Standard Errors of Item Parameter Estimates in Incomplete Designs

Dato N. M. de Gruijter
University of Leyden

Lord and Wingersky (1985) derived the asymptotic variance-covariance matrix for item and person parameters in item response models, using maximum likelihood estimation. Their results can be used in incomplete designs, in which different test forms with common subtests are administered to different groups of examinees. It is also possible to estimate the accuracy of various designs beforehand, which enables the researcher to select the best of several designs under consideration. The possibilities are demonstrated for the one- and two-parameter models. *Index terms: Asymptotic standard errors, Incomplete designs, Item banking, Information matrix, Maximum likelihood, Rasch model, Two-parameter logistic model.*

In applications of item response theory (IRT), the item information matrix has frequently been used in order to obtain the asymptotic standard errors of item parameter estimates. The use of this matrix in connection with the maximum likelihood (ML) estimation of parameters implies the assumption that person parameters are known (de Gruijter, 1984; Thissen & Wainer, 1982).

Lord and Wingersky (1985) presented the joint information matrix for item and person parameters, from which the asymptotic variance-covariance matrix for ML estimates of item and person parameters can be derived. They demonstrated that the variances and covariances depend on the restrictions used to fix the latent ability (θ) scale. In the Rasch model, for example, the latent scale can be fixed by setting one item parameter equal to 0. The errors in the parameter estimates are measured relative to this parameter. In general, the resulting error variances are larger in this frame of reference than those that would result from a scale fixed by setting the sum of all item parameters equal to 0. Thus the error variances and covariances for a particular fixation of the latent scale are relevant only in connection with parameter estimates obtained with the same fixation of the latent scale.

Lord and Wingersky's method can be applied with incomplete data when the estimation of all parameters is done in one computer run. The method was used by Wingersky and Lord (1984) in a study in which two groups of examinees were administered different test forms with a common subtest. They varied the number of items in the common subtest and found that the error variances for the unique items increased notably only when the number of common items became quite small. The lengths of the common subtest and the total test were confounded in this study, making these results even more remarkable.

Aside from the results on the importance of the choice of a scale, the method seems especially suitable for incomplete designs in equating studies and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 2, June 1988, pp. 109–116

0146-6216/88/020109-08$1.65

in item bank construction, when all parameters are simultaneously estimated by ML. In an incomplete design, the assumption of known θs on a given scale seems less tenable than in a complete design. In the extreme case of no overlap between test forms, the assumption of known θs would result in seemingly reasonable derived error variances, but estimation of all parameters on a common scale is impossible in reality.

An alternative to joint maximum likelihood is marginal maximum likelihood (MML). MML avoids some of the estimation problems which arise in the joint estimation of item and person parameters. However, for a discussion of MML in the present context, more should be known about the performance of estimation procedures based on MML in the case of incomplete data where different examinee groups are supposed to come from different examinee populations. Furthermore, results on the effects of scale fixation in ML apply as well to MML when there are a number of populations. For these reasons the present study was restricted to ML.

Unfortunately, application of Lord and Wingersky's method is hampered by the computational burden when the number of examinees is large, although grouping of θs into classes enabled them to invert a large matrix, with dimensions equal to the number of examinees, in parts. De Gruijter (1985) and Levine (1985) suggested modifications in order to make application of the method more feasible. In the present study similar modifications were introduced in connection with incomplete designs. Both the one-parameter Rasch model, in which one constraint is needed in order to fix the latent scale, and the two-parameter logistic model, in which two constraints are needed, were used. The three-parameter model, which presents no new problems with respect to scale fixation, is not discussed.

The procedure can be used with real data, in which case parameter estimates are used in the formulas instead of the unknown true parameters. Alternatively, apparently reasonable item and person parameters for hypothetical items and persons can be selected in order to obtain information on the accuracy of a particular design before any data are obtained. When this is done for several competing designs, the results can help a researcher decide which to use. In the present study all examples are based on hypothetical situations.

The Asymptotic Item Error Matrix for the Rasch Model

In the Rasch model the probability of a correct response by examinee a to item i can be written as

$$P_{ia} = \{1+\exp[a(b_i-\theta_a)]\}^{-1} \quad , \tag{1}$$

where b_i is the item difficulty parameter,
$a = 1$ is the common slope, and
θ_a is the person parameter.

