Sensitivity of the Linear Logistic Test Model to Misspecification of the Weight Matrix

Frank B. Baker

University of Wisconsin

Under the linear logistic test model, a weight is assigned to each cognitive operation used to respond to an item. The allocation of these weights is open to misspecification that can result in faulty estimates of the basic parameters. The effect on root mean squares (RMSs) of the difference between the parameter estimates obtained under misspecification conditions and those obtained under correct specification conditions was examined. Six levels of misspecification and four sample sizes were used. Even a small number of errors in the weight specifications resulted in large RMS values. However, weight matrices with a high proportion of nonzero elements tended to yield RMSs that were approximately half as large as those with a small number of nonzero elements. Although sample size had some effect on the RMS values, it was quite small compared to that due to the level of misspecification of the weights. The results suggest that because specifying the elements in the weight matrix is a subjective process, it must be done with great care. *Index terms: error rates, linear logistic test model, misspecification, parameter estimation, weight matrix.*

Over the past two decades, the field of cognitive science has had a major impact on the study of human learning and problem solving. Rather than emphasize the outcomes of learning in the form of test scores and/or observable behavior, cognitive science focuses on the cognitive processes (operations)—such as schema and production rules—underlying these products (see Anderson, 1990). The change in emphasis from products to processes has begun to influence test construction and analysis procedures and poses a major challenge for psychometric theory (see Frederiksen, Mislevy, & Bejar, 1992). Classical test theory, which arose out of testing programs measuring intelligence, aptitude, and subject matter achievement is a test score theory (see Gulliksen, 1950; Lord & Novick, 1968). In contrast, item response theory (IRT) emphasizes how items determine the characteristics of tests and their role in determining an examinee's trait level (Hambleton & Swaminathan, 1985; Lord, 1980). Although IRT has many theoretical and practical advantages over classical test theory, neither of these psychometric theories was designed for a psychological process orientation.

The Linear Logistic Model

An extension of the Rasch (1960) model was developed that incorporates the role of underlying cognitive processes into IRT. This extension is called the linear logistic test model (LLTM), because a linear constraint is imposed on the item difficulty parameters (Fischer, 1973, 1983). Under this approach, the difficulty parameters of the items are assumed to be linear functions of a smaller set of basic parameters that represent the cognitive operations involved in the items. This relationship is formalized by:

$$\beta_i = \sum_{j=1}^{m} q_{ij}\eta_j + C \,, \tag{1}$$

where

β_i is the difficulty parameter of item i ($i = 1, 2, 3, \ldots, n$) and is in a logit metric,

η_j is the basic parameter corresponding to the jth underlying cognitive operation ($j = 1, 2, 3,$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 3, September 1993, pp. 201-210
© Copyright 1993 Applied Psychological Measurement Inc.
0146-6216/93/030201-10$1.80

$\ldots, m$,

q_{ij} is the weight assigned to the cognitive operation j in item i, and

C is the normalizing constant and is defined as

$$C = \frac{-\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}\eta_j}{n} . \qquad (2)$$

C is simply the mean of the β estimates before the identification problem is solved. Thus, the average β yielded by Equation 1 is 0. This is the same normalization used in the BICAL (Wright & Mead, 1978) and MICROSCALE (Wright & Linacre, 1984) computer programs for joint maximum likelihood estimation of the item βs under the Rasch model.

q_{ij} is an element of an $n \times m$ matrix of weights, denoted **Q**. Each row in this matrix corresponds to an item; within each row the contribution of each underlying cognitive operation to the item difficulty is given a value. If only the presence or absence of the operation is of interest, the weights are either 1 or 0. In some applications (see Spada, 1977), the weight is the number of times the operation is involved in the item. Given **Q** and the examinee's vector of dichotomously scored item responses, the values of the m η are estimated; then Equation 1 can be used to obtain the β estimates.

The conditional maximum likelihood procedure for estimating η and β under the LLTM was first implemented in a computer program by Fischer & Formann (1972), and was rewritten for personal computers by Fischer, Formann, & Wild (1989). The program has been extended by Fischer & Parzer (1991a) to include the rating scale model. This latter program was employed to estimate the parameters in a measurement of change situation (Fischer & Parzer, 1991b).

