ED 424 238 TM 027 364

AUTHOR	Glas, Cees A. W.; Beguin, Anton A.
TITLE	Appropriateness of IRT Observed Score Equating. Research Report 96-04.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE	1996-10-00
NOTE	45p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Achievement Tests; *Cutting Scores; *Equated Scores; Foreign Countries; *Item Response Theory; *Scoring; Secondary Education; Secondary School Students; Test Format; *Test Results
IDENTIFIERS	Netherlands; Number Right Scoring; Polytomous Scoring; Rasch Model

ABSTRACT

Recently, L. Zeng and M. J. Kolen (1995) have introduced item response theory (IRT) observed score (OS) equating of number-correct (NC) scores for equating different forms of a test. In this paper, IRT-OS-NC equating is adapted to equating the cut-off scores of examinations. Next, the differences between results obtained using a Rasch model for polytomously scored items and results obtained via the nominal response model are evaluated. For both versions of IRT-OS-NC equating confidence intervals are derived. Finally, two procedures for testing the validity of the procedure are presented. Differences between the two versions were not very large. The methods studied here are exemplified with the results of equating a number of the examinations in secondary education in the Netherlands. Some limitations of the approach are discussed. (Contains one figure and seven tables.) (Author/SLD)

Appropriateness of IRT Observed Score Equating

Research Report 96-04

Cees A.W. Glas

and

Anton A. Béguin

Appropriateness of IRT Observed Score Equating

Cees A.W. Glas & Anton A. Béguin

Appropriateness of IRT Observed Score Equating, Cees A.W. Glas & Anton A. Béguin - Enschede: University of Twente, Faculty of Educational Science and Technology, December 1996. - 41 pages.

Abstract

Recently, Zeng and Kolen (1995) have introduced item response theory (IRT) observed score (OS) equating of number correct (NC) scores for equating different forms of a test. In the present paper, IRT-OS-NC equating is adapted to equating the cut-off scores of examinations. Next, the differences between results obtained using a Rasch model for polytomously scored items and results obtained via the nominal response model are evaluated. For both versions of IRT-OS-NC equating confidence intervals are derived. Finally, two procedures for testing the validity of the procedure are presented. The methods studied here are exemplified with the results of equating a number of the examinations in secondary education in the Netherlands.

The Design

Although much attention is given to producing equivalent examinations for secondary education from year to year, research has shown (see the Inspection of Secondary Education in the Netherlands, 1992) that the difficulty of examinations and the level of proficiency of the examinees can still fluctuate significantly over time. Therefore, an equating procedure was developed for setting the cut-off scores of examinations in such a way that some form of equity could be achieved. This is done with the following procedure. For all examinations participating in the procedure, the committee for the examinations in secondary education has chosen a reference examination where the quality and the difficulty of the items appeared to be such, that the cut-off score presented a suitable reference point. The cut-off scores of new examinations are to be equated to this reference point.One of the main difficulties of equating new examinations is the problem of secrecy: examinations cannot be made public until they are administered to the examinees. Another problem is that the examinations have no overlapping items. These problems are overcome by sampling linking groups form another stream of secondary education. These linking groups respond to items from the old and the new examination directly after the new examination has been administered. As an example, consider the design of Figure 1. This figure is a symbolic representation of an item administration design in form of a persons by items matrix; the shaded areas represent a combination of persons and items were data are available, the blank areas are unobserved.

