ED 390 944		TM 024 598

AUTHOR	De Ayala, R. J.
TITLE	Item Parameter Recovery for the Nominal Response Model.
PUB DATE	Apr 95
NOTE	32p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Adaptive Testing; *Computer Assisted Testing; Educational Diagnosis; *Estimation (Mathematics); Item Response Theory; *Sample Size; Simulation
IDENTIFIERS	Accuracy; *Item Parameters; Latent Ability Distributions; *Nominal Response Model; Performance Based Evaluation; Polytomous Items

ABSTRACT

This study extended item parameter recovery studies in item response theory to the nominal response model (NRM). The NRM may be used with computerized adaptive testing, testlets, demographic items, and items whose alternatives provide educational diagnostic information. Moreover, with the increasing popularity of performance-based assessment, the use of polytomous item response theory models, in general, and the NRM in particular, will more than likely see increased application. Establishing guidelines for reasonable item parameter estimation was seen as fundamental to the use of the NRM. Factors studied through simulation were the sample size ratio, the latent ability distribution, and item information level. Results showed that as the latent ability distribution departs from a uniform distribution the accuracy of estimating the slope parameter decreased. This decrease in accuracy may be compensated for, in part, by increasing the sample size. Moreover, more informative items tended not to be as well estimated as less informative items. The results appear to indicate that if one is interested in estimating ability, a sample size ratio of 5:1 can produce reasonably accurate item parameter estimates for this purpose. (Contains 7 figures, 7 tables, and 26 references.) (Author/SLD)

Item Parameter Recovery for the Nominal Response Model

R.J. De Ayala
University of Maryland

Please send correspondence to :
R.J. De Ayala
Measurement, Statistics, and Evaluation
Benjamin Building
University of Maryland
College Park, MD 20742

Suggested Running Header: NR item parameter recovery

ABSTRACT

This study extended item parameter recovery studies in item response theory to the nominal response model (NRM). The NRM may be used with computerized adaptive testing, testlets, demographic items, and items whose alternatives provide educational diagnostic information. Moreover, with the increasing popularity of performance-based assessment, the use of polytomous item response theory models, in general, and the NRM in particular, will more than likely see increase application. Establishing guidelines for reasonable item parameter estimation was seen as fundamental to the the use of the NRM. Factors studied were the sample size ratio, the latent ability distribution, and item information level. Results showed that as the latent ability distribution departs from a uniform distribution the accuracy of estimating the slope parameter decreased. This decrease in accuracy may be compensated for, in part, by increasing the sample size. Moreover, more informative items tended not to be as well estimated as less informative items. The results appear to indicate that if one is interested in estimating ability, a sample size ratio of 5 : 1 can produce reasonably accurate item parameter estimates for this purpose.

Item response theory (IRT) has emerged as a popular approach foı solving various measurement problems. IRT is used in state testing programs such as the Maryland State Department of Education's High School Functional Assessment program as well as in municipal programs, such as the Portland School district. Both of these programs use IRT for test equating and the Portland program also uses IRT for test design (Ferrara, personal communication, October 4, 1991; Kingsbury, personal communication, Nov. 19, 1991; Forster, 1987). The nationally available California Achievement Test and the California Test of Basic Skills (Fourth Edition) are designed and equated using IRT (CTB/McGraw-Hill, 1987; CTB/MacMillan/McGraw-Hill, 1991). Moreover, certification boards such as the American Society of Clinical Pathologists have an IRT-based adaptive testing program for certification (Bergstrom & Lunz, 1991).

Most IRT work has been based on binary models such as the one- and three-parameter logistic models. With these models an individual's response is categorized as either correct or incorrect. However, not all examinee-item interactions may be appropriately modeled by binary models. For instance, to capture the information in a Likert item or to assign credit for a partially correct answer requires a model that contains more than two categories. Moreover, because the distributions of wrong answers over the options of multiple-choice items differ across ability levels (Nedelsky, 1954; Levine & Drasgow, 1983), it is possible and may be desirable to use a model that can assess information from all item options rather than to use a model which assumes an examinee either knows the correct answer or randomly selects an incorrect alternative. In addition, the one- and three-parameter logistic models do not incorporate findings from human cognition studies (e.g., Brown & Burton, 1978; Brown & VanLehn, 1980; Lane, Stone, & Hsu, 1990; Tatsuoka, 1983). For instance, Tatsuoka's (1983) analysis of student misconceptions in performing mathematics problem showed that wrong responses could be of more than just one kind. However, dichotomous scoring uniformly assigned a score of zero to all the wrong responses. In this regard, an item's incorrect alternatives may augment our estimate of an examinee's ability by providing information about the examinee's level of understanding (i.e., provide diagnostic information).

In contrast to binary models, polytomous models contain more item parameters to estimate. Because of these additional parameters potentially larger sample sizes may be required for their accurate estimation. For example, for Masters (1982) partial credit model (PCM) a 2 : 1 or larger ratios of examinees to item parameters were needed to produce stable item and ability parameter estimates, regardless of the number of categories (Walker-Bartnick, 1990). For Samejima's (1969) graded response model (GRM), Reise and Yu (1990) recommended that at least 500 examinees are needed to achieve an adequate calibration with the GRM. Their study was conducted with a 25-item test and therefore their guidelines may only be appropriate for tests of this length. (With longer tests it may be necessary to increase the sample size.) Similar findings were reported by Ankenmann and Stone (1992).

One polytomous model for which item parameter recovery has not been studied is Bock's (1972) nominal response model (NRM). The NRM is appropriate for items with unordered responses. The NRM may be used in computerized adaptive testing (De Ayala, 1992), with testlets (Wainer & Kiely, 1987) to solve various testing issues, such as multidimensionality (Thissen, Steinberg, & Mooney, 1989), with items that do not have a "correct" response, such as demographic items (e.g., to provide ancillary information), and with items whose alternatives provide educational diagnostic information. Moreover, with the increasing popularity of performance-based assessment, the use of po' ytomous IRT models, in general, and the NRM in particular, will more than likely see increase application.

The objective of this study was to establish guidelines for obtaining reasonably accurate item parameter estimates for the NRM. Because it was believed that the ratio of the sample size to number of parameters to be estimated is more useful than the actual sample size used, one factor studied was the sample size ratio (SSR). For instance, simply because the use of 100 examinees allows accurate Rasch parameter estimation with a 20 item test does not necessarily imply that only 100 examinees are required to obtain good estimates with a 100 item test. In this study, three ratios of observations to number of item parameter to be estimated were investigated: 5 : 1, 10 : 1, and 20 : 1.

Previous parameter recovery studies (e.g., Ankenmann & Stone, 1992; Reise & Yu, 1990) have varied the discrimination parameter. For example, Reise and Yu classified item discrimination into three ranges, high, medium, and low. However, because with the NRM there are multiple discrimination parameters for each item such a scheme did not appear to be useful. Further complicating the issue is the fact that when the number of categories is three or more, different combinations of an item's slopes and intercepts can produce the same maximum amount of information (I_{max}) value. Therefore, establishing guidelines in terms of the magnitude of the slope vectors was not pursued. Rather, in order to establish a design with the characteristic of "high", "medium", "low" discrimination, it was noted that the primary importance of the discrimination parameter is its effect on item information. Therefore, one may re conceptualize the Reise and Yu study as using items that are "high", "medium", and "low" in information rather than in terms of discrimination parameters. As such, values for I_{max} were set a priori and a slope vector to obtain a specific I_{max} was determined. The I_{max} values studied were 0.25, 0.16, and 0.09; for dichotomous models these I_{max}s are equivalent to items with discriminations of 1.0, 0.8, and 0.6, respectively.

Because the accuracy of estimating items located at various points along the ability (θ) scale may be affected by the latent θ distribution (LD), a third factor investigated was the effect of the

LD. Three distributions, normal, positively skewed, and uniform, were studied. An additional factor used in the study was whether the item consisted of three or four options.

