common situation of a person answering on the basis of partial knowledge (i.e.,

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

answering by eliminating one or more of the alternatives and selecting randomly from among the remainder; Birnbaum, 1968).

The ARRG model was also designed to remove the effects of essentially random guessing from parameter estimates. However, rather than limit the modification to characteristics of the item as in the three-parameter model, the ARRG model provides a more individualized approach by focusing on the interaction between the person and the item. The adjustment made by the ARRG model is based on the idea that a person who guesses in an essentially random manner does so only on those items that are very difficult for him or her. These item-person interactions are then simply omitted during estimation of θ, so that whether a person guesses on any particular item does not affect estimation of that person's θ. This is accomplished by focusing on two characteristics inherent in all IRT models:

1. For each person, and for each item to which that person responds, an estimate of the probability of a correct response can be routinely obtained.
2. Given a set of items that fit such a model, θ estimates can be obtained on a common metric using *any* subset of such items.

The first characteristic permits identification of items that are very difficult for any particular person and therefore may attract essentially random guessing. The second characteristic allows elimination of such items from the θ estimate for each person while retaining the common metric for all resulting θ estimates even though different subsets of items may be used for different examinees.

The ARRG model uses these two characteristics by assuming, as indicated above, that an examinee guessing in an essentially random manner does so only on items with a low probability of correct response, that is, lower than the chance level of the test. Therefore, during each step of the θ estimation procedure, the ARRG model estimates the probability of a correct response for each examinee's response to each item and omits those responses for which the estimated probability of correct response is sufficiently low. The determination of an exact value below which responses are to be omitted is made empirically, constrained only by the requirement that this cutoff value must be less than or equal to the chance level of the test.

Waller (1976) applied these ideas to the Rasch model. In doing so he showed how to obtain estimates of θ free of the effects of random guessing while maintaining the Rasch model's unique characteristics. Choppin (1985) also applied some of these ideas to the Rasch model, but the details of his model are quite different. In particular, Choppin's "Rasch" model, adjusted for random guessing, estimates a different cutoff parameter for each item rather than a single cutoff for all the items. Of course, such a two-parameter model is not a true Rasch model in the sense that the raw score is not a sufficient statistic for θ estimation.

In the present application, the ARRG model consists of the well-known two-parameter, free-response model presented in Equation 1 modified to include only those interactions between an item and a person for which essentially random guessing is unlikely to occur. The model can be characterized by

$$P_{ij} = \frac{1}{1+\exp[a_j(b_j-\theta_i)]} \quad \text{for } P_{ij} > P_c \quad , \qquad (3)$$

where the ARRG cutoff value P_c is less than or equal to $1/A_j$, and A_j is the number of alternatives for item j.

For each person the ARRG model essentially divides the items into two groups: those items for which $P_{ij} > P_c$, and those items for which $P_{ij} \leq P_c$. At each step of the estimation procedure, the ARRG model uses *only* those items with a reasonable probability of a correct response from factors other than essentially random guessing (i.e., items from the first of these two groups) to estimate θ for a person. This means that if Newton-Raphson iteration is used for estimation of each person's θ_i, then every time the first and second derivatives are accumulated, only responses from the first group of items are included in each sum.

Because ARRG θ estimates are based on fewer items than are two-parameter model estimates, it is reasonable to suspect that the resulting θ estimates may be less precise. However, the resulting

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

estimated precision has been found to increase (Waller, 1973) because the noise caused by the random responses present in the data is removed during estimation of the item parameters.[1]

Estimation of the item parameters is accomplished under a similar but slightly different procedure represented by

$$P_{ij} = \begin{cases} \dfrac{1}{1+\exp[a_j(b_j-\theta_i)]} & \text{for } P_{ij} > P_c \\ 0 & \text{for } P_{ij} \leqslant P_c \end{cases} \qquad (4)$$

That is, during estimation of item parameters under the ARRG model, responses that have a small probability of being answered correctly on the basis of an examinee's θ are not omitted, but rather are assumed to be incorrect. This means, for example, that if Newton-Raphson iteration is used for joint estimation of the item parameters, then during accumulation of the first and second derivatives, the proportion correct for each θ group has been adjusted so that each examinee's response in which $P_{ij} < P_c$ is counted as an incorrect response. This procedure allows for empirical determination of a best value for P_c, as described below.