Insert Figure 1 about here

It can be seen that five linking groups were used and the design is such that the linking groups cover all items of the two examinations. The proficiency level of the linking groups and the examination populations need not be equivalent; below a marginal maximum likelihood (MML) estimation procedure will be used where every group in the design has its own ability distribution. On the other hand, the responses of the linking groups must fit the same IRT model as the responses of the examination groups. For instance, if the linking groups do not seriously respond to the items administered, equating the two examinations via these linking groups would be seriously threatened. Therefore, much attention is given to the procedure for collecting the data of the linking groups, in fact, the tests are presented to these testees as school tests with consequences for their final marks. Further, a testing procedure will be proposed below that focusses on the quality of the responses of the linking groups. The examinations considered here consist of both dichotomously and polytomously scores items. Two IRT models for performing IRT-OS-NC equating will be considered: a generalization of the Rasch model to polytomously scored items known as the generalized partial credit model (GPCM, Wilson & Masters, 1993), and the nominal response model (NRM, Bock, 1972), which can be seen as a generalization of the two-parameter logistic model (2-pl, Birnbaum, 1968) to polytomously scored items. The reasons for considering these two models are several. First, the estimation procedure of the Rasch model is quick and numerically robust. Quickness is essential in the present application because the advice concerning the new cut-off score must be given as rapidly as possible. The speed of the estimation procedure originates from the existence of minimal sufficient statistics for the parameters, which makes it possible to estimate the parameters on a high aggregation level of the data (see, for instance, Glas & Verhelst, 1989). Estimation of the parameters of the NRM, on the other hand, needs evaluation of all response patterns in every iteration step of the MML estimation procedure (see, for instance, Bock & Aitkin, 1982, or Mislevy & Bock, 1990). This results in substantially longer computing times. Further, in some instances the nominal response model suffers from identification problems, which are then solved by introducing priors on the parameters (Mislevy, 1986), which further burdens the computational task. For the Rasch model, such identification problems have not been reported. On the other hand, the NRM is more flexible, so model fit should be less a problem than with the Rasch model. Given these considerations, one of the problems studied below will be the extent to which both models produce comparable results.

The IRT Models

The design sketched above is formalized by introducing item administration variables

$$d_{bi} = \begin{cases} 1 & \text{if item i is present in test } b, \\ 0 & \text{if this is not the case.} \end{cases} \tag{1}$$

for $i = 1,...,I$ and $b = 1,...,B$. Let item i have m_i+1 response categories indexed $j = 0,1,...,m_i$, $m_i > 0$. The response to the item will be represented by an (m_i+1)-dimensional vector $\boldsymbol{x}_i' = (x_{i0},...,x_{ij},...,x_{im_i})$, where x_{ij} is defined

$$x_{ij} = \begin{cases} 1 & \text{if the respionse is in category } j, j = 0,\dots,m_i \\ 0 & \text{if this is not the case.} \end{cases} \tag{2}$$

A respondent taking test b receives a score

$$r^{(b)} = \sum_{i=1}^{I} \sum_{j=1}^{m_i} d_{bi} w_{ij} x_{ij} \,, \tag{3}$$

for $r^{(b)} = 0,1,\dots,R_b$, where R_b is the maximum score that can be obtained on test b. The score weights w_{ij} are defined by the content experts developing the examinations. One of the motivations for introducing these score weights is that some of the examinations consist of multiple choice items, where only one of the alternatives is correct and open ended questions, where the response is given an integer score. Introducing score weights opens up the possibility of differentially weighting the various items in the test. Given these scoring rules, two approaches of modelling the responses are studied, the first one is an approach where the respondent's score is the minimal sufficient statistic for ability and a model where this is not the case.

. With respect to the first approach, Andersen (1977) has shown that adopting the assumption that r is a minimal sufficient statistic for a unidimensional ability parameter theta, local stochastic independence and some technical assumptions, results in a model where the probability of a response in category j, $j = 0,\dots,m_i$, of item i is given by

$$P(X_{ij} = 1 \mid \theta, \beta_i, w_i) = \psi_{ij}(\theta) = \frac{\exp(w_{ij}\theta - \beta_{ij})}{\sum_{g=0}^{m_i} \exp(w_{ig}\theta - \beta_{ig})}, \tag{4}$$

where $\beta_i' = (\beta_{i0},\dots,\beta_{ij},\dots,\beta_{im_i})$ is a vector of item parameters and

$\mathbf{w}_i' = (w_{i0},\dots,w_{ij},\dots,w_{im_i})$ is a vector of scoring weights. The item parameter of the zero response category β_{i0} is set equal to zero to identify the model. The model is also known as the generalized partial credit model (Wilson & Masters, 1993). If the weights are { 0, 1, 2, 3,..., m_i } and a re-parametrization $\eta_{ij} = \sum_{g=1}^{j} \beta_{ig}, j = 1,\dots,m_i$ is applied, it can be easily verified that (4) specializes to the well-known partial credit model (Masters, 1982); if, further, m_i is set equal to 1, the well-known Rasch model (Rasch, 1960, 1961) follows.