Applications of the LLTM

The LLTM has been used in a variety of situations. Fischer (1973) formulated the difficulty of elementary calculus problems as a function of eight differentiation rules. Koponen (1983) used the LLTM to study errors committed in solving mathematics problems in which the cognitive operations represented common errors. Whitely & Schneider (1981) investigated geometric analogies. Paragraph comprehension items were studied by Embretson & Wetzel (1987). Mechanical rotation and balance problems were investigated by Spada (1977) and by Spada & McGaw (1985). Document literacy was studied by Sheenan & Mislevy (1990).

None of this research, however, has dealt with estimating the examinee's trait level, which is the typical application of IRT. Rather, the emphasis has been on estimating the values of the η parameters and their role in defining item difficulty. Although the elements of **Q** indicate which cognitive operations are involved in an item, the numerical values of the η parameters indicate the relative contribution of each operation to the difficulty of the item—they provide greater insight into why an item is difficult or easy. Knowing the η parameters for a set of cognitive operations also provides the test designer with a more detailed basis for constructing additional test items (see Hornke & Habon, 1986; Nährer, 1980). Thus, the LLTM holds considerable promise for providing a bridge between cognitive science and IRT.

Using the LLTM

In order to employ the LLTM, the researcher must formulate the set of m cognitive operations underlying the test items and assign values to the corresponding elements in each item's vector in **Q**. A researcher can examine existing items, decide a priori what constitutes the set of cognitive operations underlying the n items in the instrument, and then define the combinations of η parameters involved in each item by assigning the values of q_{ij} for each item (see Fischer, 1973; Spada, 1977). Or, in a test construction setting, the item writer can create items based on a known set of m cognitive operations. In this approach, the values of q_{ij} reflect how the item was constructed (e.g., Fischer & Formann, 1982; Medina-Díaz, 1993).

Thus, the weighting information in **Q** is external to the examinee's dichotomously scored item responses. However, this information is crucial to the estimation of the m η parameters and the n β parameters, because each combination of **Q** and examinee response vectors yields a unique set of parameter estimates. Thus, there is a potential for errors in the specification of the elements of **Q**. For example, a researcher could indicate that a particular cognitive operation is used to answer an item when in fact it is not (an inclusion error). Alternatively, a researcher could omit an operation when it is used to answer an item (an exclusion error). In either case, the misspecification could impact the values of the η parameter estimates, which could then lead to misinterpretation of the relative contribution the cognitive operations make to an item's difficulty.

Thus, the purpose of this study was to examine, using simulation, the effect misspecification of the elements in **Q** has on the estimates of the η and β parameters under the LLTM. The advantage of a simulation approach is that the degree of misspecification in **Q** can be controlled, and the obtained parameter estimates can be compared with those under a no-error condition.

Method

Q Conditions

Two types of **Q** matrices were used—a sparse **Q** matrix condition and a dense **Q** matrix condition. For the **Q** matrices, a 1 indicated that a given cognitive operation was necessary to answer an item correctly, and a 0 indicated that it was not necessary. (The description of the **Q** matrices as "sparse" or "dense" is used here simply to indicate the relative number of cognitive operations involved in the items of each of the tests. From a mathematical point of view, any sparse **Q** matrix can be reformulated as a dense **Q*** matrix by complementing the values of q_{ij}. Then the dense matrix $\mathbf{Q}^* = I - \mathbf{Q}$ will yield the vector of η parameters $\eta_j^* = -\eta_j$. However, such a reformulation would not be consistent with constructing test items from a known set of cognitive operations or identifying those in existing test items.)

Sparse **Q** *matrix.* A test containing 21 items and eight cognitive operations, which was used by Fischer & Formann (1972), was used for the sparse **Q** matrix condition. In this **Q** matrix, only 34 of the 168 cells contained 1s; therefore, this was designated as the *sparse* matrix ($\mathbf{Q}_s$). Under the assumption that Fischer and Formann made no specification errors, this matrix was considered error-free. The elements in $\mathbf{Q}_s$ are shown in Table 1.