Model

The NRM assumes that item alternatives represent responses which are unordered. The NRM provides a direct expression for obtaining the probability of an examinee with ability θ responding in the j-th category of item i as:

$$p_{ij}(\theta) = \frac{\exp(c_{ij} + a_{ij}\theta)}{\sum_{h=1}^{m_i} \exp(c_{ij} + a_{ij}\theta)} = \frac{\exp(a_{ij}(\theta - b_{ij}))}{\sum_{h=1}^{m_i} \exp(a_{ij}(\theta - b_{ij}))} \quad . \qquad (1)$$

where a_{ij} and c_{ij} are the slope and intercept parameters, respectively, of the nonlinear response function associated with the j-th category of item i, and m_i is the number of categories of item i (i.e., $j = 1, 2, ..., m_i$). For convenience the slope and intercept parameters are sometimes represented in vector notation, where $\mathbf{a} = (a_{i1}, a_{i2}, ..., a_{im})$ and $\mathbf{c} = (c_{i1}, c_{i2}, ..., c_{im})$. The a_{ij}s are analogous to and have an interpretation similar to traditional option discrimination indices. That is, a crosstabulation of ability groups by item alternatives shows that a category with a large a_{ij} reflects a response pattern in which as one progresses from the lower ability groups to the higher ability groups there was a corresponding increase in the number of persons who answered the item in that category and for categories with negative a_{ij}s this pattern is reversed. The intercept parameters reflect the interaction between a category's difficulty and how well it discriminates. It appears that, in general, large values of c_{ij} are associated with categories with large frequencies and as the value of c_{ij} becomes increasingly smaller the frequencies for the corresponding categories decrease.

The probability of responding in a particular category as a function of θ may be depicted by the option response function (ORF); other synonymous terms are category or option characteristic curve and trace line. Figure 1 contains the ORFs for a three-category ($m = 3$) item with $\mathbf{a} = (-0.75, -0.25, 1.0)$ and $\mathbf{c} = (-1.5, -0.25, 1.75)$.

Insert Figure 1 about here

The intersection of the ORFs can be obtained by setting adjacent category multivariate logit equal to one another and solving for θ. In general, for any item with $m_i \geq 2$ and because θ and b are on the same scale:

$$b_{k-1} = \frac{c_{(k-1)} - c_k}{a_k - a_{(k-1)}} \quad . \qquad (2)$$

where $k = 2...m_i$ and there are $m_i - 1$ ORF intersection points. This formulation is analogous to the step difficulties in the PC model.

METHOD

Programs: MULTILOG (Thissen, 1988) was used to obtain item parameter estimates for the NRM using default program parameters. A data generation program for generating responses according to the NRM was also written.

Data: A series of data sets were created. Each data set consisted of responses to 28 items and the data sets differed from one another on the basis of I_{max}, the number of item options, the form of the ability distribution from which the simulees were sampled, and the SSR. The 28 item set was created by determining for a given I_{max} level the **c** vector needed to locate the items' location (the average of the b_js) at one of the seven scale points between -3.0 to 3.0 in increments of 1 logit. For example, for a four-option item for the 0.25 I_{max} condition **a** = (0.450. -0.150, -0.100. -0.200) and to locate this item at -3.0 one would use **c** = (0.926. -0.275, -0.125.-0.525). (That is, the item's location = $(b_1 + b_2 + b_3)/3$ with b_1 = -2.00. b_2 = -3.000, and b_3 = -4.000 and the b_js are always one logit apart.) In this fashion seven items were created that spanned the usual θ range used in IRT and these items were replicated to produce the 28 item set.

For the three-option set of items the number of parameters to be estimated was 168 ((3 a_{ij}s + 3 c_{ij}s) X 28 items) and for the four-option item set there were 224 item parameters to estimate. With SSRs of 5 : 1, 10 : 1, and 20 : 1 this produced, for the three-option items, sample sizes of 840, 1680, and 3360, respectively, and for the four-option items samples of 1120, 2240, and 4480, respectively, were needed. For a given LD condition, the appropriate number of zs was sampled from a normal (0,1) distribution, a beta distribution (df_1 = 1.25. df_2 = 10), or a uniform distribution [-4, 4]. These zs were considered to be the simulees' true θs and the θs plus the 28 item parameters were used to generate polytomous response strings with a random error component for each simulated examinee. Generation of an examinee's polytomous response string was accomplished by calculating the probability of responding to each alternative of an item according to the NRM. Based on the probability for each alternative, cumulative probabilities were obtained for each alternative. A random error component was incorporated into each response by selecting a random number from a uniform distribution [0,1] and comparing it to the cumulative probabilities. The ordinal position of the first cumulative probability which was greater than the random number was taken as the examinee's response to the item.

For each of the (3 SSRs X 3 LDs X 3 I_{max}s X 2 m_js=) 54 conditions twenty-five replications were performed. That is, for a given condition (e.g., I_{max} = 0.25, normal θ distribution. 20 : 1 SSR. 4-option items), twenty-five unique response data sets were generated and each was calibrated using MULTILOG. This produced twenty-five sets of item parameter estimates for each set of item parameters. For a given combination of the LD and SSR factors, the same examinees were used for each of the I_{max} factor levels (i.e., I_{max} was a repeated measures factor).

Equating: Because of the indeterminacy of the ability scale, calibration programs define the scale so that the mean and standard deviation of θ (or b) are 0 and 1, respectively, for the calibration group. Therefore, the use of scale dependent accuracy measures, such as RMSE and average absolute deviation, require that the item parameter estimates be place on the parameter scale. The relationship between the item parameter estimate metric and the item parameter metric is a linear one. The basic transformation is:

$$\theta' = \alpha\theta + \kappa \qquad (3)$$

$$a' = \frac{a}{\alpha} \qquad (4)$$

$$b' = \alpha b + \kappa \qquad (5)$$

where θ', a', and b' are the transformed parameters corresponding to θ, a, and b, and α and κ are the slope and intercept equating constants, respectively. In the context of the present discussion θ', a', and b' are on the parameter (target) metric, whereas θ, a, and b are on the estimate (base) metric.

The determination of the α and κ may be accomplished in a number of ways. For instance, Stocking and Lord (1983) have developed a procedure for obtaining the equating constants based on test characteristic curves (TCCs); this procedure has been implemented in the EQUATE 2.0 program (Baker, 1993a) for the binary models, the GRM, and the NRM (Baker, 1992, Baker, 1993b, Baker & Al-Karni, 1991). An alternative method using the mean difficulty and the mean discrimination for obtaining α and κ was presented by Loyd and Hoover (1980).

Because the Loyd and Hoover (LH) method is more parsimonious than the Stocking and Lord approach, as well as for other pragmatic reasons[1], the LH method was generalized to the nominal response model and used for equating the NR item parameter estimates with the item parameters. The LH method specifies that:

$$\alpha = \frac{\bar{a}}{\bar{a}'} \qquad (6)$$

$$\kappa = \bar{b}' - \alpha\bar{b} \qquad (7)$$

Given that the slope-intercept form of the NRM multivariate logit for item i category j may be reparameterized as:

$$c_{ij} + a_{ij}\theta = a_{ij}(\theta - b_{ij})$$

'.nd because $c_{ij} = -a_{ij}b_{ij}$, one obtains across items that for category j:

$$\bar{b}_j = -\frac{\bar{c}_j}{\bar{a}_j}$$

Therefore, sums are taken across the common items and by substitution as well as by noting that $\bar{b}'_j = -\frac{\bar{c}'_j}{\bar{a}'_j}$, one obtains:

$$\alpha_j = \frac{\bar{a}_j}{\bar{a}'_j} \qquad (8)$$

$$\kappa_j = \bar{b}'_j - \alpha\bar{b}_j = \bar{b}'_j - \frac{\bar{a}_j}{\bar{a}'_j}\bar{b}_j = -\frac{\bar{c}'_j}{\bar{a}'_j} - \frac{\bar{a}_j}{\bar{a}'_j}\bar{b}_j = \frac{\bar{c}_j - \bar{c}'_j}{\bar{a}'_j} \qquad (9)$$

Equations (8) and (9) are the EQ-NR method. The equating constants may then be applied to transform one metric to another:

$$a'_{ij} = \frac{a_{ij}}{\alpha_j} \qquad (10)$$

$$c'_{ij} = c_{ij} - a'_{ij}\kappa_j \qquad (11)$$

where a'_{ij} and c'_{ij} are the equated (transformed) slope and intercept parameters, respectively, and a_{ij} and c_{ij} are the untransformed slope and intercept parameters, respectively.

Table 1 contains an example of the application of the EQ-NR method. NRM item parameters for four 4-option items were randomly generated and transformed to item parameter "estimates" by applying the reparameterized forms of (4) and (5) (i.e., $a'_{ij} = \frac{a_{ij}}{\alpha}$ and $c'_{ij} = c_{ij} - \kappa a'_{ij}$), where $\alpha = 0.4$ and $\kappa = 1.3$. The estimates were then transformed back to the parameter metric by application of the EQ-NR method; α = (2.5, 2.5, 2.5, 2.5) and κ = (-3.25, -3.25, -3.25, -3.25). As can be seen, the equated item parameter estimates are equal to the parameters. (The application of the LH method to ordered polytomous models, such as the PCM and the GRM, is a direct extension the binary case^2.) The major advantages of the EQ-NR method are its simplicity and that no special software is necessary for its implementation. However, its robustness in real-world applications needs to be investigated.