With n items and N examinees there are $M = n + N$ parameters. In a complete design, estimation accuracy of ML parameter estimation is reflected by the $M \times M$ information matrix for item and person parameters,

$$\mathbb{H} = \begin{bmatrix} \mathbb{S} & \mathbb{F} \\ \mathbb{F}' & \mathbb{T} \end{bmatrix} \quad , \tag{2}$$

where $\mathbb{S}$ is the sum of N $n \times n$ diagonal matrices $\mathbb{S}_a$ with elements

$$s_{ii(a)} = D_{ia}^{-1}\left(\frac{\partial P_{ia}}{\partial b_{ia}}\right)^2 = D_{ia} \quad , \tag{3}$$

where $D_{ia} = P_{ia}(1 - P_{ia})$. The diagonal elements of $\mathbb{S}$ contain the Rasch model item informations, the inverse values of which frequently are used as error variances. Matrix $\mathbb{T}$ in Equation 2 is an $N \times N$ diagonal matrix with elements

$$t_{aa} = \sum_i D_{ia}^{-1}\left(\frac{\partial P_{ia}}{\partial \theta_a}\right)^2 = \sum_i D_{ia} \quad , \tag{4}$$

and $\mathbb{F}$ has elements

$$f_{ia} = D_{ia}^{-1}\left(\frac{\partial P_{ia}}{\partial b_i}\right)\left(\frac{\partial P_{ia}}{\partial \theta_a}\right) = -D_{ia} \quad . \tag{5}$$

In order to fix the latent scale, one restriction is needed. The choice of the restriction is important, as it has an impact on the error variance-covariance matrix beyond the mere fact that different restrictions imply different scalings (Lord & Wingersky, 1985). In the Rasch model the obvious choice is

to set the average b equal to 0. Parameter b_n can then be regarded as a function of the other bs and accordingly can be eliminated. The introduction of the restriction can be viewed as a transformation to new parameters $b_1^*, b_2^*, \ldots, b_{n-1}^*$ which are related to the old parameters according to the equation

$$\mathbb{b} = \mathbb{E}\mathbb{b}^* = \begin{bmatrix} \mathbb{I} \\ \cdots\cdots\cdots\cdots \\ -1 \; -1 \ldots -1 \end{bmatrix} \mathbb{b}^* \quad , \tag{6}$$

where $\mathbb{I}$ is the $(n-1) \times (n-1)$ identity matrix. The transformation results in an $(M-1) \times (M-1)$ information matrix,

$$\mathbb{H}^* = \begin{bmatrix} \mathbb{S}^* & \mathbb{F}^* \\ \mathbb{F}^{*\prime} & \mathbb{T} \end{bmatrix} = \begin{bmatrix} \mathbb{E}'\mathbb{S}\mathbb{E} & \mathbb{E}'\mathbb{F} \\ \mathbb{F}'\mathbb{E} & \mathbb{T} \end{bmatrix} \quad . \tag{7}$$

The $(n-1) \times (n-1)$ leading matrix of $\mathbb{H}^{*-1}$, $\mathbb{C}^*$, gives the asymptotic variance-covariance matrix for the first $n-1$ items. This matrix can be obtained from

$$\begin{aligned} \mathbb{C}^{*-1} &= \mathbb{S}^* - \mathbb{F}^*\mathbb{T}^{-1}\mathbb{F}^{*\prime} \\ &= \mathbb{E}'(\mathbb{S} - \mathbb{F}\mathbb{T}^{-1}\mathbb{F}')\mathbb{E} \\ &= \mathbb{E}'\left[\sum_{a=1}^{N} (\mathbb{S}_a - t_{aa}^{-1}\mathbb{f}_a\mathbb{f}_a')\right]\mathbb{E} \quad , \end{aligned} \tag{8}$$

where $\mathbb{f}_a$ is the ath column of $\mathbb{F}$. In this way the problem of obtaining the variance-covariance matrix for item parameters reduces to the problem of inverting a symmetric matrix with dimensions equal to $n-1$; a large value of N is no problem, although a further simplification can be obtained by grouping the person parameters.