Notice that in the parametrization of (4), it is possible to have an item with, say $m_i = 2$, and score weights { 1, 2, 3 }, that is, the zero score cannot be obtained on this item. For practical purposes, such as not having to down-code data in case of an unobserved zero category, and for communication of results to the practitioner, this may be quite convenient and all theory to be presented below applies to the general parametrization of (4). However, it must be stressed that subtracting a weight equal w_{i0} from all category weights within the item, such that w_{i0} itself will be transformed to zero, will not alter the likelihood equations. With this alteration the denominator of (4) will equal $1 + \sum_{g=1}^{m_i} \exp(w_{ig}\theta - \beta_{ig})$, while the nominator of the probability of scoring in the zero category will equal one.

The paradigm that the scoring rule must be equivalent with the sufficient statistic for ability is abandoned by replacing these weights in (4) by unknown item parameters alpha_{ij} that must be estimated. In the framework of dichotomous items this approach results in the two-parameter logistic model (2-pl) by Birnbaum (1968). The nominal response model by Bock (1972) can be viewed as a generalization of the 2-pl to polytomous items. This model can be derived from (4) by replacing $\mathbf{w}_i$ by α_i, $\alpha_i' = (\alpha_{i0},\dots,\alpha_{ij},\dots,\alpha_{im_i})$, and setting α_{i0} equal to zero to identify the model.

As already mentioned above, a marginal maximum likelihood (MML)estimation procedure will be used where every group in the design isassumed to be sampled from a specific ability distribution, so, for instance, the data in the design depicted in Figure 1 are evaluated using seven ability distributions, that is, one distribution for the reference group, one for the examinees of the first examination, and five for the linking groups. Let the ability parameters of the respondents of test b have a normal distribution with density $g(\theta|\mu_b,\sigma_b)$. Then the probability of observing a response pattern $\mathbf{x}^{(b)}$ as a function of the item parameters of test b, say α_b and β_b and the population parameters μ_b and σ_b is given by

$$p(\mathbf{x}^{(b)}|\alpha_b,\beta_b,\mu_b,\sigma_b) = \pi_{\mathbf{x}^{(b)}} = \int p(\mathbf{x}^{(b)}|\theta,\alpha_b,\beta_b)g(\theta|\mu_b,\sigma_b)d\theta. \quad (5)$$

MML estimation boils down to maximizing the loglikelihood

$$L(\alpha,\beta,\boldsymbol{\mu},\sigma) = \sum_b \sum_{\mathbf{x}^{(b)}} n_{\mathbf{x}^{(b)}} \ln \pi_{\mathbf{x}^{(b)}}, \quad (6)$$

with respect to all item parameters α and β and all population parameters $\boldsymbol{\mu}$ and σ; the second summation runs over the set of all possible response patterns of test b and $n_{\mathbf{x}}^{(b)}$ is the number of respondents with response pattern $\mathbf{x}^{(b)}$. Of course, due to the large number of possible response patterns, these counts will usually be either equal to zero or one. The important point here is that with the present procedure all item and population parameters are simultaneously estimated on a common scale (Bock & Aitkin, 1982, Mislevy & Bock, 1990, Glas & Verhelst, 1989), so the procedure of estimating parameters for each test form separately and subsequently combining these estimates to derive a common scale (Kolen & Brennan, 1995, Chapter 6) is not necessary here.

The Equating Procedure

Once the data have been gathered and the IRT model has been estimated,the next step in the equating procedure is estimating the frequency distributions performing equipercentile equating. Consider the example of Table 1. The example concerns a reference examination and a new examination of 50 score points. The second and fourth column concern the cumulative relative frequency distributions of the reference and new examination produced by the populations actually administered these two tests. These two distributions could be either the actually observed distributions or their expected values, this will be commented upon later. In the third column an estimate of the cumulative score distribution of the reference population on the new examination is given. This estimate is computed as follows.