The primary focus of this study was to compare the behavior of the three-parameter and ARRG models empirically. The ARRG model might be expected to produce the better fit because it individualizes the adjustment for random guessing by examining each item-person interaction, whereas the three-parameter model produces one estimate of each item's guessing parameter c_j and uses these values for every person's responses. However, the results of this study indicate that this view is too simplistic.

Several issues that naturally arise when using the three-parameter model are also addressed in this study. First, there are two questions concerning estimation of the guessing parameter: what method of estimation will be used, and the related question of whether to estimate these parameters individually or at one constant value. Another issue concerns problems with convergence during estimation of all parameters of the three-parameter model. Samejima (1973) has shown that for some response vectors, maximum likelihood estimates of θ are not necessarily unique and may not exist at finite values. Similar convergence problems have been found to exist during estimation of the item parameters. The results of this study yield some suggestions for solving these convergence problems.

Instruments and Data

The measures used were the different content area examinations that comprise the Comprehensive Tests of Basic Skills (CTBS; CTB McGraw-Hill, 1974), Form S, Levels 2, 3, and 4. This battery of tests provides measures of the following cognitive abilities: reading vocabulary, reading comprehension, spelling, mechanics of language, language expression, mathematical computation, mathematical concepts, mathematical applications, reference skills, science, and social studies.

The examinees consisted of samples from three cohorts of children obtained by stratified random sampling representative of the population of children in three grades from a southeastern state; each sample size was just under 1,000. The three cohorts of children were elementary school (Grade 4), middle school (Grade 7), and high school (Grade 10). Measuring 11 content areas for each of these cohorts yielded a total of 33 different examinations on which to compare the models.

A second independent sample was taken from each of the three cohorts for the purpose of replication of the results on the entire set of 33 comparisons.

[1]This is clarified in Waller (1973) in the following manner: Consider estimates of the accuracy of θ estimates, $\hat{\sigma}_\theta$. The value of $\hat{\sigma}_\theta$ will be estimated by the negative of the square root of the reciprocal of the information function. For the θ parameter in the two-parameter model, this is given by the reciprocal of the second derivative of the log-likelihood function: $\hat{\sigma}_\theta = [-1/(-\sum a_j^2 P_{vj} Q_{vj})]^{1/2}$. Clearly any decrease in the values of the estimated discrimination parameters, a_j, will result in a corresponding increase in the estimated standard deviation. Because guessing responses at low θ levels will reduce the estimated value of a_j for any item that attracts random guessing, removing these guessing responses from the estimation procedure will result in a corresponding increase in the estimated value of a_j, and a corresponding decrease in the estimated standard deviation of the θ estimates.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Method

The computer program used was an augmented version of LOGOG (Kolakowski & Bock, 1973). In the procedure used by this program, joint estimation of the two item parameters common to both models (difficulty and discrimination) is accomplished through a quasi-marginal maximum likelihood procedure. This procedure does not employ EM estimation (Bock & Aitkin, 1981), but rather simply assumes a particular distribution of abilities. A normal distribution of θ has been assumed throughout this study.

In other words, for estimation of difficulty and discrimination for all analyses performed in this study, the examinees were rank ordered and distributed into 10 score groups or fractiles, where the number in each fractile was chosen to reflect a normal distribution. This assumption of a prior distribution of θ made during estimation of the item parameters allows for estimation of all parameters of the two-parameter model without any problems of convergence.[2]

For the three-parameter model, the guessing parameter or lower asymptote for each item is estimated separately and remains constant throughout the entire maximum likelihood estimation procedure. Maximum likelihood estimation of this parameter has proven to be less than satisfactory in that the convergence problems referred to above are magnified when this additional parameter is included in maximum likelihood estimation, as in the LOGIST computer program.

At least two methods for estimating the lower asymptote parameter are currently in use. The simplest and perhaps most frequently used method can be called the *constant* method. In the constant method the guessing parameter is set to the same value for all items, $1/A$, where A is the number of alternatives. The second method can be called the *individualized* method. In this method the value of c_j for each item is obtained by ranking the examinees, distributing them into some number of score groups, and then estimating the lower asymptote graphically by examining the proportion correct in each of these score groups. The individualized method with 10 score groups was used in this study for the comparisons between the two models. A second set of analyses using the constant method was also performed, and the resulting fits were compared to the three-parameter model using the individualized method and to the ARRG model.