The Bayesian analysis with a prior for the bs (Swaminathan & Gifford, 1982) can be contrasted with ML. In the Bayesian analysis, the bs usually have independent priors with a common, fixed mean and a common variance. This is enough to fix the latent scale. The prior structure with independent priors does not, however, produce a proper solution for the error variance-covariance matrix of item parameter estimates, a matrix in which the elements sum to 0, reflecting the restriction $\sum b = 0$. With a small number of items this might be a problem.

The derivation of $\mathbb{C}^{*-1}$ for incomplete designs proceeds in a way similar to the derivation for the complete design. In the general case, there are K test forms linked by common subtests. The tests are administered to K different groups of examinees. Let $\mathbb{e}_k$ $(k = 1, \ldots, K)$ be a vector of length n, with element i equal to 1 when item i is included in test k and 0 otherwise. All elements D_{ia} used in Equation 2 must be replaced by $e_k(i)D_{ia(k)}$, where $a(k)$ designates examinee a in group k. Matrix $\mathbb{F}$ can be partitioned as $[\mathbb{F}_1, \ldots, \mathbb{F}_K]$ and $\mathbb{T}$ can be written as $\mathbb{T} = \text{diag}[\mathbb{T}_1, \ldots, \mathbb{T}_K]$.

The equivalent of Equation 8 for the incomplete design can be written as

$$\begin{aligned} \mathbb{C}^{*-1} &= \mathbb{E}'(\mathbb{S} - \mathbb{F}\mathbb{T}^{-1}\mathbb{F}')\mathbb{E} \\ &= \mathbb{E}'\left[\sum_k (\mathbb{S}_k - \mathbb{F}_k\mathbb{T}_k^{-1}\mathbb{F}_k')\right]\mathbb{E} \\ &= \mathbb{E}'\left[\sum_k \sum_{a(k)} (\mathbb{S}_{a(k)} - t_{aa(k)}^{-1}\mathbb{f}_{a(k)}\mathbb{f}_{a(k)}')\right]\mathbb{E} \quad . \end{aligned} \tag{9}$$

After $\mathbb{C}^*$ has been computed, the variance-covariance matrix for all n parameters is obtained as

$$\mathbb{C} = \mathbb{E}\mathbb{C}^*\mathbb{E}' \quad . \tag{10}$$

The Asymptotic Item Error Matrix for the Two-Parameter Model

In the two-parameter logistic model, items differ not only with respect to the difficulties b, but also with respect to the slope parameters a_i. This has two consequences. First, matrix $\mathbb{H}$ also contains derivatives with respect to these parameters; its size is $M \times M$, where $M = 2n + N$ in a complete design. Second, two restrictions are needed in order to fix the latent scale. Lord and Wingersky (1985) began with the assumption that there are real differences in b values. They suggested setting the mean b of a number of discriminating, moderately easy items equal to 0, and setting the mean b of a number of discriminating, moderately difficult items equal to 1. These restrictions correspond to a transformation in which one b parameter from the easy items and one b parameter from the difficult items are eliminated:

$$\begin{bmatrix} \mathbb{b} \\ \mathbb{a} \end{bmatrix} = \mathbb{E}\begin{bmatrix} \mathbb{b}^* \\ \mathbb{a}^* \end{bmatrix} \quad , \tag{11}$$

where $\mathbf{b}^*$ has $n-2$ elements and $\mathbf{a}^*$ has n elements. The transformation on matrices $\mathbb{S}$ and $\mathbb{F}$ in Equation 7, resulting in matrices $\mathbb{S}^*$ and $\mathbb{F}^*$, is easily implemented. Arguments will be given for another choice, however.

Consider the following alternative restrictions: $\Sigma b = 0$, $\Sigma a = n$. With these restrictions the $(2n-2) \times (2n-2)$ matrix $\mathbb{C}^{*-1}$ can be partitioned as

$$\mathbb{C}^{*-1} = \begin{bmatrix} \mathbb{P}_{11} & \mathbb{P}_{12} \\ \mathbb{P}_{21} & \mathbb{P}_{22} \end{bmatrix} , \quad (12)$$