Dense **Q** *matrix.* Tests containing complex items can require an examinee to employ many cognitive operations for each item to obtain a correct response. Medina-Díaz (1993) analyzed a 29-item 9th grade algebra test on solving linear equations. Eight production rules were considered the set of cognitive operations needed to answer the items. The number of production rules needed to answer each item ranged from three to seven; multiple items involved the same number of production rules but not necessarily the same rules. Thus, for a given level of item difficulty, each cognitive operation made a smaller relative contribution and the numerical values of the η_j would be smaller than in the sparse **Q** matrix.

In order to compare results based on complex items with those from the $\mathbf{Q}_s$ condition, a test based on Medina-Díaz's (1993) algebra test was created that contained 21 items and eight cognitive operations. The **Q** matrix shown in the lower section of Table 1 was constructed using six items with three production rules, four items with four production rules, five items with five production rules, five items with six production rules, and one item with seven production rules. Because 96 of the 168 weights were 1s, this set of weights was designated as the *dense* **Q** matrix ($\mathbf{Q}_d$).

Procedures

Baseline conditions. For the $\mathbf{Q}_s$ condition, the numerical values of the η parameter estimates obtained by Fischer et al. (1989) were $\eta_1 = 2.152$, $\eta_2 = 1.229$, $\eta_3 = -.468$, $\eta_4 = 1.907$, $\eta_5 = 1.051$, $\eta_6 = .086$, $\eta_7 = .141$, and $\eta_8 = -.474$. For the

Table 1
Weights for the **Q** Matrices

	Cognitive Operation							
Item	1	2	3	4	5	6	7	8
Sparse Matrix ($\mathbf{Q}_s$)								
1	0	0	0	0	0	0	1	0
2	0	0	0	1	0	0	1	0
3	0	0	0	0	0	1	1	0
4	1	0	0	0	0	0	0	1
5	0	1	0	0	0	0	0	1
6	0	0	0	0	1	0	0	1
7	0	0	1	0	0	0	0	1
8	0	1	0	0	0	0	1	0
9	1	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	0
11	0	0	0	1	0	0	0	0
12	0	0	0	0	0	1	0	0
13	0	0	1	0	0	0	1	0
14	0	0	0	0	0	0	0	1
15	0	0	0	1	0	0	0	1
16	0	0	0	0	1	0	1	0
17	0	1	0	0	0	0	0	0
18	0	0	1	0	0	0	0	0
19	0	0	0	0	0	1	0	1
20	1	0	0	0	0	0	0	1
21	0	0	0	0	1	0	0	1
Dense Matrix ($\mathbf{Q}_d$)								
1	1	0	1	0	1	0	0	0
2	0	0	0	1	0	1	0	1
3	0	1	0	1	0	0	1	0
4	1	1	0	0	1	0	0	0
5	0	0	1	0	0	0	1	1
6	0	1	0	1	0	1	0	1
7	1	0	1	0	1	0	1	0
8	1	0	0	1	0	0	1	0
9	0	1	1	0	1	1	0	1
10	1	1	0	1	0	1	0	0
11	1	1	0	1	0	1	0	1
12	0	0	1	1	1	0	1	0
13	0	1	1	0	1	0	1	1
14	1	0	1	1	0	1	0	1
15	1	1	0	1	1	0	1	0
16	0	1	1	1	0	1	1	1
17	1	1	0	1	1	0	1	1
18	1	1	0	1	1	1	0	1
19	1	0	1	1	1	1	1	0
20	1	1	1	0	1	0	1	1
21	1	1	1	0	1	1	1	1

$\mathbf{Q}_d$ condition, the following values of the η parameters were established: $\eta_1 = -.75$, $\eta_2 = -.3$, $\eta_3 = -.1$, $\eta_4 = -.08$, $\eta_5 = -.05$, $\eta_6 = .2$, $\eta_7 = .6$, and $\eta_8 = 1.2$. For this study, these values were considered the true values of η in each **Q** condition. Using these values and the error-free **Q** matrices for each **Q** condition, the 21 β parameters for each **Q** condition were computed using Equation 1.