Insert Table 1 about here

Analysis: The accuracy of item parameter estimation was assessed by root mean square error (RMSE). RMSE was calculated according to:

$$\text{RMSE}(\Lambda_{ij}) = \sqrt{\frac{\sum(\hat{\Lambda}_{ij} - \Lambda_{ij})^2}{n_r}} \qquad (12)$$

where $\hat{\Lambda}_{ij}$ is the equated item parameter estimate (either $\hat{a}_{ij}$ or $\hat{c}_{ij}$) for item i option j, Λ_{ij} is the corresponding item parameter (either a_{ij} or c_{ij}), and n_r, the number of replications, equaled 25.

The analysis of the 3- and 4-category cases were treated separately as were the slope and intercept parameters. The basic design was a two-group repeated measures with LD and SSR as the between subjects factors and I_{max} as the within subjects factor. Because $\sum^{m_i} a_j = 0$ and $\sum^{m_i} c_j = 0$. a and c do not consist of m_i independent item parameter estimates and the RMSE for each item option parameter estimate could not be used as the dependent variable. Therefore, the mean RMSE(Λ) across item options and across replicates was used as the dependent variable.

It was expected that the accuracy of item parameter estimation would be related to the distribution of responses across item options. A measure of the distribution of responses across item options was obtained by using the index of dispersion. D:

$$D = \frac{m_i(N_i^2 - \sum_{j=1}^{m_i} n_{ij}^2)}{N_i^2(m_i - 1)} . \qquad (13)$$

where N_i is the number of examinees responding to item i and n_{ij} is the number of examinees responding in option j for item i. D has a range from 0.0 to 1.0 (inclusive) with D = 0.0 indicating that all responses to an item are in one option and D = 1.0 signifying that responses are evenly distributed across all options.

RESULTS

Table 2 contains descriptive statistics on the latent ability distributions for each SSR as well as the mean correlation between the item parameter and its estimate (i.e.. the average correlation between the option parameter and its estimate across the number of item options, $\bar{r}_{\hat{a}a}$ and $\bar{r}_{\hat{c}c}$: the correlations were converted to zs before averaging). As can be seen for a given LD, increasing the SSR was associated with an increase in $\bar{r}_{\hat{a}a}$. regardless of the number of item options. Similarly. increasing the SSR produced an increase $\bar{r}_{\hat{c}c}$. however these increases were not as dramatic due to the strong linear relationship between $\hat{c}$ and c at the 5 : 1 SSR. For a given LD and SSR level the $\bar{r}_{\hat{a}a}$s were consistently larger for the three category condition than for the four-option category. For a given SSR condition the $\bar{r}_{\hat{a}a}$s were largest for the uniform θ distributions and smallest for the positively skewed θ distributions. regardless of the number of item options.

Insert Table 2 about here

Figure 2 contains plots of D versus an item's average RMSE(a) for the 5 : 1 and 20 : 1 SSRs for the three- and four-option item sets; the 10 : 1 SSR plot falls predictably between the 5 : 1 and 20 : 1 SSR plots. As can be seen there is an inverse relationship between D and the mean RMSE(a). the average RMSE for an item decreased as the distribution of responses across an item's option increased. Specifically. for the three-option item set the correlations between D and the mean RMSE(a) were -0.597. -0.647. 0.647. for the 5 : 1. 10 : 1. and 20 : 1 SSRs and the corresponding correlations for the four-option items were -0.348. -0.438. and 0.485. For the intercepts the correlations for the 5 : 1, 10 : 1, and 20 : 1 SSRs/three-option items were -0.647, -0.569, and -0.512 and for the four-option items -0.331, -0.324, and -0.303, respectively. In general, the lowest RMSEs and larger Ds were associated with the uniform θ distribution, whereas the highest RMSEs and smaller Ds occurred with the positively skewed θ distribution, regardless of SSR. Moreover, for a given SSR level the mean D was less for the three-option item set ($\bar{D}_{5:1}$ = 0.863, $\bar{D}_{10:1}$ = 0.869, $\bar{D}_{20:1}$ = 0.864) than for the four-option item set ($\bar{D}_{5:1}$ = 0.927, $\bar{D}_{10:1}$ = 0.928, $\bar{D}_{20:1}$ = 0.928). The uniform θ distribution resulted in the greatest distribution of responses across item options ($\bar{D}_{4\text{-option}}$ = 0.940, $\bar{D}_{3\text{-option}}$ = 0.902), with the normal and positively skewed θ distributions having approximately the same average D values (normal: $\bar{D}_{4\text{-option}}$ = 0.922, $\bar{D}_{3\text{-option}}$ = 0.848; positively skewed: $\bar{D}_{4\text{-option}}$ = 0.920, $\bar{D}_{3\text{-option}}$ = 0.840).

Insert Figure 2 about here

The repeated measures analysis of the slope parameter (4-option items) is presented in Table 3. As can be seen, the accuracy of estimating the slope parameters was influenced by the interaction of the LD with the SSR and the I_{max}. Post hoc comparisons for the I_{max} factor showed that the slope parameters for items with I_{max} = 0.16 or the I_{max} = 0.09 (mean RMSE(a) = 0.060 and mean RMSE(a) = 0.057, respectively) were estimated significantly more accurately than for items with I_{max} = 0.25 (mean RMSE(a) = 0.071).

Analysis of the LD X SSR interaction showed that the average RMSE(a) for the uniform θ distribution was significantly less than that for either the normal or positively skewed distributions for all levels of the SSR factor and that the normal distribution mean RMSE(a) was significantly less that of the positively skewed ability distribution for the 5 : 1, 10 : 1, and 20 : 1 SSRs. Moreover, doubling the SSR led to significant reductions in the average RMSE(a) for the normal and the positively skewed θ distributions. Roughly speaking, quadrupling the sample size led to a halving of the average RMSE(a) for the 5 : 1 ratio. However, despite the increase in examinees for the positively skewed θ distribution the accuracy of estimation using the 20 : 1 ratio (mean RMSE(a) = 0.0636) only approximated that for the normal θ distribution using a 10 : 1 ratio (mean RMSE(a) = 0.0629). In addition, it took a 20 : 1 ratio with the normal θ distribution to produce an average RMSE(a) (mean RMSE(a) = 0.0430) approaching that for a uniform θ distribution based on a 5 : 1 SSR (mean RMSE(a) = 0.0427).

Insert Table 3 about here

Figure 3 contains the mean RMSE(a) for the SSR X LD interaction for the slope parameters for the four-option item sets. As can be seen, when θ is positively skewed twice as many subjects are needed in order to estimate the slope parameters approximately as accurately as when θ is normally distributed. For example, the mean RMSE(a) for the positive skewed LD condition using a 20 : 1 SSR is comparable to that with normal distribution and a 10 : 1 SSR. Similarly, with the positive skewed LD condition a 10 : 1 SSR results in a mean RMSE(a) that is slightly better than obtained using half as many subjects from a normal distribution. With a uniform distribution of ability even a 5 : 1 SSR provides more accurate estimation than can be obtained with four times as many subjects from a positively skewed ability distribution and almost comparable to that obtained when ability is normally distributed.

Insert Figure 3 about here

The analysis of the intercept parameters (4 option items) showed significant main effects for both the I_{max} and SSR factors (Table 4). Post hoc analyses showed that doubling the SSR did not lead to a significant reduction in the mean accuracy with which the intercept parameters were estimated. However, increasing the SSR from 5 : 1 to 20 : 1 led to almost halving the average RMSE(c): mean RMSE(c) for the 5 : 1, 10 : 1, 20 : 1 levels were 0.085, 0.061, 0.044, respectively). Similar to the case with RMSE(a), increasing I_{max} levels were associated with increases in the mean RMSE(c). As the item information increased from 0.09 to 0.16 to 0.25, there were significant decreases in the accuracy with which the intercept parameters were estimated (for I_{max} =0.09: mean RMSE(c) = 0.049, for I_{max} =0.16: mean RMSE(c) = 0.059, and for I_{max} =0.25: mean RMSE(c) = 0.081).