For the ARRG model, one value of P_c was used for each content area examination in each cohort. The estimation procedure for this parameter is a gross minimum goodness-of-fit procedure. Values for P_c are chosen from 0 to $1/A$ in steps of .05; for example, for a CTBS-type examination consisting of four-choice items, the trial values would be .00, .05, .10, .15, .20, and .25. For each value of P_c a complete item analysis and test of fit is performed. The final estimate of P_c is that value yielding the minimum goodness-of-fit value used to examine the fit of the instrument as a whole.

If there is random guessing present in the data, resetting very low-probability responses to incorrect (as is done during estimation of the item parameters) and examining the fit of the model can result in an improvement in fit as compared to the free-response model's fit. This occurs because as P_c gradually increases across analyses, the number of responses that are due to random guessing (and therefore appropriately set to incorrect) also increases. In other words, with values of P_c that are too low, the proportion of correct responses in low-θ groups is higher than would be found under the free-response model, and as P_c is gradually increased these correct responses that are due to random guessing are appropriately set to incorrect. However, as P_c continues to increase, the fit is seen to deteriorate. This occurs because when P_c is too large, the proportion of correct responses in low- to middle-θ groups is lower than would be found under the free-response model.

The test statistic used in the LOGOG computer program is the Bock (1972) chi-square goodness-

[2]The prior assumption of a normal distribution of θ may introduce some bias into the estimated values of these parameters in cases where the distribution of θ is sufficiently skewed. However, it seems reasonable to assume that the distribution of θ in many of these cognitive abilities is nearly normal, and that any resulting bias in these estimated values would affect both models in essentially the same manner and can therefore be ignored in the comparisons made in this study.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

of-fit statistic. The cells required for this statistic are formed during estimation of the item parameters as described above, that is, by ranking the examinees by θ and then dividing the sample into 10 fractiles, where the number of examinees in each group is chosen to reflect a normal distribution of θ. The assumption of normality does not affect the validity of this test statistic in the present context. Although the distribution of this statistic may be chi-square only under certain circumstances (Yen, 1981), the *p* values that would result are not necessary for this study because comparing the relative fit of the two models was the primary concern rather than examining their fit in a formal hypothesis-testing framework. Accordingly, this statistic is referred to as chi-square without any reference to *p* values.

One comparison of fit was made for each content area examination for each of the three cohorts. The number of degrees of freedom used to examine the fit of each item of each examination is equal to the number of cells (10) minus the number of parameters estimated, yielding 7 for the three-parameter model and 8 for the ARRG model. The value of the Bock chi-square for each item was summed to obtain a value of chi-square that may be used for examining the fit of each of the two models for each content area examination as a whole. The number of degrees of freedom in this case is the sum of the degrees of freedom for each item minus two. These two additional degrees of freedom are lost because the mean and standard deviation of the distribution of θ are assigned values of 0 and 1, respectively (Kolakowski & Bock, 1973). For the ARRG procedure, an additional degree of freedom is also subtracted when examining the test of fit of each instrument as a whole because this procedure requires estimation of one cutoff value P_c for each examination.

Results and Discussion

Comparison of the Two Models

Results. The results of the 33 comparisons between the models are presented in Table 1 in terms of the ratio of chi-square to degrees of freedom. Table 1 presents the results of the analyses of the first group of datasets, one dataset for each cohort, as well as the results for the replication datasets.

The results using the first group of datasets indicate that in 61% of these examinations the three-parameter model with individually estimated guessing parameters produced the best fit to the data (20/33). The free-response model produced the best fit in 21% (7/33). The ARRG model produced the best fit in the remaining 18% (6/33).

In 17 of the 33 examinations, the results using the replication datasets are the same as the results for the original datasets in the sense that the same model produced the best fit. In this subset of 17 examinations the three-parameter model fit best in 76% (13/17), the free-response model fit best in 18% (3/17), and the ARRG model fit best in the remaining 6% (1/17). Considering the entire set of 33 examinations in the replication datasets, the three-parameter model fit best in 48% (16/33), the free-response model fit best in 30% (10/33), and the ARRG model produced the best fit in 21% (7/33).