For both **Q** conditions, GENIRV (Baker, 1986) used the β parameters to generate basal sets of dichotomously scored item responses to the 21-item tests for samples of size 20, 50, 100, and 1,000. Lord (1983) indicated that when small sample sizes are used, the Rasch model is the most appropriate IRT model. Thus, the first three sample sizes (N) represented those often encountered in practical applications of IRT. Because IRT is based on asymptotic results, an N of 1,000 was used to approximate asymptotic conditions.

Using the error-free **Q** matrix from both conditions, the basal dataset for each sample size was analyzed using Fischer et. al's (1989) LLTM microcomputer program. This resulted in estimates of η_j and β_i for each of the four Ns within each **Q** condition. These estimates served as the baseline values in subsequent comparisons. The LLTM computer program yields the logarithm of the product normalized item easiness parameters. The signs of these estimates must be reversed to obtain item difficulty estimates that correspond to those yielded by other IRT computer programs such as BICAL or MULTISCALE.

Misspecification conditions. Misspecification of the elements in a **Q** matrix may have an impact on the estimates of η. In the case of a sparse **Q** matrix, each item's β estimate, obtained using Equation 1, would depend on only a few cognitive operations and any error in the item's vector of weight elements could change the value of the β. In a complex item, such as in a dense matrix, such misspecification could have lesser impact on the value of β.

To introduce such errors, a computer program was written that randomly changed a specified percent of the elements in an error-free **Q** matrix. The program first selected an item at random and then selected a cognitive operation at random and reversed the current value of that cell (0 to 1, or 1 to 0) in **Q**. Thus, both inclusion and exclusion errors were simulated. This resulted in an adjusted **Q** matrix for each error rate. The number of weights changed in the error-free **Q** matrices varied from 2 to 17—that is, the error rate × the number of items (21) × the number of cognitive operations (8).

Analysis

The basal dataset for a given sample size and a given adjusted **Q** matrix was analyzed using the LLTM program, which produced estimates of η_j and the β_i, as well as the value of the log-likelihood function. The root mean square (RMS) values of the difference between the η_js based on the error-free **Q** matrix (the baseline value) and those yielded by the adjusted **Q** matrices for each error rate and sample size condition were computed (within each **Q** condition). The RMS values of the difference between the values of the β_is based on the error-free and the adjusted **Q** matrices also were computed for each error rate and sample size condition (within each **Q** condition).

The lower levels of error (1%, 2%, and 3%) corresponded to a realistic amount of potential misspecification of **Q** with errors of inclusion and exclusion. The higher levels (5%, 7.5%, and 10%) involved a large number of misspecifications that would be unlikely in practice. However, they were used to provide some information about the impact of high levels of error and to determine if there was an interaction of error rate with sample size. Because the errors were randomly introduced into the **Q** matrices, where they fell was important. To evaluate this factor, the overall process described above was replicated 10 times for each sample size-error rate combination within each **Q** condition. The end product for each **Q** condition was 4 × 6 × 10 = 240 sets of η and β estimates and their corresponding RMSs.

Results

Sparse Q Matrix Condition

η *parameters.* The average values of the RMSs for η under the sparse matrix condition are reported in Table 2 and are plotted in Figure 1a. The 1% error rate introduced only 2 errors into $\mathbf{Q}_s$. Yet over the four *N*s, the mean RMS for this error rate ranged from .311 to .459. These values are quite large considering the few errors introduced. For five of the six error rates, the average RMSs were actually larger at $N = 1{,}000$ than at $N = 20$. When $N = 20$, the average RMSs ranged from .364 to .689 over the six error rates. When $N = 1{,}000$, it ranged from .364 to .832. Using the RMSs as the dependent variable, a two-way ANOVA was performed under a fully fixed effects model with sample size and error rate as the main effects and 10 replications per cell. There were significant main effects ($\alpha = .05$) for error rate and sample size but the interaction was nonsignificant. The sample size accounted for 19% of the total sum of squares (SS), but the error rate accounted for 36%. Thus, the sample size effect was considerably less than the error rate effect.

