The Use of Item Statistics in the Calibration of an Item Bank

Dato N. M. de Gruijter
University of Leyden

An IRT analysis based on p (proportion correct) and r (item-test correlation) is proposed for a group of tests having items in common. The procedure is a generalization of a procedure proposed by De Gruijter and Mooijaart (1983) which is related to procedures for the factor analysis of dichotomous data. The procedure results in IRT item parameters using data from examinee groups with subsets of common items; it is, therefore, particularly appropriate for calibrating items for use in small-scale item banks. Simulated data are used to illustrate the procedure.

In many item banks the statistics p, item proportion correct, and r, item-test correlation, are stored as item indices. The values of these statistics depend on the group which took the test. They are, therefore, imperfect indices of item difficulty and item discriminating power.

In order to obtain indices which are independent of a specific group, an analysis based on an item response theory (IRT) model is required. One such model is the three-parameter logistic model, in which the probability of a correct response on item i given latent ability θ equals

$$P_i(\theta) = c_i + (1 - c_i)/\{1 + \exp[-Da_i(\theta - b_i)]\} \quad , \tag{1}$$

where b_i is the difficulty parameter of item i,
a_i is the item discrimination parameter,
c_i is the pseudo-guessing parameter, and
D is a constant equal to 1.7.

The item parameters a_i, b_i, and c_i ($i = 1, \ldots, n$) of the items in an n-item test and the person parameters θ (i.e., characteristics of the distribution of θ) are estimated from the observed item scores. The latent scale has an interval scale characteristic: Values $a^* = x^{-1}a$, $b^* = xb + y$, and $\theta^* = x\theta + y$ satisfy Equation 1, as do the original values a, b, and θ. Therefore, when the parameters must be estimated, two restrictions on the parameters are needed in order to fix the latent scale. This can be done, for example, by setting the mean θ to 0 and the standard deviation to 1.

The item statistics p and r can be used in the estimation of the item parameters. Under the assumption that θ has a standard normal distribution, Urry (1974) gave formulas for the estimation of a and b from p and r, with corrections for attenuation and spuriousness if needed, given a value for c. The estimates of a, b, and c are used as starting values in estimation programs such as ANCILLES (Urry, 1976).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 3, September 1986, pp. 231–237

0146-6216/86/030231-07$1.60

The IRT analysis in ANCILLES is restricted to items in a single test form. ANCILLES shares this restriction with most procedures for the estimation of item parameters. When several test forms have been used, a procedure must be chosen which guarantees that the estimates $\hat{a}$ and $\hat{b}$ for all items are on a common scale. When test forms have items in common, this goal can be achieved in various ways. One possibility is to fix the b values in new analyses to the values obtained in a previous analysis. Other procedures use the item characteristic curves (Haebara, 1980; Stocking & Lord, 1983). In an analysis by Petersen, Cook, and Stocking (1983), these IRT procedures did not appear to be as efficient as concurrent calibration, i.e., the simultaneous estimation of all parameters. This is not too surprising: It is doubtful that the a and b estimates can be brought on a common scale when the value of c for an item may differ from occasion to occasion, whereas fixing c may result in other problems when the fixed value does not coincide with the true value.

Here, a concurrent analysis for the three-parameter logistic model with a common value for c is proposed. The procedure uses only information residing in p and r, and is therefore suitable with item banks in which these item statistics are stored. It is a generalization to more test forms of a procedure suggested by De Gruijter and Mooijaart (1983).

De Gruijter and Mooijaart's Proposal

De Gruijter and Mooijaart (1983) proposed to estimate item parameters from item proportion correct p_i and marginal proportions for item pairs p_{ij}, where p_{ij} is the proportion of examinees with both i and j correct. The procedure is related to factor analysis of dichotomous data (Christoffersson, 1975; Muthén, 1978).