Insert Table 1 about here

Let b be the reference examination and let b^* be the new examination. The proportion of respondents in the reference population obtaining a score $r^{(b^*)}$ on the new examination, say $P_r^{(b^*)}$, is estimated by its expected value, that is, as the expected proportion of respondents of a population characterized by population parameters μ_b and σ_b obtaining a score $r^{(b^*)}$ on a test characterized by item parameters α_{b^*} and β_{b^*}. Using (5), this expectation is given by

$$E(P_r^{(b^*)}|\alpha_{b^*},\beta_{b^*},\mu_b,\sigma_b) = \sum_{x^{(b^*)}} \int p(x^{(b^*)}|\theta,\alpha_{b^*},\beta_{b^*})g(\theta|\mu_b,\sigma_b)d\theta. \quad (7)$$

Of course, it is also possible to calculate the expected value of the proportion of respondents of the reference population obtaining a score $r^{(b)}$ on the reference test, say $P_{r^{(b)}}$, using (7) with b^* substituted by b.

Returning to Table 1, the third columns contains the cumulative distribution of respondents of the response population on the new examination as computed by (7). The cut-off score for the new examination is set in such a way that the expected percentage of respondents failing the new examination in the reference population is approximately equal to the percentage of examinees in the reference population failing the reference examination. In the example of Table 1, the cut-off score of the reference examination was 24; as a result 21.0% failed the exam. If this percentage is held constant for the reference population, the new cut-off score should be 18. Obviously, the new examination is more difficult, which is also reflected in the mean score of the two examination displayed at the bottom of the table. The old and the new cut-off scores are marked with a straight line in the first column. It can be seen that the percentage of students in the new population failing the new examination is 15.8%. This suggests that the new population is more proficient than the reference population, also this is reflected in the difference between the mean scores of the two populations if the examination is held constant. An interesting aspect of the procedure is that the cut-off scores of the two examinations could also have been equated conditional on the new population. Further, the actual observed distributions could be replaced by their expected values. These two topics will be retumed to in the sequel.

Results of the Equating Procedure

In the examination campaign of 1995, the cut-off scores of eight examinations where equated to the cut-off scores of older examinations, the topics of the examinations are listed under the heading "Topic" of Table 2. There are seven examinations in language comprehension and one in music. The examinations are administered at two levels, topics labeled "D" in Table 2 are at MAVO-D-level, topics labeled "H" are at HAVO-level. The reference examinations were originally administered between 1989 and 1993. All examinations consist of dichotomous selected response items, except the examination for Dutch language comprehension, which has both selected and constructed response formats. The selected response items where dichotomous, but a correct response was given two score points, on the constructed response items two to six points could be obtained; the total number of score points for both the reference and the new examination was 90.

Insert Table 2 about here

The examination data consisted of samples of candidates from the complete examination populations, the sample sizes are shown in the columns 4 and 8 of Table 2. The means and standard deviations of the observed frequency distributions of the examinations are shown in the columns 5, 6, 8 and 9. For each design there were 5 linking groups, every linking group made approximately the same number of items and all items were used in the link. The total numbers of respondents in the linking groups are shown in the last column of Table 2.

Insert Table 3 about here

In Table 3, the results of the equating procedure are given for the version of the procedure where all distributions are estimated by their expected values. For each topic, four possible cut-off points are evaluated, $r^{(b)}$ = 20, 25, 30, 35 for examinations with 50 score points and $r^{(b)}$ = 45, 55, 65, 75 for the examination with 90 score points, these scores are listed in the column labeled $r^{(b)}$. As mentioned above, the associated scores on the new test could be computed using either the reference or the new population, these scores on the new test will be denoted $\phi_R(r^{(b)})$ and $\phi_N(r^{(b)})$, respectively. The results obtained via the reference population are listed in the columns 3 to 5, the results obtained via the new population are listed in columns 6 to 8. The third column contains the scores $\phi_R(r^{(b)})$ computed using the GPCM, in the next column the resulting scores are given as they are obtained using the NRM. Column 5 contains the difference between these two sets of scores. For convenience, the sum of these absolute values of these differences is given at the bottom line of the table. The following two columns give the scores $\phi_N(r^{(b)})$, that is, the scores on the new test computed via the new population, in column 8 the difference between these two scores are given. Finally, the differences in results obtained using either the reference or new population, $\phi_R(r^{(b)})-\phi_N(r^{(b)})$ are shown in column 9 for the GPCM and column 10 for the NRM, respectively. Two conclusions can be drawn from this table. First, the GPCM and the NRM do produce different results, but these differences are not spectacular: the sum of the absolute values of the differences given at the bottom of the table are 13 and 11 score points over all examinations and equated scores, and the absolute difference is never more than two score points. The second conclusion is that using either the reference or new population for determining the difference between the examination makes little difference, at the bottom of the table it is shown that the sum of the absolute values of the differences are 0 and 4 score points.