The standard deviations (SDs) of the RMSs for η for each sample size and rate combination for the $\mathbf{Q}_s$ condition also are reported in Table 2. When $N = 20$, the SDs ranged from .111 to .265 over the six error rates. When $N = 1{,}000$, the SDs ranged from .125 to .223. In three of the four sample sizes, the smallest value was yielded by the 3% error rate. The SDs of the RMSs appeared to be insensitive to *N*—low or high SDs were likely to appear at any sample size within an error rate. Within a sample size, there was a slight trend for the SDs to increase as the error rate increased. For example at $N = 50$ and a 1% error rate, SD = .132; at 10% it was .166. At $N = 1{,}000$ and 1%, SD = .140 and at 7.5% SD = .181.

β *parameters.* The average RMSs for the β parameters obtained under the $\mathbf{Q}_s$ condition also are reported in Table 2. Again, the RMS at the 1% error rate was large, ranging from .206 to .324 over the four sample sizes. When $N = 20$, the average RMSs ranged from .278 to .675 over the six error rates. When $N = 1{,}000$, the RMSs ranged from .280 to .571. Although there was a slight trend toward lower average RMSs as sample

Table 2
Mean (M) and SD of RMS Values for η and β Parameter Estimates for the Q_s and Q_d Matrices, at Four Sample Sizes

Matrix, Parameter, and % Error	Sample Size							
	N = 20		*N* = 50		*N* = 100		*N* = 1,000	
	M	SD	M	SD	M	SD	M	SD
Q_s: η Parameter								
1%	.382	.156	.311	.132	.459	.174	.395	.140
2%	.364	.153	.278	.115	.440	.179	.364	.141
3%	.420	.111	.345	.130	.648	.154	.506	.125
5%	.581	.265	.438	.155	.828	.184	.668	.207
7.5%	.655	.218	.523	.172	.901	.167	.737	.181
10%	.689	.179	.553	.166	1.036	.267	.832	.223
Q_s: β Parameter								
1%	.278	.053	.206	.081	.324	.090	.280	.077
2%	.340	.105	.241	.094	.346	.114	.287	.105
3%	.422	.093	.311	.090	.425	.104	.362	.100
5%	.508	.101	.393	.097	.498	.099	.453	.109
7.5%	.621	.106	.453	.100	.566	.104	.515	.101
10%	.675	.129	.514	.098	.611	.141	.571	.126
Q_d: η Parameter								
1%	.260	.119	.158	.080	.163	.080	.134	.066
2%	.297	.194	.170	.071	.126	.052	.114	.052
3%	.340	.219	.208	.145	.216	.127	.171	.129
5%	.345	.077	.238	.064	.220	.069	.191	.077
7.5%	.324	.098	.284	.082	.300	.114	.244	.063
10%	.442	.087	.350	.079	.385	.113	.311	.084
Q_d: β Parameter								
1%	.203	.087	.160	.083	.167	.086	.145	.086
2%	.266	.073	.198	.088	.188	.084	.169	.086
3%	.263	.102	.196	.096	.210	.093	.179	.102
5%	.341	.081	.264	.087	.297	.074	.246	.091
7.5%	.384	.097	.306	.093	.320	.096	.292	.085
10%	.466	.108	.368	.087	.406	.087	.356	.081