Given a latent trait model for $P_i(\theta)$, the model proportions correct p_i^* and marginals for item pairs p_{ij}^* are

$$p_i^* = \int p_i(\theta)g(\theta)d\theta \qquad (i = 1, \ldots, n) \tag{2}$$

and

$$p_{ij}^* = \int P_i(\theta)P_j(\theta)g(\theta)d\theta \qquad (i = 2, \ldots, n; j = 1, \ldots, i-1) \quad , \tag{3}$$

where $g(\theta)$ is the density of the latent ability distribution and n is the test length. The density can be approximated by a discrete distribution with K latent classes (Bock & Aitkin, 1981). The standard normal distribution, for example, can be approximated by a discrete distribution with K values θ_k and K relative frequencies v_k using Gaussian quadrature formulas (Stroud & Secrest, 1966). The proportions p_i^* and p_{ij}^* can be rewritten as

$$p_i^* = \sum_{k=1}^{K} v_k P_i(\theta_k) \qquad (i = 1, \ldots, n) \tag{4}$$

and

$$p_{ij}^* = \sum_{k=1}^{K} v_k P_i(\theta_k)P_j(\theta_k) \qquad (i = 2, \ldots, n; j = 1, \ldots, i-1) \quad . \tag{5}$$

De Gruijter and Mooijaart proposed to obtain item parameters which minimize the differences between observed values p_i and p_{ij} and model proportions p_i^* and p_{ij}^*. They also considered the possibility of estimating the discrete distribution of θ; in the present context the distribution is assumed to be fixed.

In order to be able to analyze large tests, De Gruijter and Mooijaart chose ordinary least squares (OLS) instead of generalized least squares. Muthén (1978) reported satisfactory results with OLS. Con-

sequently, the function to be minimized with respect to the item parameters is

$$F = N\left[\sum_{i=1}^{n}(p_i-p_i^*)^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1}(p_{ij}-p_{ij}^*)^2\right] \quad , \tag{6}$$

where N is the number of examinees. De Gruijter and Mooijaart actually used a modification of F in Equation 6, but the modification is not relevant in the present context.

Simultaneous Estimation of Item Parameters With Several Test Forms

Assume that $L > 1$ test forms have been administered to L groups of examinees. The tests have items in common. The examinee groups may differ, but it is assumed that the distributions of θ have the same form, that is,

$$\theta_{k\ell} = \mu_\ell + \sigma_\ell\theta_k \qquad (\ell = 1, \ldots, L) \quad . \tag{7}$$

The values θ_k and v_k are chosen in such a way as to approximate the standard normal distribution. The μs and σs must be estimated along with the item parameters. In order to fix the latent interval scale, two restrictions are needed. For the sake of simplicity the following restrictions are used:

$$\sum_{\ell=1}^{L}\mu_\ell = 0 \tag{8}$$

and

$$\sum_{\ell=1}^{L}\sigma_\ell = L \quad . \tag{9}$$

The item and distribution parameters can be obtained from the minimization of the function

$$F = \sum_{\ell=1}^{L} N_\ell\left[\sum_{i=1}^{n\ell}(p_{i\ell}-p_{i\ell}^*)^2 + \sum_{i=2}^{n\ell}\sum_{j=1}^{i-1}(s_{ij\ell}-s_{ij\ell}^*)^2\right] \quad , \tag{10}$$

where $p_{i\ell}$ is the proportion correct of the ith item in test ℓ, the test administered to examinee group ℓ, and $s_{ij\ell}$ is the covariance of the ith and the jth items in test ℓ. F in Equation 10 is a generalization to the situation with more than one test form of F in Equation 6, with cross-products p_{ij} replaced by covariances s_{ij}. The generalization is similar to the analysis of covariance structures with structural means (Sörbom, 1982).