Insert Table 4 about here

Tables 5 and 6 contain the repeated measures analyses for the slope and intercept parameters for the three-option item set, respectively. Analysis of the significant SSR main effect for the slope parameter showed that doubling the SSR from 5 : 1 to 10 : 1 led to a significant reduction in the mean RMSE(a) (0.086 and 0.062, respectively), however, no significant improvement was realized by doubling the 10 : 1 SSR; mean RMSE(a) for 20 : 1 level was 0.047. Quadrupling the 5 : 1 SSR also led to significantly more accurate slope parameter estimates, on average. However, given the above results this would appear to be unnecessary to use an SSR greater than 10 : 1 with 3 option items. The significant I_{max} X LD interaction showed that, regardless of I_{max} level, that the uniform LD resulted, on average, in the most accurate RMSE(a) and the positively skewed LD the least accurate (Figure 4). In general, the slope parameters for the I_{max} = 0.25 level items were significantly more poorly estimated than for the I_{max} = 0.16 level items for all LDs and, except for the normal LD level, the average RMSE(a)s for the I_{max} = 0.16 level items were significantly greater than for the I_{max} = 0.09.

Insert Table 5 and Figure 4 about here

Analysis of the RMSE(c) for the three-option items revealed results that paralleled those for the four-option items. Specifically, increasing the SSR from 5 : 1 to 20 : 1 led to a significant reduction in the average RMSE(c); the mean RMSE(c)s for the 5 : 1, 10 : 1, and 20 : 1 SSR levels were 0.088, 0.062 and 0.045, respectively. As was the case with the four-option items, more informative item sets were not as well estimated, on average, as the less informative item sets; mean RMSE(c)s were 0.048, 0.063, and 0.083 for the 0.09, 0.16, and 0.25 I_{max} levels, respectively. The mean RMSE(a) and RMSE(c) for the three-option items were comparable in magnitude to those of the four-option item set.

Insert Table 6 about here

DISCUSSION

The use of marginal maximum likelihood estimation allows one to obtain item parameter estimates prior to estimating the examinees' θs. Obtaining the $\hat{\theta}$s may be performed using maximum likelihood, expected a posteriori (EAP), or maximum a posteriori estimation techniques and treating the item parameter estimates as known quantities. As such, SSR and LD's effect on the $\hat{\theta}$s will be indirect (if at all) and only through their effect on the accuracy of estimating the item parameters. For this reason this study focused only on the accuracy of estimation of NRM's item parameters.

Results showed that as the latent θ distribution departs from a uniform distribution the accuracy of estimating the slope parameter decreases. In these cases, in order to increase the accuracy of estimating the slope parameter one needs to increase the sample size. The effects of the form of θ distribution on RMSE may, in part, be related to the distribution of responses across item options. It was found that the uniform LD produced the greatest dispersal of responses across item options and that the positively skewed LD produced least variability in the examinees responses. Therefore, if there are insufficient numbers of examinees responding to a particular item option, then that option will not be as accurately estimated as other options that have attracted more examinee responses. (It should be noted that poor estimation of an option's parameters may affect the estimation of the other options' parameters.) Short of rewriting the option, increasing the sample size is one means of increasing the number examinees responding to a particularly unattractive option. Moreover, more informative items (i.e., items with larger slope parameters) tended not to be as well estimated as less informative items. However, the RMSE(a) observed for these informative items (e.g., I_{max} = 0.25) may be considered adequate by some. Similar findings were found with the intercept parameter. In particular, the more informative the items the greater the number of subjects required in order to estimate the intercepts with a degree of accuracy comparable to that of less informative items. This was true for both three- and four-option items.

Given the magnitude of the average RMSEs observed, were the significantly more accurate item parameter estimates obtained by increasing the SSR meaningfully more accurate? In a real world application there could be substantial costs involved in doubling or quadrupling the SSR (if it could be done at all). To answer this question an additional set of analyses based on confidence intervals (CIs) were performed.

For each of the original six item parameter pools, a data set was generated according to the NRM that contained the responses of 1100 simulees. These simulees were distributed such that 100 simulees were located at each of 11 θ points between -2.5 and 2.5 in 0.5 logit increments (i.e., 100 simulees had θ = -2.5, 100 simulees had θ = -2.0, ... 100 simulees had θ = 2.5). For each of these 1100 simulees the EAP $\hat{\theta}$ and its standard error of estimation were obtained using the item parameter estimates from each replication as well as the item parameters used to generate the data; for EAP estimation 80 quadrature points and a uniform prior was used. Because there were 25 replications for each condition there were 25 $\hat{\theta}$s for each simulee and for a given condition there was a total of 1100 X 25 = 27,500 $\hat{\theta}$s. For each of these $\hat{\theta}$s a 95% CI was calculated and for a given condition the number of times the CI contained θ was recorded. Table 7 contains the results of these analyses.

Insert Table 7 about here

As can be seen from top half of Table 7, while there were differences in the proportion of 95% CIs containing θ across I_{max}, for a given LD and I_{max} condition increasing the SSR did not appear to result in meaningful differences in the proportion of CIs containing θ. In general, the entries approxii..ated the expected value of 0.950. Alternatively, the CIs based on the item parameters give an indication of how well one could expect to do given the sample size used. The differences between the CIs calculated on the basis of the item parameters and their estimates are presented in the bottom half of Table 7. These differences are typically on the order of one one thousandths. Overall, the largest differences are found for the 5 : 1 SSRs. However, these are small differences. In this regard, it appears that if one's focus is to use item parameters for ability estimation, a 5 : 1 SSR may produce item parameter estimates that are reasonably accurate.

This CI approach (the top half of Table 7) may be used to compare different sets of item parameter estimates for meaningful differences with respect to $\hat{\theta}$s. (If competing models are to be compared, then a model independent simulation data set would be used.) The CI method has the advantages of simplicity, a clearly define and objective goal, an indication of how well or poorly one is doing in ability estimation, and, if desired, the possibility of significance testing.

References

Ankenmann, R.D., & Stone, C. (1992, April). *A monte carlo study of marginal maximum likelihood parameter estimates for the graded model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Baker, F. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Baker, F.B. (1993a). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17*, 20.

Baker, F.B. (1993b). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239-251.

Baker, F.B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

Bergstrom, B. & Lunz, M. (1991, April). *Confidence in pass/fail decisions for computer-adaptive and paper-and-pencil examinations.* Paper presented at the annual meeting of the American Educational Research Association, Chicago. IL.

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

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Brown, J.S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4*, 379-426.

CTB/MacMillan/McGraw-Hill (1991). *Comprehensive Tests of Basic Skills. Fourth Edition.* (Technical Report, June, 1991). Monterey, CA: CTB/MacMillan/McGraw-Hill.

CTB/McGraw-Hill (1987). *California Achievement Tests, Forms E and F, Levels 10-20.* (Technical Report). Monterey, CA: CTB/McGraw-Hill.

De Ayala, R.J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement,16*, 327-343.

Forster, F. (1987, April). *Riding the Rasch tiger: Laying the item bank foundation.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Lane, S., Stone, C.A., & Hsu, H. (1990). *Diagnosing students' errors in solving algebra word problems.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Levine, M., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating under the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Nedelsky, L. (1954). Ability to avoid gross ever as a measure of achievement. *Educational and Psychological Measurement, 14*, 459-472.

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Tatsuoka K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Thissen, D.J. (1988). *MULTILOG-User's Guide* (Version 5.1). Scientific Software, Inc. Mooresville, IN.

Thissen, D.J., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247-260.

Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Walker-Barnick, L.A. (1990). *An investigation of factors affecting invariance of item parameter estimates in the partial credit model.* (Doctoral dissertation, University of Maryland, College Park, 1990).

Footnotes

[1] According to the documentation which accompanies EQUATE 2.0 (Baker, 1993a) the "nominally scored test equating is quite sensitive to the values of the initial estimators [α and κ]". Moreover, "the interaction among sample size, the estimation techniques employed in MULTILOG, and the equating coefficients yielded by EQUATE are in need of further investigation" (Baker, 1993b, p 248).

[2] The application of the EQ-NR approach to ordered polytomous models (EQ-OR) is done threshold-wise for obtaining the κs for the thresholds ($\kappa_j = \bar{b}'_j - \alpha\bar{b}_j$) and item-wise for obtaining α ($\alpha = \frac{\bar{a}}{\bar{a}'}$). The transformation of the item discrimination and the thresholds is performed by $a'_i = \frac{a_i}{\alpha}$ and $b'_{ij} = \alpha b_{ij} + \kappa_j$.