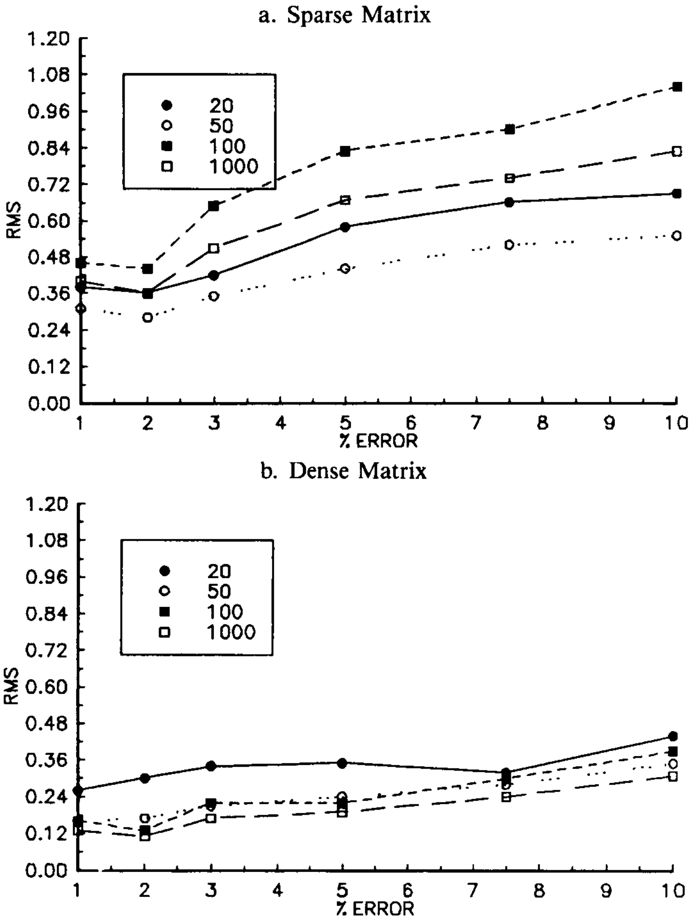
size increased, it was not consistent within error rates. In particular, *N* = 50 yielded the smallest average RMSs over all error rates. A two-way ANOVA was performed and there were significant main effects (α = .05) for error rate and sample size, but the interaction was nonsignificant. Sample size accounted for only 9% of the total SS and error rate accounted for 54%, indicating again that the dominant effect was error rate.

The SDs of the RMSs for β, under the Q_s condition, also are reported in Table 2 by sample size and error rate. These SDs appeared to be somewhat insensitive to either sample size or error rate. When *N* = 20, the SDs ranged from .053 to .129 over the six error rates. At *N* = 1,000, they ranged from .077 to .126. There was no clear pattern to the values of the SDs across sample size and error rates. However, the 10% error rate yielded the largest SDs for all sample sizes, except at *N* = 50.

Dense Q Matrix Condition

η *parameters.* The average RMSs for the η parameters yielded under the Q_d condition are reported in Table 2 and plotted in Figure 1b. For all error rates, there was an overall decrease in RMSs as the sample size increased. For example, at the 1% error rate, the mean RMS decreased from .260 to .134 over the four sample sizes. Within all sample sizes, the values of the aver-

Figure 1
Average RMSs of the η Parameter Estimates



age RMS increased as the error rate increased. When $N = 20$ the values went from .260 to .442 over the six error rates. At $N = 1{,}000$, they increased from .134 to .311. Using the RMSs as the dependent variable, a two-way ANOVA was performed as for the previous conditions. There were significant main effects ($\alpha = .05$) for error rate and sample size, but the interaction was nonsignificant. The sample size accounted for 16% of the total SS, but the error rate accounted for 25%.

The SDs of the average RMS values for the η parameters for the Q_d condition also are reported in Table 2. Within the sample sizes, there was no clear pattern in the values of the SDs as the error rate increased. With the exception of the 7.5% and 10% error rates, there was a general decrease in the SDs as sample size increased. The 3% error rate yielded the largest SDs across all four sample sizes, ranging from .219 to .129. At the 10% error rate, the magnitude of the SDs across the four sample sizes was quite consistent with a SD of .087 for $N = 20$ and .084 for $N = 1{,}000$. However, $N = 100$ yielded a SD of .113.

β *parameters.* The average RMSs for the β parameters obtained under the Q_d condition show that within all sample sizes, there was a general pattern of increase in average RMS with increased error rate (Table 2). When $N = 20$, the values ranged from .203 at the 1% error rate to .466 at the 10% error rate; at $N = 1{,}000$, they ranged from .145 to .356 over the six error rates. As was the case with the η parameters for the $\mathbf{Q}_d$ condition, there was a discernible decrease in the average RMSs across the four sample sizes within all error rates. The two-way ANOVA showed significant main effects ($\alpha = .05$) for error rate and sample size, but the interaction was nonsignificant. Again, the dominant effect was error rate: Sample size accounted for only 8% of the total SS, and error rate accounted for 43%. The SDs of the RMSs for the β estimates also are reported in Table 2. The values were uniformly small, ranging from .073 to .108. Inspection of the values of the SDs did not reveal any pattern as a function of either sample size or error rate.