Discussion. The fact that the two groups of datasets produced consistent results in only 16 of the 33 possible examinations indicates that the comparison procedures used here are not accurate enough for precise conclusions to be drawn concerning particular examinations. However, three general conclusions can be appropriately drawn from these results:

1. The three-parameter model can be expected to produce better fit to more than twice as many datasets as the ARRG model (in this study, to 55% of the datasets).
2. The free-response model can be expected to produce better fit than either guessing model for a significant proportion of datasets (here, the figure was approximately 25% of the datasets).
3. The ARRG model can be expected to produce the best fit to only a small proportion of datasets (about 20% of those examined in this study).

Modeling Partial-Knowledge Guessing

These results clearly indicate that the individualization accomplished by the ARRG model did

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Goodness-of-Fit Ratios (χ^2/df) for the Three-Parameter (3P) and ARRG Models and Numbers of Items (k) by Cohort for Original and Replication Datasets

Content Area	Cohort: Elementary (4th) k	3P	ARRG	P_c	Middle (7th) k	3P	ARRG	P_c	High (10th) k	3P	ARRG	P_c
Original Datasets												
Reading												
Vocabulary	40	2.03*	2.58	.15	40	1.56*	1.89	0	40	1.67*	2.41	.15
Comprehension	45	2.33*	2.60	0	45	1.63	1.60*	.05	45	1.80	1.70*	.05
Language												
Spelling	50	1.86*	1.87	.10	30	1.25*	1.56	0	30	1.21*	1.40	0
Mechanics	20	1.90	1.63*	.05	20	1.39	1.39	0**	20	2.07	1.71	0**
Expression	35	1.88*	2.08	0	35	1.12*	1.28	0	35	2.15	1.81*	.15
Mathematics												
Computation	48	1.59*	1.65	.05	48	2.14*	2.63	.15	48	1.48*	1.68	0
Concepts	25	2.90	2.90	0**	25	1.51	1.43	0**	25	1.48*	1.51	0
Application	25	1.59*	1.63	0	25	1.81*	2.44	.05	25	1.63*	1.88	.15
Reference Skills	20	1.89	1.75*	.05	20	1.67	1.50	0**	20	2.22	1.93	0**
Science	36	2.02	1.90	0**	41	1.76	1.70*	.05	40	1.42*	1.61	.05
Social Studies	37	1.59*	1.88	.05	40	1.71*	1.87	0	39	1.58*	1.81	.05
Replication Datasets												
Reading												
Vocabulary	40	2.75	2.52*	.05	40	1.78*	1.99	0	40	2.21*	2.65	0
Comprehension	45	1.74	1.72*	.15	45	1.91	1.85	0**	45	1.75*	1.95	0
Language												
Spelling	50	2.06*	2.19	.05	30	2.39	2.07	0**	30	1.70	1.69	0**
Mechanics	20	1.79	1.58*	.05	20	2.29	1.59	0**	20	2.37	1.95	0**
Expression	35	1.88*	1.99	0	35	1.17*	1.45	0	35	2.19	1.78	0**
Mathematics												
Computation	48	1.74	1.72*	.05	48	1.83*	2.25	.05	48	1.77*	1.98	0
Concepts	25	2.64	2.19*	.15	25	1.41	1.18*	.05	25	2.02	1.83	0**
Application	25	1.84	1.82	0**	25	2.12*	2.41	0	25	2.11*	2.17	.15
Reference Skills	20	1.81	1.56	0**	20	1.76*	1.82	0	20	1.97	1.67	0**
Science	36	1.87	1.52*	.15	41	1.65*	1.69	.05	40	1.55*	1.61	.05
Social Studies	37	1.71*	1.97	.05	40	1.66*	1.84	.05	39	1.58*	1.81	0

*Indicates the model with the better fit.
**Indicates that the free-response model produced the best fit.

not fulfill the expectation that this model would tend to produce a better fit to empirical data than would the three-parameter model. A thorough examination of the rationale underlying the two models with respect to their treatment of other types of guessing, in particular partial-knowledge guessing, provides an interesting interpretation of these results. Perhaps more importantly, this examination reveals that the three-parameter model can be interpreted as modeling partial-knowledge guessing behavior as well as random guessing behavior.