where $\mathbb{P}_{11}$ is the $(n-1) \times (n-1)$ matrix corresponding to parameters b^*, and $\mathbb{P}_{22}$ is the matrix corresponding to parameters a^*. When all as are equal to 1, $\mathbb{P}_{11}$ is equal to $\mathbb{C}^{*-1}$ in Equation 8 for the one-parameter model. Again using a well-known result on partitioned matrices, the $(n-1) \times (n-1)$ leading matrix of $\mathbb{C}^*$, $\mathbb{C}^*_{11}$, can be obtained as

$$\mathbb{C}^*_{11} = \mathbb{P}_{11}^{-1} + \mathbb{P}_{11}^{-1}\mathbb{P}_{12}\mathbb{C}^*_{22}\mathbb{P}_{21}\mathbb{P}_{11}^{-1} , \quad (13)$$

where

$$\mathbb{C}^*_{22} = (\mathbb{P}_{22} - \mathbb{P}_{21}\mathbb{P}_{11}^{-1}\mathbb{P}_{12})^{-1} . \quad (14)$$

Hence the error variance-covariance matrix for b^*, and for that matter b, can be written as the sum of a matrix for known as (which for $a_i = a = 1$ means the one-parameter model) and a term with contributions due to the fact that the as must also be estimated. The total inverse of $\mathbb{C}^{*-1}$ can be obtained using

$$\mathbb{C}^*_{12} = -\mathbb{P}_{11}^{-1}\mathbb{P}_{12}\mathbb{C}^*_{22} . \quad (15)$$

The computation of $\mathbb{C}^*$ using the partition of Equation 12 has another advantage: Two $(n-1) \times (n-1)$ matrices must be inverted instead of a $(2n-2) \times (2n-2)$ matrix.

The two different sets of restrictions are related by a linear transformation. The relationship between the corresponding error matrices is more complex, as mentioned above; in order to obtain an error matrix for another specification of the latent scale, the "delta" method (Kendall & Stuart, 1969, chap. 10) must be used.

The Error Matrix for Person Parameters

The matrix result used in Equation 13 can also be used in order to find the asymptotic error matrix for person parameters:

$$\mathbb{V} = \mathbb{T}^{-1} + \mathbb{T}^{-1}\mathbb{F}^{*\prime}\mathbb{C}^*\mathbb{F}^*\mathbb{T}^{-1} = \mathbb{T}^{-1} + \mathbb{T}^{-1}\mathbb{F}'\mathbb{C}\mathbb{F}\mathbb{T}^{-1} , \quad (16)$$

where $\mathbb{C}$ is defined as in Equation 10. The error variances are

$$v_a = t_a^{-1} + t_a^{-2}\mathbf{f}_a'\mathbb{C}\mathbf{f}_a , \quad (17)$$

where t_a^{-1} equals the traditional estimate of the error variance for θ_a. Equation 17 is similar to an equation for complete designs given by Wright and Panchapakesan (1969). Their Equation 29 is based on the assumption that $\mathbb{C}$ is diagonal.

Errors in linking test forms may result in correlated errors for item parameter estimates of items in a particular test form, and consequently in correlated errors in person parameter estimates. In incomplete designs, the second factor in Equations 16 and 17 should not be simply neglected. However, when the second factor is negligible for vectors $\mathbf{f}$ based on different θ values and different item selections, the traditional test information function and its useful additive property (i.e., the fact that it is the sum of item information functions) can be used in various applications.

Illustrative Examples

The Effect of Common Subtest Size in the Rasch Model

Wingersky and Lord (1984) found that the common subtest in linking two tests could be rather small without the results deteriorating notably. In their study, common subtest size was confounded with total test length. In the present simulation study the effect of subtest size was investigated for fixed test length.

The simulation considered two 30-item tests with all b values equal to 0, and two examinee groups with 100 examinees each. The two θ distributions were identical, with mean $\theta = 0$ and a variance close to 1. The overlap between the two test forms

Table 1
Variances of Item Parameter Estimates

Number of Items		Error Variance	
Total	Common	Unique Items	Common Items
59	1	.072	.024
58	2	.060	.024
57	3	.056	.024
56	4	.054	.024
55	5	.053	.024
50	10	.050	.024
40	20	.048	.024

was varied, with common subtest sizes equal to 1, 2, 3, 4, 5, 10, and 20 items.