Discussion

A small degree of misspecification of **Q** had a large impact on the parameter estimates. This impact was indexed by the RMS values of the difference between the parameter estimates obtained when **Q** was misspecified and when it was not. At low rates of misspecification (1%, 2%, and 3%), the sparse **Q** matrix condition yielded an average RMS of .41 for the η parameter estimates. This is quite large considering the relatively few cells (2 to 5 of 168) that were misspecified. At this same level of misspecification, the $\mathbf{Q}_d$ condition yielded an average RMS of .20. At higher levels of misspecification (5%, 7.5%, and 10%), the average RMSs were .70 and .30 for the sparse and dense **Q** matrix conditions, respectively. Thus, at both levels of misspecification, the RMS values yielded under the sparse matrix condition were approximately 1.5 times those obtained under the dense matrix condition. This suggests that the density of **Q** is an important factor in how misspecification affects the η parameter estimates. In a sparse **Q** matrix, only one or two cognitive operations are typically involved in a given item. Thus, even a low level of misspecification alters the structure of the estimation process and distorts the values of the η estimates. In a dense **Q** matrix, a larger number of cognitive operations are involved in each item and a low level of misspecification tends to get ‘‘smoothed out’’ over the test items. Because of this, the consequences of a low level of misspecification were not quite as serious in the dense matrix condition as they were in the sparse matrix condition.

Even though the two-way ANOVAs for the η parameter RMSs indicated a statistically significant sample size effect for both the sparse and dense **Q** matrix conditions, only a small proportion (19% for $\mathbf{Q}_s$ and 16% for $\mathbf{Q}_d$) of the total SS could be attributed to this factor. In both conditions, the error rate factor accounted for a larger proportion (36% for $\mathbf{Q}_s$ and 25% for $\mathbf{Q}_d$) of the total SS.

The ANOVAs based on the β RMSs showed a similar pattern—sample size accounted for 9% ($\mathbf{Q}_s$) and 8% ($\mathbf{Q}_d$), and error rate accounted for 54% ($\mathbf{Q}_s$) and 43% ($\mathbf{Q}_d$) of the SS. Thus, the effect of sample size on the RMSs was minimal, and the dominant effect for both η and β was error rate. The lack of a significant interaction between sample size and error rate in all four ANOVAs was interesting. It suggests that using large samples will not reduce the impact of the error rates on the agreement of the estimates with the baseline values.

An interesting effect of sample size on the η parameter estimates was noticed only in the dense matrix results. Once sample size was greater than 20, it had only a modest impact on the decrease in average values of the RMSs. This suggests that there was a step function operating between $N = 20$ and $N = 50$ that was not present in the sparse matrix results. In the dense matrix results, there was a large decrease in average RMS of η when N increased from 20 to 50. But further increases in sample size resulted in a slight fluctuation about a lower average value of the RMSs. This suggests that in order to obtain proper estimates of the η parameters, there must be some minimum number of examinees correctly responding to those items involving each cognitive operation. In addition, the cognitive operation must appear in a sufficient number of items.

In a sparse matrix, there are only a few items involving the same cognitive operation and in order to obtain enough data relative to each cognitive operation many examinees would be required. However, in a dense matrix, many items involve the same cognitive operations and a smaller sample size would yield sufficient data for each cognitive operation. It appears that in the dense matrix situation, once the minimum sample size was reached, further increases in size had little additional effect on the η parameter estimates. This suggests that increasing the total number of items in tests employing a sparse **Q** matrix might improve the estimation of the η parameters. But due to the usual constraints on test length, this is not normally possible. Overall, the minimal effect of sample size when a sparse **Q** matrix is employed suggests that some parameter recovery studies need to be performed to examine this aspect of parameter estimation in the LLTM.

Given the η parameter estimates and a **Q** matrix, the β parameter estimate for an item can be obtained using Equation 1. There are two sources of error in this estimate that reflect the effect of the misspecification of the elements in the **Q** matrix. First, the values of the η parameters might be in error. Second, the particular combination of η parameters involved in an item can be in error. However, the latter source of error occurs only in those items containing misspecified **Q** matrix elements. Despite these two sources of errors, the general pattern of average RMSs for the β estimates under both **Q** matrix conditions generally paralleled those for the η parameter estimates.