Item responses may be divided into three groups: those arising from random guessing behavior, those arising from knowledge or the lack thereof (e.g., being misled by an attractive distractor), and those based on partial-knowledge guessing as described above. The following interpretation of these results assumes that if a response is due at least in part to guessing behavior, it is reasonable to attribute such a response to essentially random guessing when $P_{ij} \leq 1/A_j$, and to partial-knowledge guessing when $P_{ij} > 1/A_j$. Although partial-knowledge guessing behavior may have played a part in producing a response when $P_{ij} \leq 1/A_j$, there is no practical reason for differentiating the two types of guessing behavior here in the chance-level range of probability.

Note that because the ARRG model has no effect

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

on item responses for which the probability of a correct response is greater than $1/A_j$, it has no effect on responses that might be due purely to partial-knowledge guessing behavior. The view that the three-parameter model also ignores partial-knowledge guessing behavior can be argued as follows: If P_{ij} in Equation 1 is replaced by F_{ij}, F_{ij} can then be viewed as representing the probability that examinee i knows the correct answer to item j; hence the three-parameter model of Equation 2 may be written, after some algebraic manipulation, as

$$P_{ij} = F_{ij}(1) + (1 - F_{ij})c_j \quad . \tag{5}$$

This equation can be viewed as dividing the act of a person responding correctly to an item into two mutually exclusive events whose probabilities sum to give the probability of a correct response. The first term represents the probability of a correct response when the examinee knows the correct answer, while the second term represents this probability when the examinee knows nothing concerning the correct answer, and guesses randomly.

Both terms in Equation 5 use the defining equation for conditional probability of event A given event B (Feller, 1968) for each of the two mutually exclusive events in the equation:

$$P(\text{A,B}) = [P(\text{B})][P(\text{A given B})] \quad . \tag{6}$$

That is, the right side of Equation 5 may be interpreted as the probability that the examinee knows the correct answer (F_{ij}) multiplied by the probability of a correct response given knowledge of the answer (1.0), plus the probability that the examinee *does not* know the correct answer ($1 - F_{ij}$) multiplied by the probability of a correct response to the item by guessing randomly, given lack of knowledge of the correct answer (c_j). It is this view of the three-parameter model that has been interpreted to mean that partial-knowledge guessing is ignored by this model.

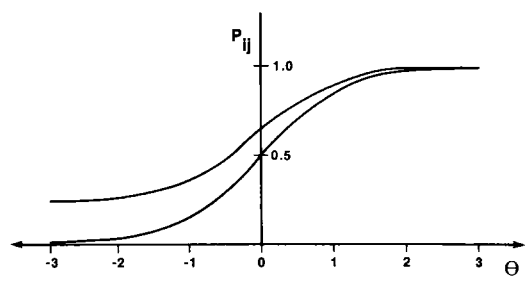
This view focuses on the fact that c_j is a guessing probability that is constant and independent of θ. Although this is true, a comparison of two- and three-parameter ogives clearly reveals that the effect of raising the lower asymptote by inclusion of a positive c_j in the model has a differential effect on P_{ij} depending on the examinee's position on the θ continuum.

Figure 1 presents two ogives for an item of middle difficulty ($b_j = 0.0$) and discrimination ($a_j = 1.7$). The upper curve represents the three-parameter model with guessing parameter simulating a four-choice item ($c_j = .25$). The lower curve represents the free-response, two-parameter model for the same item.

Proceeding along the curve, starting at the lowest θs and proceeding to middle and higher θs, it can be seen that in comparison to the free-response model of Equation 1, P_{ij} is raised in these portions of the curve as well. In fact, P_{ij} is raised all along the curve with the increment becoming smaller as θ increases. Only the asymptote represents the probability of a correct response from essentially random guessing.

Although the effect on P_{ij} of raising the lower asymptote in this middle part of the θ continuum has not been the focus of much discussion, the increase in P_{ij} caused by the three-parameter model all along the θ continuum can be viewed as representing the probability of a correct response due to partial-knowledge or random guessing. That is, for any specific θ level the probability dimension between 0 and 1 can be viewed as being composed of three intervals. The distance between a probability of 0 and the probability under the two-parameter model, ${}_2P_{ij}$, represents the probability of a correct response assuming that the examinee knows the correct answer. The distance between ${}_2P_{ij}$ and the probability represented by the three-parameter model, ${}_3P_{ij}$, represents the probability of a correct response by partial-knowledge or random guessing;

Figure 1
Two-Parameter and Three-Parameter Ogives for a Four-Choice Item of Medium Difficulty



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

the distance between ${}_3P_{ij}$ and a probability of 1 represents the probability of an incorrect response.