The procedure can be simplified by replacing the $n_\ell - 1$ covariances where item $i\ell$ is involved, by their average

$$s_{i.\ell} = (n_\ell-1)^{-1}\sum_{j\neq i} s_{ij\ell} \quad . \tag{11}$$

This allows the selection of

$$F = \sum_{\ell=1}^{L} N_\ell \sum_{i=1}^{n\ell}[(p_{i\ell}-p_{i\ell}^*)^2 + (n_\ell-1)(s_{i.\ell}-s_{i.\ell}^*)^2] \tag{12}$$

as the function to be minimized. The purpose of the simplification is twofold. First, it agrees with the frequent use of p and r as item statistics in item banks. The average covariance $s_{i.\ell}$ is easily obtained from the item-total correlation of the ith item in test ℓ:

$$(n_\ell-1)s_{i.\ell} = s_{i\ell}(s_{t\ell}r_{it\ell} - s_{i\ell}) = s_{i\ell}\left[\left(\sum_{j=1}^{n\ell} s_{j\ell}r_{jt\ell}\right)r_{it\ell} - s_{i\ell}\right] \quad , \tag{13}$$

where $s_{i\ell} = p_{i\ell}^{1/2}(1 - p_{i\ell})^{1/2}$, (14)

$s_{i\ell}$ is the standard deviation of total test scores on test ℓ, and

$r_{it\ell}$ is the item-total correlation of item i in test ℓ.

Secondly, Equation 12 has fewer terms than Equation 10; the iterations in the estimation procedure can therefore be performed more quickly. This is important when the use of microcomputers is considered.

The large reduction in terms reduces estimation accuracy. For this reason, only a common value for c is estimated. Further, it is important that there are real differences between at least some examinee groups, because when there are no group differences only two different statistics are available for each item. This is enough to estimate the difficulty and discrimination parameters in the two-parameter model, but insufficient when information on a third parameter in the three-parameter model is also needed. Information on group differences is obtained in the noniterative procedure to obtain starting values, as discussed below.

A further reduction of terms might be considered in some circumstances. In multiple-matrix sampling where each examinee is administered only one item, covariances between items are not obtained. Mislevy (1983) and Reiser (1983) proposed item response models for grouped data in such a sampling design, where item parameter estimates are obtained from p. These authors proposed to use maximum likelihood estimation. Notice, however, that only in some special cases can the grouped data models be related to IRT models on the individual level (Mislevy, 1983).

Starting Values

There are several ways to obtain reasonable starting values for the three-parameter logistic model with a common starting value c. One approach is first to choose a value c and to use formulas similar to Urry's (1974) formulas for the normal ogive in order to obtain estimates $\hat{a}$ and $\hat{b}$ for each test under the assumption of a standard normal distribution of θ. Linear transformations of the latent scales of the different tests are needed in order to obtain a common scale. The necessary transformations are

$$b'_{i\ell} = \sigma_\ell b_{i\ell} + \mu_\ell \tag{15}$$

and

$$a'_{i\ell} = \sigma_\ell^{-1} a_{i\ell} \quad , \tag{16}$$

under the restrictions $\sum\mu_\ell = 0$ and $\sum\sigma_\ell = L$.

The values σ_ℓ ($\ell = 1, \ldots, L$) and μ_ℓ ($\ell = 1, \ldots, L$) must be chosen to ensure the similarity of the multiple item parameter estimates of items which have been used in more than one test form. More specifically, the values σ and μ are chosen in such a way that the function

$$G = \sum_i \Big[\sum_{\ell | i \in S\ell} N_\ell (b'_{i\ell} - b'_{i.})^2 \Big] \quad , \tag{17}$$

where

$$b'_{i.} = c_i \sum_{\ell | i \in S\ell} N_\ell b'_{i\ell} \tag{18}$$

and

$$c_i^{-1} = \sum_{\ell | i \in S\ell} N_\ell \quad , \tag{19}$$

is minimized (De Gruijter, 1986). Index i in Equation 16 refers to item i, not to the ith item in test ℓ as in the previous section. $S\ell$ is the set of all items in test ℓ. This procedure can be viewed as a generalization of the estimation of shifts in a web of tests (see Engelhard & Osberg, 1983).

When the bs have been obtained, the multiple values for items which have been used in more than one test are averaged. Further, starting values a'_i are obtained.