Table 1: Example equating using EQ-NR method.

Parameter Item	a_1	a_2	a_3	a_4	c_1	c_2	c_3	c_4
1	-1.937	-0.702	0.516	2.123	-2.040	-0.781	1.333	1.488
2	-1.549	-1.361	1.039	1.870	-1.615	-0.913	0.719	1.809
3	-2.126	-0.829	1.216	1.739	-2.434	-1.185	0.949	2.670
4	-2.326	-1.155	1.131	2.350	-1.527	-1.379	0.527	2.378
Mean	-1.984	-1.012	0.976	2.020	-1.904	-1.065	0.882	2.087
Estimates								
1	-4.842	-1.755	1.289	5.308	4.255	1.500	-0.343	-5.412
2	-3.871	-3.402	2.598	4.675	3.418	3.509	-2.658	-4.269
3	-5.314	-2.073	3.040	4.347	4.475	1.510	-3.004	-2.981
4	-5.815	-2.887	2.828	5.874	6.033	2.375	-3.149	-5.258
Mean	-4.961	-2.529	2.439	5.051	4.545	2.223	-2.289	-4.480
Equated								
1	-1.937	-0.702	0.516	2.123	-2.040	-0.781	1.333	1.488
2	-1.549	-1.361	1.039	1.870	-1.615	-0.913	0.719	1.809
3	-2.126	-0.829	1.216	1.739	-2.434	-1.185	0.949	2.670
4	-2.326	-1.155	1.131	2.350	-1.527	-1.379	0.527	2.378
Mean	-1.984	-1.012	0.976	2.020	-1.904	-1.065	0.882	2.087

Table 2: Descriptive Statistics on Ability Distributions and Item Parameters[a]

		Mean	SD	Skew	3-option items[b]		4-option items[b]	
Distribution	SSR	θ	θ	θ	$\bar{r}_{\hat{a}a}$	$\bar{r}_{\hat{c}c}$	$\bar{r}_{\hat{a}a}$	$\bar{r}_{\hat{c}c}$
Normal	5 : 1	0.001	1.001	-0.024	0.668	0.991	0.621	0.988
	10 : 1	-0.003	0.999	-0.005	0.779	0.995	0.735	0.995
	20 : 1	-0.002	0.998	-0.011	0.870	0.998	0.844	0.997
PS	5 : 1	-0.001	0.731	1.291	0.529	0.991	0.500	0.989
	10 : 1	-0.003	0.733	1.332	0.652	0.996	0.637	0.995
	20 : 1	0.002	0.737	1.297	0.733	0.998	0.733	0.997
Uniform	5 : 1	-0.024	2.308	0.011	0.862	0.989	0.845	0.987
	10 : 1	0.002	2.315	0.004	0.920	0.994	0.897	0.993
	20 : 1	-0.005	2.309	0.003	0.946	0.997	0.929	0.996

[a] SD: Standard Deviation. PS: Positively Skewed

[b] correlations converted to z-scores before taking the average

Table 3: RMSE Repeated Measures Analyses for slope parameters (4 options)[a].

Source	SS	df	MS		p
Between Subjects					
LD	0.099	2	0.049	70.389	0.000
SSR	0.050	2	0.025	35.836	0.000
LD X SSR	0.010	4	0.002	3.554	0.012
Subj w/i Groups	0.038	54	0.001		
Within Subjects					
I_{max}	0.006	2	0.003	38.419	0.000
I_{max} X LD	0.001	4	0.000	1.580	0.185
I_{max} X SSR	0.000	4	0.000	0.442	0.778
LD X SSR X I_{max}	0.000	8	0.000	0.355	0.942
I_{max} X Subj w/i Groups	0.009	108	0.000		

Post Hoc Comparison ts for LD:

		SSR	
Hypothesis	5 : 1	10 : 1	20 : 1
μ_{nml} vs μ_{ps}	4.240*	2.588*	2.522*
μ_{nml} vs μ_{unif}	5.450*	3.682*	2.043*
μ_{ps} vs μ_{unif}	9.690*	6.269*	4.565*

Post Hoc Comparison ts for SSR:

		LD	
Hypothesis	Normal	PS	Unif
$\mu_{5:1}$ vs $\mu_{10:1}$	2.978*	4.630*	1.210
$\mu_{5:1}$ vs $\mu_{20:1}$	5.415*	7.133*	2.008*
$\mu_{10:1}$ vs $\mu_{20:1}$	2.437*	2.502*	0.798

Post Hoc Comparison ts for I_{max}:

Hypothesis	
$\mu_{0.09}$ vs $\mu_{0.16}$	2.391
$\mu_{0.09}$ vs $\mu_{0.25}$	11.730*
$\mu_{0.16}$ vs $\mu_{0.25}$	9.339*

[a] nml: Normal. ps: Positively Skewed. unif: Uniform

Table 4: RMSE Repeated Measures Analyses for intercept parameters (4 options)[a].

Source	SS	df	MS	F	p
Between Subjects					
LD	0.005	2	0.002	1.692	0.194
SSR	0.053	2	0.026	18.492	0.000
LD X SSR	0.001	4	0.000	0.164	0.956
Subj w/i Groups	0.077	54	0.001		
Within Subjects					
I_{max}	0.033	2	0.016	66.624	0.000
I_{max} X LD	0.001	4	0.000	1.524	0.200
I_{max} X SSR	0.002	4	0.000	1.783	0.138
LD X SSR X I_{max}	0.001	8	0.000	0.386	0.926
I_{max} X Subj w/i Groups	0.026	108	0.000		

Post Hoc Comparison ts for SSR:

Hypothesis	
$\mu_{5:1}$ vs $\mu_{10:1}$	-2.966
$\mu_{5:1}$ vs $\mu_{20:1}$	-4.932*
$\mu_{10:1}$ vs $\mu_{20:1}$	-1.966

Post Hoc Comparison ts for I_{max}:

Hypothesis	
$\mu_{0.09}$ vs $\mu_{0.16}$	4.891*
$\mu_{0.09}$ vs $\mu_{0.25}$	15.934*
$\mu_{0.16}$ vs $\mu_{0.25}$	11.042*

[a] nml: Normal. ps: Positively Skewed. unif: Uniform

Table 5: RMSE Repeated Measures Analyses for slope parameters (3 options)^a.

Source	SS	df	MS	F	p
Between Subjects					
LD	0.119	2	0.060	66.458	0.000
SSR	0.051	2	0.025	28.271	0.000
LD X SSR	0.007	4	0.002	1.889	0.126
Subj w/i Groups	0.048	54	0.001		
Within Subjects					
I_{max}	0.013	2	0.006	75.923	0.000
I_{max} X LD	0.001	4	0.001	3.696	0.007
I_{max} X SSR	0.000	4	0.000	0.449	0.773
LD X SSR X I_{max}	0.000	8	0.000	0.299	0.965
I_{max} X Subj w/i Groups	0.009	108	0.000		

Post Hoc Comparison ts for SSR:

Hypothesis	
$\mu_{5:1}$ vs $\mu_{10:1}$	-3.807*
$\mu_{5:1}$ vs $\mu_{20:1}$	-6.075*
$\mu_{10:1}$ vs $\mu_{20:1}$	-2.268

Post Hoc Comparison ts for LD:

		I_{max}	
Hypothesis	0.09	0.16	0.25
μ_{nml} vs μ_{ps}	4.272*	5.023*	6.737*
μ_{nml} vs μ_{unif}	5.740*	5.061*	4.895*
μ_{ps} vs μ_{unif}	10.012*	10.084*	11.632*

Post Hoc Comparison ts for I_{max}:

		LD	
Hypothesis	Normal	PS	Unif
$\mu_{0.09}$ vs $\mu_{0.16}$	1.344	2.886*	2.738*
$\mu_{0.09}$ vs $\mu_{0.25}$	4.721*	9.779*	6.455*
$\mu_{0.16}$ vs $\mu_{0.25}$	3.377*	6.893*	3.717*

[a] nml: Normal. ps: Positively Skewed. unif: Uniform

Table 6: RMSE Repeated Measures Analyses for intercept parameters (3 options)[a].