Stating this in the notation of Equation 5, the probability of a correct response, P_{ij}, may be viewed as consisting of two mutually exclusive parts as in Equation 7:

$$P_{ij} = {}_2P_{ij} + {}_3P_{ij} \tag{7}$$

where ${}_2P_{ij} = (F_{ij})1$ and ${}_3P_{ij} = (1 - F_{ij})g_{ij}$.

The new parameter g_{ij} represents the probability of a correct response of examinee i to item j in the event of any form of guessing behavior, and its estimation is not routinely required.

Because the ARRG model only affects responses in which the probability of a correct response is clearly below the chance level of an item, the comparisons made in this study allow identification of four types of datasets. This can be accomplished through judicious use of P_c and an examination of the relative fit of the two models. The four types are

1. Datasets that contain essentially no guessing responses (free-response datasets);
2. Datasets that contain only partial-knowledge guessing responses, as defined above;
3. Datasets that contain only essentially random guessing responses, as defined above; and
4. Datasets that contain both random guessing and partial-knowledge guessing responses.

Of course, as indicated by the replication datasets, the results of such analyses are only general in nature and must await more precise goodness-of-fit procedures before they can be applied to any individual dataset.

The first and second types of datasets contain no essentially random guessing of the kind defined above. Such datasets are identified by the fact that the fit of the ARRG model *degenerates* as soon as the value of P_c is raised from 0. After identifying such datasets, these two types can be differentiated by comparing the fit of the ARRG model with $P_c = 0$ to the fit of the three-parameter model. When the ARRG model produces the better fit, a dataset can be said to have no guessing responses of any kind. When the three-parameter model yields the better fit, a dataset can be said to contain some other type of item responses, namely partial-knowledge guessing responses.

The third and fourth types of datasets contain some amount of essentially random guessing responses. Such datasets are identified by the fact that the fit of the ARRG model *improves* as the value of P_c increases from 0. Once identified, these two types can be differentiated by comparing the fit of the best-fitting ARRG model with $P_c > 0$ to the fit of the three-parameter model. When the ARRG model produces the better fit, a dataset can be said to have only essentially random guessing responses. When the three-parameter model produces the better fit, a dataset can be said to have both random guessing responses and partial-knowledge guessing responses.

For the purpose of general conclusions, each of the 66 examinations in this study can be considered as a different dataset. The 36 datasets in which the three-parameter model produced the best fit as compared to the ARRG model can be said to contain responses that are due at least in part to partial-knowledge guessing behavior. The best-fitting ARRG analysis occurred with $P_c = 0$ in 19 of these datasets, indicating no essentially random guessing responses present in these data. In the other 17 datasets the best-fitting ARRG analysis occurred with $P_c > 0$, indicating that these datasets contained both random and partial-knowledge guessing responses. These results suggest that the three-parameter model does take partial-knowledge guessing into account.

The ARRG model produced the better fit in 30 datasets, including for this discussion the 17 in which the best-fitting value for P_c was 0. These datasets have heretofore been labeled "free response" because the fact that $P_c = 0$ in a dataset for which the ARRG model fits better than the three-parameter model indicates no guessing responses of any kind (at least any kind modeled by either of these models). In the remaining 13 datasets the best-fitting analysis occurred with $P_c > 0$, indicating that the only guessing responses were due to essentially random guessing.

Estimating the Guessing Parameter

Results. A second set of results yielded by this study is a comparison of the fit of the three-parameter model using the constant method for estimating

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

the guessing parameter of all items as opposed to the three-parameter model using the individualized method to estimate guessing parameters. The results from the original and replication datasets are presented in Table 2 in terms of the ratio of chi-square to degrees of freedom.