The resulting error variances are given in Table 1. The values can be compared with those for a complete design. In a complete design the error variance would be about .048 for a sample size of 100 examinees (the exact value depends on the number of items). The variance for the unique items is always at least twice the value of the common item variance: The estimates for unique items are based on half the number of examinees. Additionally, there is extra error variance due to inaccurate linking of test forms. The error variance of the common items does not change much with an increase in the strength of the link between the two tests: Within test forms the accuracy of the θ estimates, needed for the estimation of item parameters, does not change.

The linking effect is, of course, strongest with only one common item. In this case, the excess variance amounts to about half the value to be expected without linking errors. This is clarified in Figure 1, where two three-item tests with one common item are schematically represented. In the middle section, a random error is introduced in the common item: On the basis of the response in the first examinee group, the item seems too difficult compared with the other items in the first test. The lower section of the figure shows how this error is redistributed over all items after the rescaling $\Sigma b = 0$. When the tests are of equal length and have a large number of items, the errors have a size of approximately $\frac{1}{2}d$, where d is the size of the original error: The items in the first test have a bias of $\frac{1}{2}d$ in one direction, and those in the other test have a bias of $\frac{1}{2}d$ in the other direction.

Now think of d as a random variable, the difference between the difficulties of the common item in Test 1 and Test 2. Its variance should be equal to $\text{var}(\hat{b}_1 - \hat{b}_2) \approx 2\ \text{var}(\hat{b})$, twice the error variance of b for the common item in one group. The linking variance for two equally sized tests is .25 times this value, or half the variance of the common item based on one group. For the data of Table 1 this variance is .024. The total error variance for unique items equals .048 + .024 = .072.

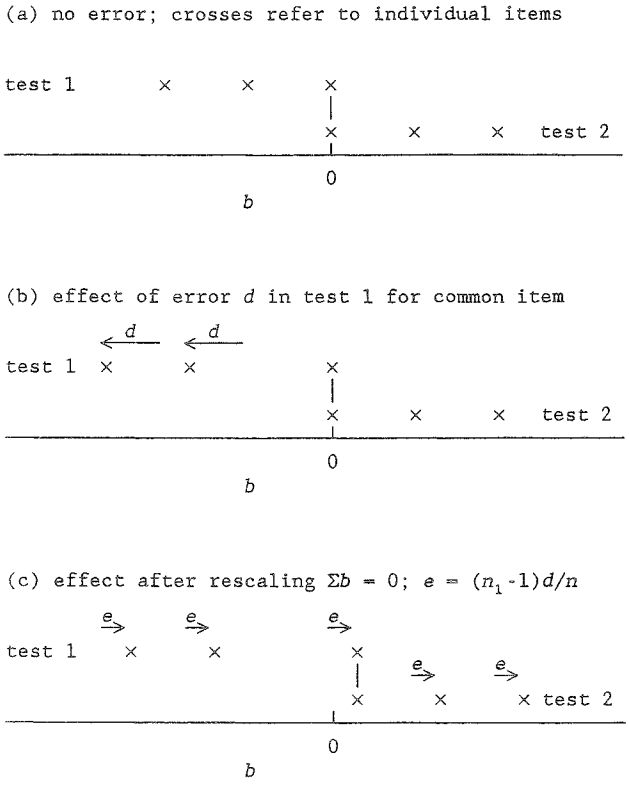
The effect can be demonstrated more clearly with another example. Set the b value of the common item equal to 1.00. In this case the error variance for the common item is .028. The error variance for the other items is .076. This value results from the addition of .048 (the error variance for a complete design and $N = 100$) and .028 (half the error variance of the common item for $N = 100$).

From this example it is clear that the linking error diminishes quickly with an increase in the number of common items, at least in a joint analysis of all data. The results of Wingersky and Lord are corroborated.

Alternative Designs With More Than Two Test Forms

When there are more than two test forms, the

Figure 1
Linking Error With Error in the Common Item
(n is the Total Number of Items;
$n_1 - 1$ is the Number of Unique Items in Test 1)



way in which tests overlap can be varied while holding the total overlap constant. The overlap may have an impact on the size of the error variances of the difficulty parameters, as will be demonstrated.