A Simulation

In order to demonstrate the procedure, a simulation study was done with five hypothetical 50-item tests, administered to five samples of 250 hypothetical examinees. The five tests were constructed to have common subtests according to the design ABCDE, AFGHI, BFJKL, CGJMN and DHKMO, where each letter stands for a different 10-item subtest. The items satisfied the three-parameter logistic model with a common value for c equal to .25. The subtests were all equal with a values of .6, .6, .8, .8, 1.0, 1.0, 1.2, 1.2, 1.4, 1.4 and b values of $-.4$, .4, $-.8$, .8, 0.0, 0.0, $-.8$, .8, $-.4$, .4. The θs of the examinee groups were sampled from normal distributions with means $-.5$, $-.2$, .1, .5, and .1 respectively, and a common standard deviation equal to 1.

Starting values for the parameters a, b, μ, and σ were obtained from p and item-total correlations corrected for attenuation and spuriousness, according to the procedure described in the previous section. Next, F from Equation 12—with seven latent classes, chosen in order to approximate the normal distribution—was minimized with a program run on a microcomputer. In this program the method of steepest descent was used for the minimization of F with respect to the as and bs, next with respect to the μs and σs, and finally with respect to c. This process was repeated until the relative decrease in F was less than 1%. This stopping rule was chosen in order to keep the time needed for the computation within reasonable limits: The function decreases very slowly after the first iterations.

The final $\hat{c}$ was equal to .216. The estimated bs correlated .97 with the true values; the estimated as correlated .74 with the true values. The latter correlation is disappointing, but Hulin, Lissak, and Drasgow (1982) had already noticed that for accurate estimation of as, relatively large samples are needed (see their Table 7, based on as with a somewhat smaller standard deviation than in the present study).

With the sample sizes in this study, individual item parameters are not accurately estimated, but many applications of this procedure do not require such accuracy. One such application is to verify whether items recently added to the item bank are relatively more or less difficult than the older items. This can be investigated by comparing the relative true score on a group of old items (i.e., the true score on these items divided by the number of items) with the relative true score on a group of recently added items, for a range of θ values. The "old" items from subtests A,B, and C and the more "recent" items in subtests M, N, and O were compared in this manner; the results are presented in Figure 1. The difference $\hat{\tau}_{ABC} - \hat{\tau}_{MNO}$ should be 0, as all subtests are equal. The obtained differences are quite small.

Discussion

A simple procedure has been suggested for the simultaneous estimation of item parameters when there are several test forms. This procedure, which can be implemented on microcomputers, seems useful under the following conditions:

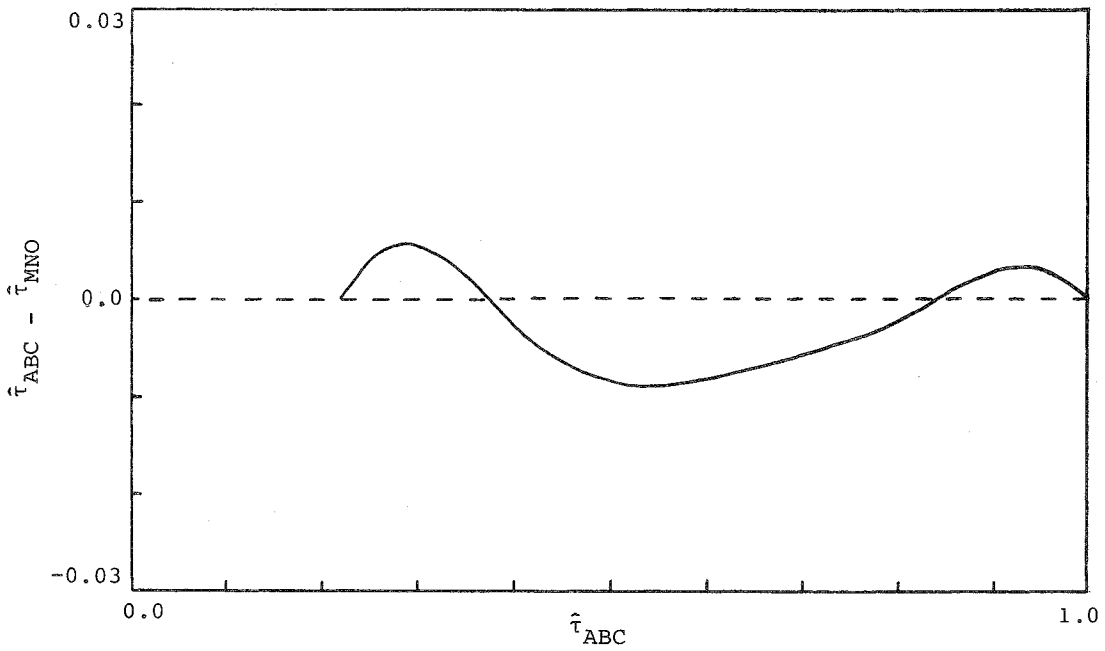
1. The researcher is interested in the relation between test forms measuring the same instructional objectives, and not in a detailed analysis of particular test forms.
2. Accurate item parameter estimates are not needed for the application in question.
3. The use of more sophisticated procedures is not indicated: A large computer is not available or too expensive in view of small sample sizes, or the available data are insufficient, which is the case when only indices such as p and r have been obtained.

The procedure seems especially suitable in connection with small scale item banking projects in educational settings.

The results obtained with this procedure are reasonable. It might be argued, however, that the procedure was tested under ideal conditions: All true c values were equal and all population distributions were normal. The equality restriction on the cs seems not too troublesome, however. The estimation of individual cs is difficult, and in programs like LOGIST, in which the individual cs can be estimated, a

Figure 1

Plot of the Estimated Difference Between Relative True Scores on Two Groups of Items



common value frequently is substituted for individual values because of low estimation accuracy. Of course, when there are different item types, and the average c is likely to differ between item types, it is possible to adapt the estimation procedure and to estimate a different c for each item type.

The normal distribution can be viewed as an approximation for distributions with low to moderate degrees of skew that are frequently found in practice. When there is evidence of strong deviations from normality, due to the existence of subgroups of varying ability, problems may arise. These can be avoided by computing ps and rs for each of the subgroups separately.

Another potential threat to estimation accuracy is speededness. The ps of the last test items are underestimates when a substantial number of examinees does not reach these items. In such a case it might be better to eliminate the last items before performing an analysis of the type suggested here.

References

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

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–32.

De Gruijter, D. N. M. (1986). Item banking with random or stratified tests. *Tijdschrift voor Onderwijsresearch, 11,* 61–66.

De Gruijter, D. N. M., & Mooijaart, A. (1983). Least squares estimation of the item parameters in the three-parameter logistic model. *Tijdschrift voor Onderwijsresearch, 8,* 218–223.

Engelhard, G., & Osberg, D. W. (1983). Constructing a test network with a Rasch measurement model. *Applied Psychological Measurement, 7,* 283–294.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

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

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics, 8,* 271–288.

Muthén, B. (1978). Contributions to factor analysis of dichotomous data. *Psychometrika, 43,* 551–560.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Reiser, M. (1983). An item response model for the estimation of demographic effects. *Journal of Educational Statistics, 8,* 165–186.

Sörbom, D. (1982). Structural equation models with structured means. In K. G. Jöreskog (Ed.), *Systems under indirect observation—Part I.* Amsterdam: North-Holland.

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

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs NJ: Prentice-Hall.

Urry, V. W. (1974). Approximation to item parameters of mental test models and their use. *Educational and Psychological Measurement, 34,* 253–269.

Urry, V. W. (1976). Ancillary estimators for the item parameters of mental test models. In *Computers and testing: Steps toward the inevitable conquest.* Washington DC: Research Section, Personnel Research and Development Center, U.S. Civil Service Commission.

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.