Source	SS	df	MS	F	p
Between Subjects					
LD	0.007	2	0.003	2.278	0.112
SSR	0.059	2	0.030	19.720	0.000
LD X SSR	0.001	4	0.000	0.085	0.987
Subj w/i Groups	0.081	54	0.001		
Within Subjects					
I_{max}	0.039	2	0.019	94.249	0.000
I_{max} X LD	0.001	4	0.000	1.755	0.143
I_{max} X SSR	0.002	4	0.000	2.109	0.085
LD X SSR X I_{max}	0.000	8	0.000	0.245	0.981
I_{max} X Subj w/i Groups	0.022	108	0.000		

Post Hor Comparison ts for SSR:

Hypothesis	
$\mu_{5:1}$ vs $\mu_{10:1}$	-3.017
$\mu_{5:1}$ vs $\mu_{20:1}$	-5.099*
$\mu_{10:1}$ vs $\mu_{20:1}$	-2.083

Post Hoc Comparison ts for I_{max}:

Hypothesis	
$\mu_{0.09}$ vs $\mu_{0.16}$	8.040*
$\mu_{0.09}$ vs $\mu_{0.25}$	19.326*
$\mu_{0.16}$ vs $\mu_{0.25}$	11.285*

[a] nml: Normal. ps: Positively Skewed. unif: Uniform

Table 7: Proportion of times 95% confidence intervals contained θ.

			3-options			4-options	
			I_{max}			I_{max}	
Distribution	SSR	0.09	0.16	0.25	0.09	0.16	0.25
Normal	parameters	0.952	0.956	0.943	0.946	0.936	0.952
	5 : 1	0.950	0.956	0.941	0.943	0.931	0.952
	10 : 1	0.952	0.957	0.942	0.945	0.935	0.952
	20 : 1	0.952	0.956	0.944	0.944	0.935	0.952
PS	parameters	0.952	0.956	0.943	0.946	0.936	0.952
	5 : 1	0.944	0.952	0.938	0.936	0.933	0.946
	10 : 1	0.952	0.954	0.941	0.945	0.937	0.950
	20 : 1	0.953	0.955	0.942	0.948	0.937	0.951
Uniform	parameters	0.952	0.956	0.943	0.946	0.936	0.952
	5 : 1	0.951	0.957	0.939	0.941	0.929	0.947
	10 : 1	0.952	0.957	0.941	0.942	0.931	0.947
	20 : 1	0.952	0.958	0.941	0.941	0.931	0.947

Differences between CIs based on item parameter estimates and CIs based on item parameters

Normal	5 : 1	-0.002	0.000	-0.002	-0.003	-0.005	0.000
	10 : 1	0.000	0.001	-0.001	-0.001	-0.001	0.000
	20 : 1	0.000	0.000	0.001	-0.002	-0.001	0.000
PS	5 : 1	-0.008	-0.004	-0.005	-0.010	-0.003	-0.006
	10 : 1	0.000	-0.002	-0.002	-0.001	0.001	-0.002
	20 : 1	0.001	-0.001	-0.001	0.002	0.001	-0.001
Uniform	5 : 1	-0.001	0.001	-0.004	-0.005	-0.007	-0.005
	10 : 1	0.000	0.001	-0.002	-0.004	-0.005	-0.005
	20 : 1	0.000	0.002	-0.002	-0.005	-0.005	-0.005

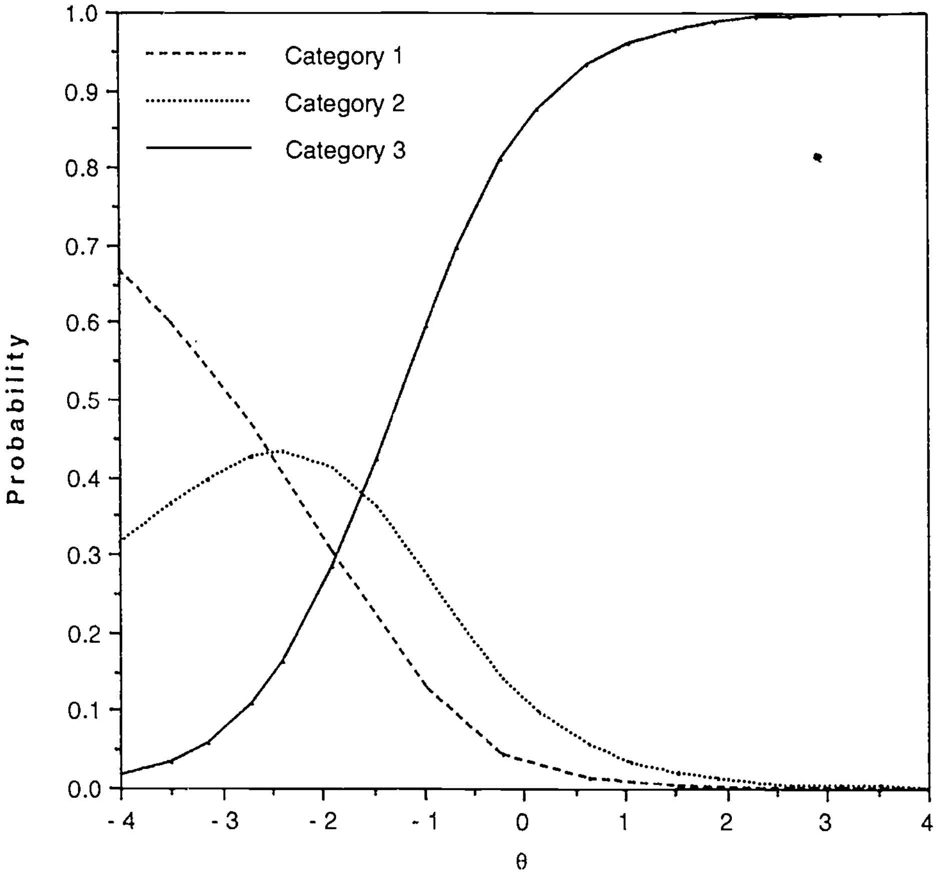
Figure Captions

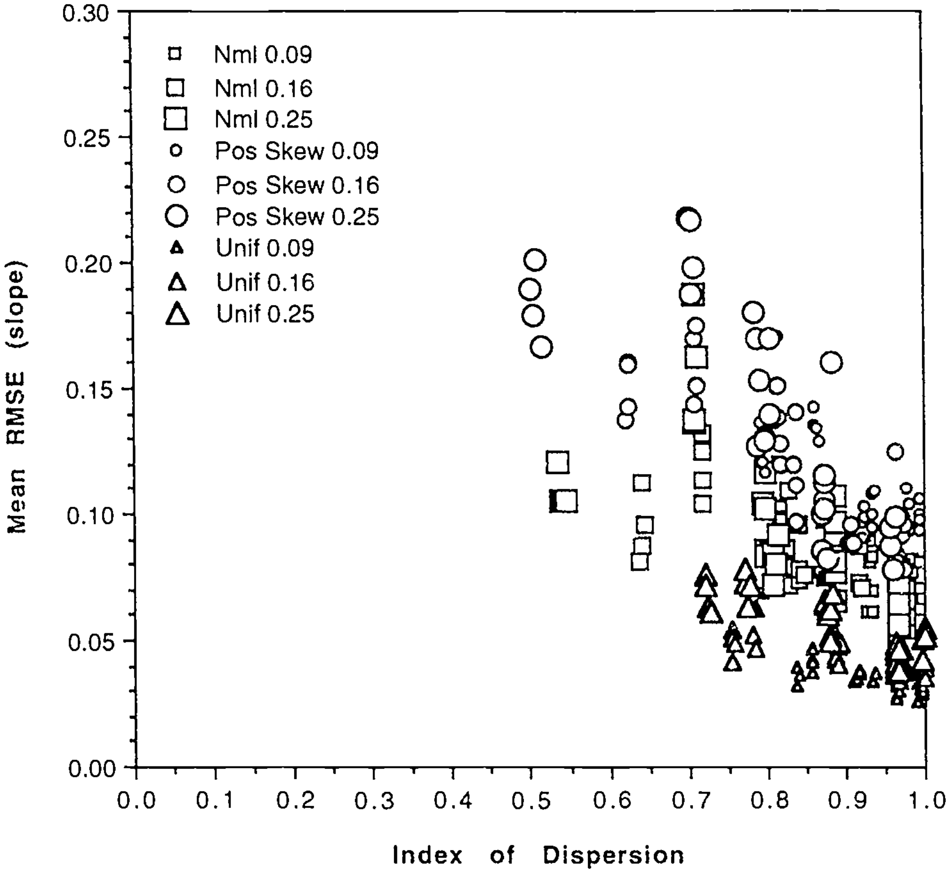
Figure 1. Example ORFs for a three-category item. a = (-0.75, -0.25, 1.0) and c = (-1.5, -0.25, 1.75).