This last result depreciated when the expected distributions of the two examinations were replaced with the actual observed distributions. This can be seen in Table 4. Column 3 contains the differences between the scores $\phi_R(r^{(b)})$ as computed using the GPCM and the NRM, respectively. In column 4 the a comparable result is displayed for the scores $\phi_N(r^{(b)})$. Comparing these two columns labeled $\omega_R^1-\omega_R^2$ and $\omega_N^1-\omega_N^2$ with the columns labeled $\phi_R^1-\phi_R^2$ and $\phi_N^1-\phi_N^2$ in Table 3, it can be seen that using observed or expected scores makes little difference if the two models are contrasted. The columns 5 and 6 contain information analogous to the information in the two last columns of Table 3, so the entries are the difference between the computed scores on the new test using either the reference or new population, the differences of column 5 concern the GPCM, the next column concerns the NRM. At the bottom line it can be seen that the sum of absolute differences is clearly increased. The reason is that the expected distribution can be seen as a smoothed version of the observed distribution. In other words, the results of the first procedure are more parsimonious because it is based on four model-conform expected distributions, while the latter procedure uses more irregular observed distributions. This is further confirmed by the results of

----

Insert Table 4 about here

----

the last four columns of the table. Here the differences between the scores computed using the observed and expected distribution are listed for the GPCM and NRM applied using the reference and new population, respectively. Though the absolute difference is never greater than two score points, the occurrence of differences is such, that their absolute sums range from 9 to 22. So summing up, using expected distributions for all combinations of tests and populations resulted in a more parsimonious results, mainly due to the fact that expected distributions are smoother than the observed distributions from which they emanate. Further, the GPCM and NRM produce quite similar results.

Some Computational Considerations

Computing expected distributions defined by (7) involves summing over the set of all possible response patterns $\mathbf{x}^{(b)}$ of some test b. Dropping the indices b and b^*, for the GPCM, (7) can be written as

$$E(P_r|\beta,\mu,\sigma) = \sum_{\mathbf{x}} \exp(-\mathbf{x}'\beta)\int \exp(r\theta)P_0(\theta,\beta)g(\theta|\mu,\sigma)d\theta$$

$$= \gamma(r,\beta)\zeta(r,\beta,\mu,\sigma) \qquad (8)$$

where $P_0(\theta,\beta)$ is the probability of a zero response pattern as a function of ability, $\gamma(r,\beta)$

is a combinatorial function of all response patterns resulting in r and $\zeta(r,\beta,\mu,\sigma)$ is a function which does not depend on response patterns but only on r. In the framework of the Rasch model and its generalizations, combinatorial functions and their computation have been extensively studied (Fischer, 1974, Verhelst, Glas & van der Sluis, 1981, Verhelst & Veldhuijzen, 1991, Liou, 1994) and they can be evaluated fast and accurate. The function $\zeta(r,\beta,\mu,\sigma)$ contains an integration over a normal distribution which can be evaluated using Gauss-Hermite quadrature (Abramowitz & Stegun, 1970). Applications of Gaussian quadrature in IRT are numerous (Bock & Aitkin, 1981, Mislevy & Bock, 1990, Zeng & Kolen, 1995), but it must be pointed out that for the integrals evaluated here the number of quadrature points must be large to obtain acceptable numerical precision (Verhelst & Verstralen, personal communication). In the examples of this paper, the number of quadrature points was set equal to 180.