The results of the analyses of the original datasets as presented in Table 2 are that in 82% (27 of 33 examinations), the three-parameter model using the individualized method produced a better fit to the data than did the three-parameter model using the constant method. Comparing Table 2 to Table 1 reveals that the ARRG model also produced a better fit to these data than the three-parameter model using the constant method in 75% of 32 of these examinations (in one examination, both produced the same fit). The results of the replication datasets concerning these questions are essentially the same: In 88% (29 of 33 examinations), the three-parameter model using the individualized method produced a better fit to the data than did the three-parameter model using the constant method.

Table 2
Goodness-of-Fit Ratios (χ^2/df) for the Three-Parameter Model and Numbers of Items (k) with Individualized and Fixed Guessing Parameters by Cohort, for Original and Replication Datasets

Content Area	Cohort: Elementary (4th) k	Ind	Fixed	Middle (7th) k	Ind	Fixed	High (10th) k	Ind	Fixed
Original Datasets									
Reading									
Vocabulary	40	2.03*	2.29	40	1.56*	1.89	40	1.67*	2.04
Comprehension	45	2.33	2.17*	45	1.63*	3.02	45	1.80*	1.89
Language									
Spelling	50	1.86	1.79*	30	1.25	1.23*	30	1.21*	6.99
Mechanics	20	1.90*	1.95	20	1.39*	1.88	20	2.07	1.87*
Expression	35	1.88*	2.14	35	1.12*	2.01	35	2.15	2.05*
Mathematics									
Computation	48	1.59*	1.83	48	2.14*	2.58	48	1.48*	2.28
Concepts	25	2.90*	3.79	25	1.51*	3.40	25	1.48*	1.79
Application	25	1.59	1.48*	25	1.81*	2.12	25	1.63*	2.84
Reference Skills	20	1.89*	2.54	20	1.67*	1.69	20	2.22*	7.45
Science	36	2.02*	2.80	41	1.76*	2.98	40	1.42*	1.59
Social Studies	37	1.59*	2.31	40	1.71*	4.98	39	1.58*	5.33
Replication Datasets									
Reading									
Vocabulary	40	2.75*	2.77	40	1.78*	6.47	40	2.21*	9.64
Comprehension	45	1.74*	1.83	45	1.91*	2.61	45	1.75*	2.38
Language									
Spelling	50	2.06	1.84*	30	2.39*	2.91	30	1.70*	5.10
Mechanics	20	1.79	1.74*	20	2.29	1.88*	20	2.37	1.93*
Expression	35	2.01*	2.34	35	1.17*	2.25	35	2.19*	2.64
Mathematics									
Computation	48	1.74*	2.01	48	1.83*	2.47	48	1.77*	4.41
Concepts	25	2.64*	2.78	25	1.41*	7.60	25	2.02*	7.18
Application	25	1.84*	2.40	25	2.12*	2.61	25	2.11*	6.11
Reference Skills	20	1.81*	7.10	20	1.76*	3.72	20	1.97*	2.31
Science	36	1.87*	2.85	41	1.65*	2.12	40	1.55*	1.83
Social Studies	37	1.71*	7.14	40	1.66*	1.97	39	1.58*	6.23

*Indicates better fit.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Note that none of these results affects, in any way, the results of the comparison between the three-parameter model and the ARRG model. Rather, the comparisons between the ARRG model and the three-parameter model using the constant method strongly reinforce the contention that the guessing parameters should be estimated individually rather than at a constant value.

As might be expected, there were convergence problems in a number of testing situations. With the three-parameter model these problems tended to occur with items that were at least moderately difficult—indicating that some guessing may be present in the data—and with items that at the same time had low overall discriminations. For each of these problem items the guessing parameter was reset to 0, thus fitting a two-parameter model for that item and allowing that item's discrimination to remain relatively low. Using the quasi-marginal maximum likelihood procedure as implemented in LOGOG of assuming a prior distribution of θ when estimating the item parameters in the two-parameter model, convergence was necessarily obtained in all such situations when the method of estimating the guessing parameter was the individualized method.

These problems were much more troublesome when using the constant method. In order to obtain convergence, it was often necessary to reset as many as half the items' guessing parameters to 0 while leaving the remaining items at the constant value of .25. In the replication sample analyses, for example, only 6 of the 33 examinations came to stable item parameters with all $c_j = .25$. In the remaining 27 analyses, an average of 25% of the guessing parameters had to be reset to 0 in order to achieve stable item parameters.