In addition, the ANOVA results for the β estimates yielded results similar to those for the η estimates. The pattern of SDs of the RMSs for η and β were quite similar under both matrix conditions. Thus, it appears to make little difference in interpretation whether the RMSs for the η parameter or the β estimates are analyzed because both provide a similar picture. Given the more fundamental nature of the η parameters, these results should be focused upon.

Several general conclusions can be drawn from these results:

1. Even a small amount of misspecification (1%-3%) of **Q** can have a considerable impact on the values of the η and β parameter estimates.
2. A larger level of misspecification (5%, 7.5%, or 10%) can seriously degrade the parameter estimates to the point where they are not credible.
3. The sample size has a modest effect on the RMSs of the estimates and there was no interaction between sample size and error rate.
4. For a fixed set of cognitive operations, increasing the number of the operations involved in each item of a test tends to reduce the impact of misspecification. Thus, estimates based on misspecified *sparse* **Q** matrices will differ more from the baseline estimates than will those employing *dense* **Q** matrices.

Because it yields estimates of parameters corresponding to the cognitive operations involved in a test, the LLTM provides a bridge between cognitive science and measurement theory. However, the utility of the parameter estimates depends on the appropriateness of the **Q** matrix. Estimates of the η and β parameters yielded by the LLTM are crucially dependent on the definition of the **Q** matrix. Any misspecification in this matrix will be reflected in deviant values of the η parameter estimates and hence in their interpretation. Because specifying the **Q** matrix is a judgmental task, it must be done with great care.

References

Anderson, J. R. (1990). *Cognitive psychology and its implications.* New York: Freeman.

Baker, F. B. (1986). *GENIRV: A computer program for generating item responses.* Unpublished manuscript, University of Wisconsin.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175-193.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Fischer, G. H., & Formann, A. (1972). *An algorithm and a Fortran program for estimating the item parameters of the linear logistic test-model* (Research Bulletin No. 11). Vienna: Institute fur Psychologie der Universitat Wien.

Fischer, G. H., & Formann, A. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6,* 397-416.

Fischer, G. H., Formann, A., & Wild, B. (1989). *Program LLTM.* Unpublished manuscript, Institute fur Psychologie der Universitat Wein.

Fischer, G. H., & Parzer, P. (1991a). LRSM: Parameter estimation for the linear rating scale model. *Applied Psychological Measurement, 15,* 138.

Fischer, G. H., & Parzer, P. (1991b). An extension of the rating scale model with an application to the measurement of change. *Psychometrika, 56,* 637-651.

Frederiksen, N., Mislevy, R. J., & Bejar, I. (1992). *Test theory for a new generation of tests.* Hillsdale NJ: Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

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

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369-380.

Koponen, R. (1983). *An item analysis of tests in mathematics applying logistic test models.* Unpublished doctoral dissertation, University of Jyväsklä, Finland.

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

Lord, F. M. (1983). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 51-61). New York: Academic Press.

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

Medina-Dı́az, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement, 17,* 117-130.

Nährer, W. (1980). Zur Analyse von matrizentestaufgaben mit dem linearen logistischen testmodell [Analysis of test response matrices with the linear logistic test model]. *Zeitschrift fur Experimentelle and Angewandete Psyhchologie, 27,* 553-564.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Sheenan, K., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27,* 255-272.

Spada, H. (1977). Logistic models of learning and thought. In H. Spada & L. F. Kempf (Eds.), *Structural models of thinking and learning* (pp. 227-261). Bern: Huber.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 169-194). Orlando FL: Academic Press.

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement, 5,* 383-397.

Wright, B. D., & Linacre, J. M. (1984). *MICROSCALE manual* [Computer program manual]. Westport CT: Mediax Interactive Technologies.

Wright, B. D., & Mead, R. J. (1978). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23C). Chicago IL: University of Chicago, Statistical Laboratory.

Author's Address

Send requests for reprints or further information to Frank B. Baker, Department of Educational Psychology, Educational Sciences Building, University of Wisconsin, Madison WI 53705, U.S.A. E-mail: fbaker@macc.wisc.edu.bitnet