Three groups of 100 examinees with identical θ distributions and three 30-item tests with b equal to 0 were selected. Error variances for two designs were computed, and are presented in Tables 2 and 3. In the first design, Tests 1 and 2 were linked by items 26 through 30, and Tests 2 and 3 by items 51 through 55. There was no direct link between Test 1 and Test 3. This had an effect on the error variances of the unique items in these two tests, which are larger than the error variances of the unique items in Test 2. In the second design, all tests overlapped through the same common subtest. The largest error variance was smaller than the largest error variance in the first design. In vertical equating, when the examinee groups differ in average ability, other results might be expected.

Comparison of a Rasch Analysis and a Two-Parameter Analysis

The error variances in this study were derived under the condition that the model assumptions were fulfilled. The error variances under two models can be compared only when both models are true, which in this connection implies the validity of the Rasch model. In this section, violation of model assumptions is briefly discussed, but first the one-

Table 2
Variances of Item Parameter Estimates in Design 1

Item	Test Form	Error Variance
1-25	1	.059
26-30	1, 2	.027
31-50	2	.053
51-55	2, 3	.027
56-80	3	.059

and two-parameter models are compared given the validity of the one-parameter model.

The comparison involved two groups of 100 examinees (again with identical θ distributions) and two hypothetical 30-item tests. In the first test, 10 items had $b = -1.0$; in the second test, 10 items had $b = 1.0$. The remaining items, among which five items were common, had $b = 0$. Table 4 presents the theoretical error variances for the difficulty parameters, using Equation 13, based on the restrictions $\sum b = 0$ and $\sum a = n$. The impact of errors in as on the error variances of bs is clear. Should the Rasch model be true, a Rasch analysis should be performed rather than a two-parameter analysis.

Unfortunately, in practice, the true model is unknown. In many cases, however, items are likely to differ somewhat in discrimination and the Rasch model analysis is in error. This error can be serious in incomplete designs when the different examinee groups have various ability levels, as in vertical equating. This is demonstrated in Figure 1, which displays a design with only one common item. Assume that all items except this item conform to the Rasch model. When it has a steeper slope (it is likely that the common items are relatively more discriminating), it will be relatively easy in a higher-ability group and relatively difficult in a lower-ability group. There is only one parameter estimate for the common item; the different behavior of the item in the two groups appears in systematic differences between $\hat{b}$s of the other items in both test forms, and consequently between $\hat{\theta}$s of the two different groups. From Figure 1 it can be concluded that the systematic difference equals roughly the size of the difference between the apparent difficulties of the item.

Table 3
Variances of Item Parameter Estimates in Design 2

Item	Test Form	Error Variance
1-5	1, 2, 3	.016
6-30	1	.055
31-55	2	.055
56-80	3	.055

Table 4
Variances of Difficulty Parameters in a Two-Parameter Analysis

	Error Variance	
Items	α fixed	α free
Common	.024	.027
Unique: $b = 0.0$	.060	.062
Unique: $b = \pm 1.0$	.068	.110

Discussion

With an incomplete design it seems important to derive the error variance-covariance matrix for item and/or person parameters from the joint item and person parameter information matrix, as suggested by Lord and Wingersky (1985). Here it has been demonstrated that the item error matrix can be obtained rather easily, even when there is a large number of examinees in an incomplete design. Next, the error variances for person parameters can easily be obtained.

Marginal maximum likelihood estimation involves estimation of distribution parameters rather than person parameters. Nevertheless, some of the arguments presented here apply equally well to MML with item parameters and population parameters for a number of populations. Also in this estimation procedure the item parameter error matrix depends on all parameters and on the types of restrictions used to fix the θ scale. With MML a researcher might decide to place restrictions on the population parameters. In the Rasch model, for example, the sum of the population means might be set equal to 0.

The robustness of the parameter estimates under violation of the model assumptions needs more

careful study. In incomplete designs, deviations from the model assumptions can result in more or less serious errors.

References

de Gruijter, D. N. M. (1984). A comment on "Some standard errors in item response theory". *Psychometrika, 49,* 269–272.

de Gruijter, D. N. M. (1985). A note on the asymptotic variance-covariance matrix of item parameters in the Rasch model. *Psychometrika, 50,* 247–249.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (Vol. I, 3rd ed.). New York: Hafner.

Levine, M. V. (1985). Discussion. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Lord, F. M., & Wingersky, M. S. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175–191.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, University of Leyden, Boerhaavelaan 2, 2334 EN Leyden, The Netherlands.