With so many items' guessing parameters set equal to 0 rather than the constant, it seems inappropriate to refer to these analyses as using the constant method. However, because less than 20% of the examinations achieved stable item parameters with all $c_j = .25$, it is the closest approximation to constant c_j analyses that can be obtained.

Discussion. The lack of convergence under the three-parameter model is not necessarily the result of estimation problems; rather, it may be the result of data indicating that one or more of the assumptions made by this IRT model are being violated. Samejima (1973) showed that the second derivative of the log-likelihood for the three-parameter model is not everywhere negative. In particular, Samejima demonstrated that for some response vectors, estimates of θ can be infinite. There is every reason to assume that with the similarly configured second derivatives of the item parameters, the structure of the data for a particular item will yield infinite estimates of the discrimination and/or difficulty. That is, for a particular dataset the likelihood function for some item may be structured so that the maximum occurs with one or both item parameters approaching infinity. In this case, failure to converge on maximum likelihood estimates does not represent a convergence problem, but is simply an accurate attribute of the model for that dataset. Because this is unlikely to happen if all assumptions of the model are met, lack of convergence might simply imply failure to meet the assumptions of the three-parameter model.

Whatever the reason, if for a particular dataset the likelihood function for some item approaches a maximum as one or both parameters approach infinity, then in order to obtain maximum likelihood estimates of the difficulty and discrimination parameters for that item, the two-parameter model must be fitted. That is, for some datasets it may be necessary to accept the reduced information provided by such items, which will be reflected in the lower discriminations that accompany resetting the guessing parameter to 0 and fitting a two-parameter model to these items in order to obtain convergence.

The ARRG model also has stabilization problems. That is, unique maximum likelihood estimates are assured within any cycle where the θs are estimated assuming the item parameters or where the item parameters are estimated assuming a normal distribution of θ.

However, across cycles there is enough movement of examinees between adjacent fractiles as a result of the ARRG model cutoff to prevent convergence to absolutely stable item parameters. Nevertheless, as indicated by Waller's (1973) simulation studies, the movement between fractiles is

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

minimal. Working with the same structure for parameter estimation but with the normal ogive model, that study indicated that stopping the estimation process after four cycles produced parameter estimates that are unbiased. Concomitantly, that study also indicated that the standard errors of the parameters are underestimated in the ARRG model in the sense that $(1 - \alpha)$-level confidence intervals should probably be considered to have at most $(1 - 2\alpha)$ confidence. This is probably a direct result of the additional error introduced by the presence of random guessing within the data.

Conclusions

The results of this study indicate that the three-parameter model *can* be viewed as making an adjustment for partial-knowledge guessing behavior. The results also indicate two other implications for the practical use of IRT models:

1. The guessing parameter in the three-parameter model should be estimated for each item separately and not with one value equal to $1/A$.
2. Using the quasi-marginal maximum likelihood procedure (as implemented in LOGOG) of assuming a prior distribution of θ, problems of convergence with items possessing a guessing parameter greater than 0 can be solved by simply resetting the guessing parameter for those items to 0 and accepting the resulting lower discrimination as a necessary result when the data indicate that the three-parameter model is inappropriate.

References

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

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

Choppin, B. H. L. (1985). A two-parameter latent trait model. *Evaluation in Education, 9,* 43–62.

CTB McGraw-Hill. (1974). *Comprehensive Tests of Basic Skills*. Monterey CA: Author.

Feller, W. (1968). *An introduction to probability theory and its applications, volume 1*. New York: Wiley.

Kolakowski, D., & Bock, R. D. (1973). *LOGOG: Maximum likelihood item analysis and test scoring: Logistic model for multiple item responses*. Ann Arbor MI: National Educational Resources.

Samejima, F. (1973). A comment on Birnbaum's three-parameter model in the latent trait theory. *Psychometrika, 38,* 221–233.

Waller, M. I. (1973). *Removing the effects of random guessing from latent trait ability estimates*. Unpublished doctoral dissertation, University of Chicago.

Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing* (Research Bulletin RB-76-8). Princeton NJ: Educational Testing Service.

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

Author's Address

Send requests for reprints or further information to Michael I. Waller, Department of Educational Psychology, University of Wisconsin—Milwaukee, Box 413, Milwaukee WI 53201, U.S.A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/