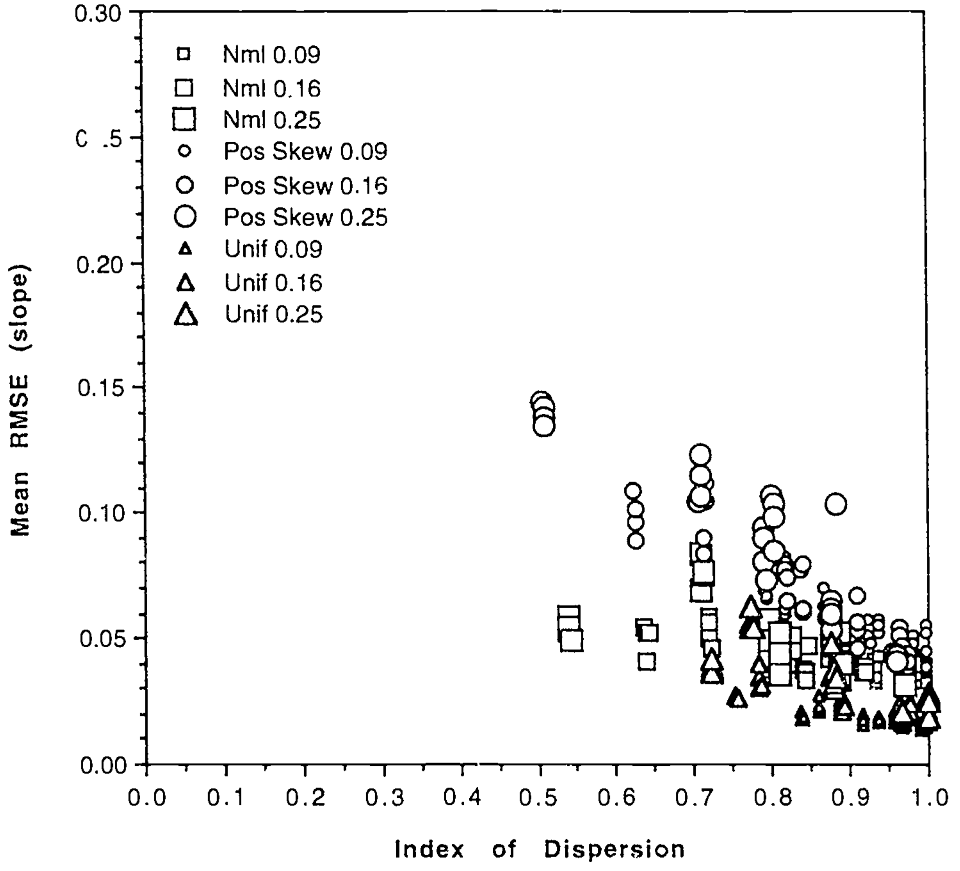
Figure 2a. D vs the mean RMSE(a)/item for 5:1 SSR, 3-option items.
Figure 2b. D vs the mean RMSE(a)/item for 20:1 SSR, 3-option items.
Figure 2c. D vs the mean RMSE(a)/item for 5:1 SSR, 4-option items.
Figure 2d. D vs the mean RMSE(a)/item for 20:1 SSR, 4-option items.

Figure 3. Mean RMSE(a) for the SSR and LD interaction, 4-option items.

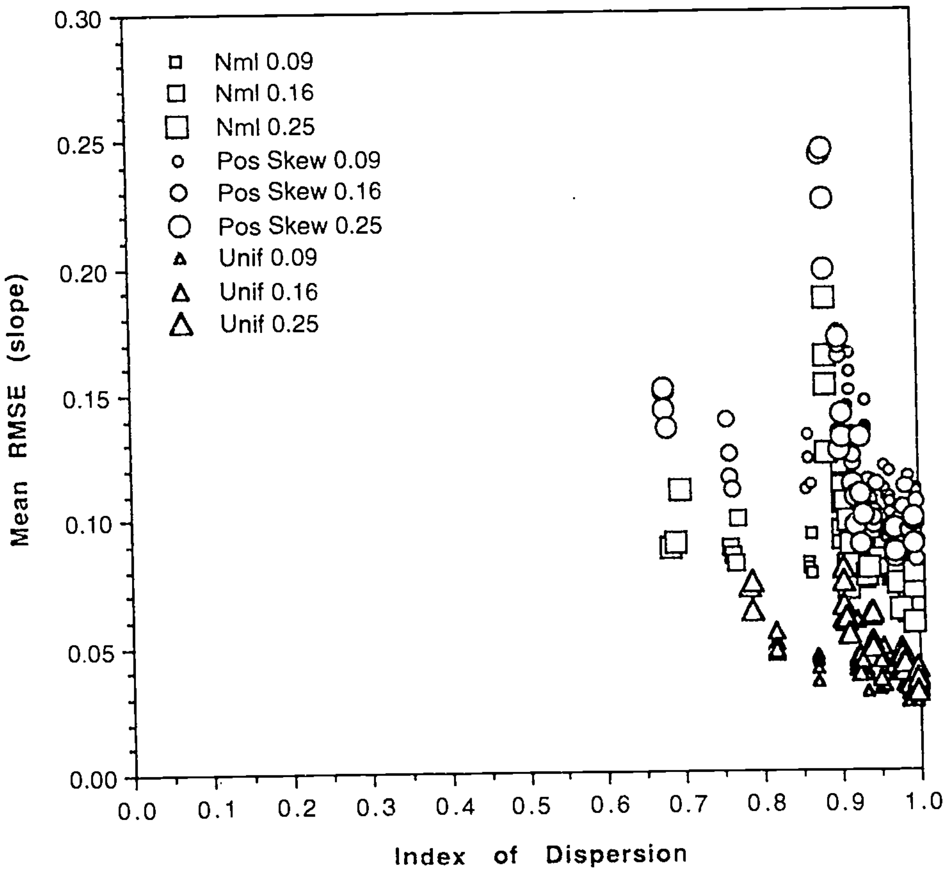
Figure 4. Mean RMSE(a) for the I_{max} by LD interaction.

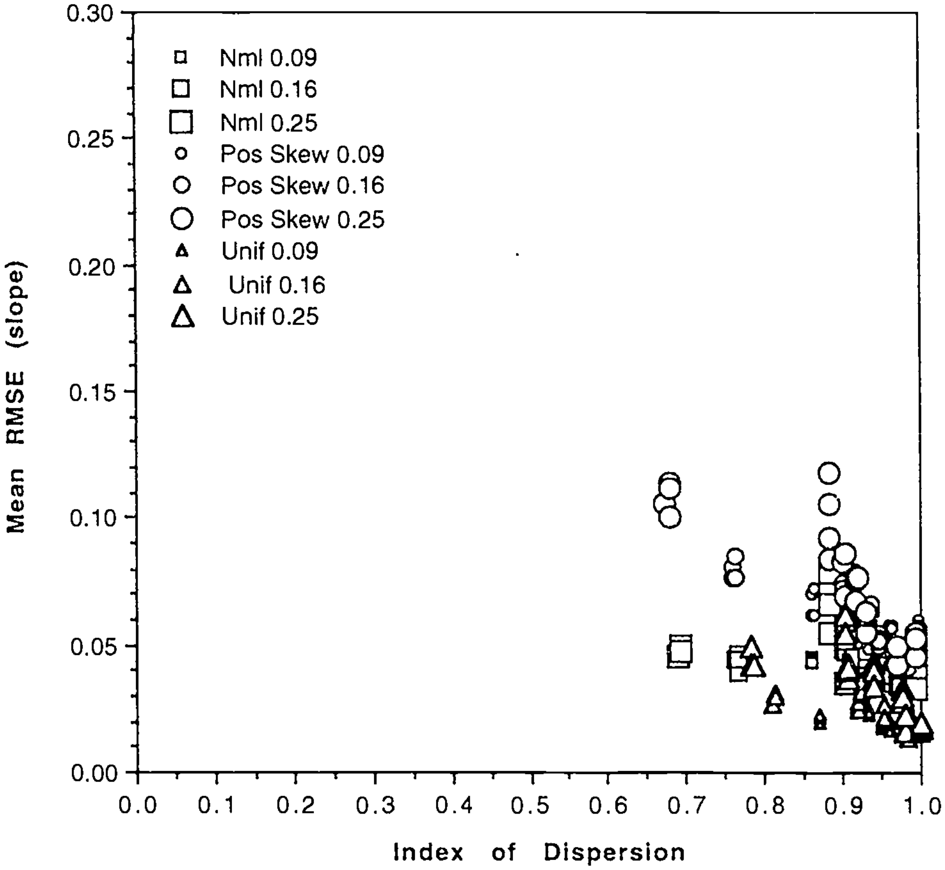

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Probability
Category 1
Category 2
Category 3
- 4
- 3
- 2
- 1
0
1
2
3
4
θ

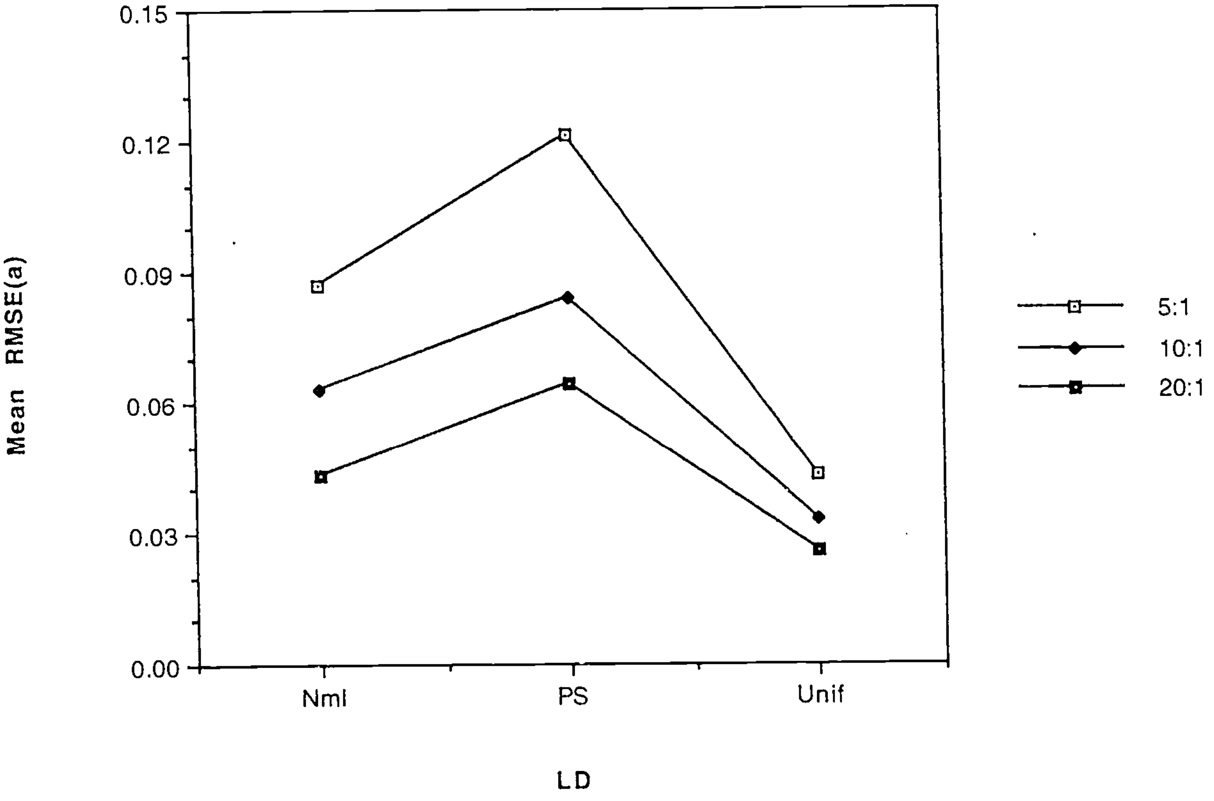

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean RMSE (slope)
□ Nml 0.09
□ Nml 0.16
□ Nml 0.25
○ Pos Skew 0.09
○ Pos Skew 0.16
○ Pos Skew 0.25
△ Unif 0.09
△ Unif 0.16
△ Unif 0.25
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Index of Dispersion

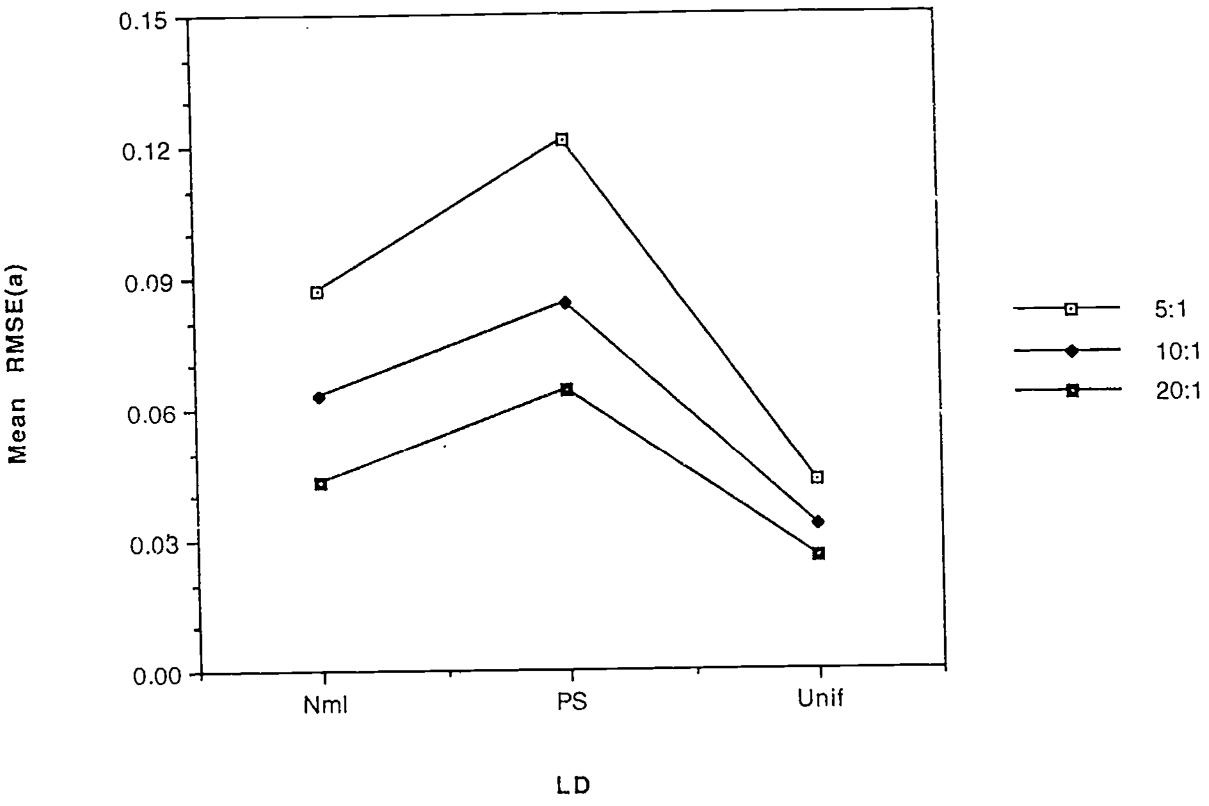


□ Nml 0.09
□ Nml 0.16
□ Nml 0.25
○ Pos Skew 0.09
○ Pos Skew 0.16
○ Pos Skew 0.25
△ Unif 0.09
△ Unif 0.16
△ Unif 0.25
Mean RMSE (slope)
0.30
C .5
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Index of Dispersion


0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean RMSE (slope)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Index of Dispersion
□ Nml 0.09
□ Nml 0.16
□ Nml 0.25
○ Pos Skew 0.09
○ Pos Skew 0.16
○ Pos Skew 0.25
△ Unif 0.09
△ Unif 0.16
△ Unif 0.25


0.30
0.25
0.20
0.15
0.10
0.05
0.00
Mean RMSE (slope)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Index of Dispersion
Nml 0.09
Nml 0.16
Nml 0.25
Pos Skew 0.09
Pos Skew 0.16
Pos Skew 0.25
Unif 0.09
Unif 0.16
Unif 0.25


0.15
0.12
0.09
0.06
0.03
0.00
Mean RMSE(a)
Nml
PS
Unif
LD
5:1
10:1
20:1


0.15
0.12
0.09
0.06
0.03
0.00
Mean RMSE(a)
Nml
PS
Unif
LD
5:1
10:1
20:1