For the NRM, expression (7) can be written as

$$\sum_{\boldsymbol{x}} \int \exp(\boldsymbol{x}'(\alpha\theta-\beta)P_0(\theta,\alpha,\beta)g(\theta|\mu,\sigma)d\theta =$$

$$\int \sum_{\boldsymbol{x}} \exp(-\boldsymbol{x}'\delta(\theta))P_0(\theta,\alpha,\beta)g(\theta|\mu,\sigma)d\theta, \tag{9}$$

where $P_0(\theta,\alpha,\beta)$ is the probability of a zero response pattern as a function of ability and $\delta(\theta) = (\alpha\theta-\beta)$. An important difference between (8) and (9) is that in the former expression a factor depending on response patterns can be placed before the integration sign, while this is not possible in (9).

One way to compute (9) is to introduce combinatorial functions $\gamma(r,\delta(\theta)) = \sum_{\boldsymbol{x}} \exp(-\boldsymbol{x}'\delta(\theta))$ which are defined conditionally on θ, so that (8) generalizes to

$$E(P_r|\alpha,\beta,\mu,\sigma) = \int \gamma(r,\delta(\theta))P_0(\theta,\alpha,\beta)g(\theta|\mu,\sigma)d\theta.$$

Computing (10) boils down to evaluating the combinatorial functions in every quadrature point. However, as was mentioned above, the number of quadrature points needed is quite large, so this approach is quite time consuming. As an alternative, (10) can be evaluated using a Monte Carlo procedure, where response patterns are generated using the relevant item and population parameters to approximate the distribution of sum scores on a test for a certain population. Also this approach requires a substantial amount of computer time. For the examples in the present paper both methods are used; details on the relative merits of the two procedures are beyond the scope of the present paper.

Confidence Intervals

When the practitioner is confronted with the need to adjust the cut-off score of some examination, the first question that comes to mind is about the reliability of the estimated new cut-off score. In this section, two methods for computing confidence intervals for all relevant estimates will be considered: the delta method and the bootstrap method. The delta method (see, for instance, Bishop, Fienberg & Holland, 1975) will be described first. This method is based on the fact that if $\lambda-\hat{\lambda}$ has an asymptotic normal distribution with mean **0** and covariance matrix Σ_λ, and f is a differentiable real-valued function, then $f(\lambda)-f(\hat{\lambda})$ has an asymptotic normal distribution with mean **0** and covariance matrix

$$\Sigma_f = (\partial f/\partial \lambda)\Sigma_\lambda(\partial f/\partial \lambda)'. \tag{11}$$

In the present case, all inferences, such as the expected cumulative score distributions and the mean and variance of the expected score distributions, are based on (7), which, in turn, is a function of estimated item- and population parameters. Therefore, first the standard errors of (7) will be derived. Let λ be a vector of all item and population parameters and $f(\lambda)$ will be a vector of one or more expected score distributions. So, in general $f(\lambda)$ will have elements $E(P_r|\alpha,\beta,\mu,\sigma)$. Consider the GPCM. To derive an expression for the derivative of (8) with respect to an item parameter, notice that

$$\frac{\partial \gamma(r,\beta)}{\partial \beta_{ij}} = -\exp(-\beta_{ij})\gamma(r-j,\beta^{(i)}), \tag{12}$$

where $\gamma(r-j,\beta^{(i)})$ is a combinatorial function over all possible response patterns on the test without item i resulting in score $r-j$, so this is a function of all item parameters minus the parameters of item i (see, for instance, Fischer, 1974, Liou, 1994, Verhelst & Glas, 1995). Further,

$$\frac{\partial P_0(\theta,\beta)}{\partial \beta_{ij}} = \psi_{ij}(\theta)P_0(\theta,\beta), \tag{13}$$

and so

$$\frac{\partial E(P_r|\beta,\mu,\sigma)}{\partial \beta_{ij}} =$$

$$-\exp(-\beta_{ij})\gamma(r-j,\beta^{(i)})\zeta(r,\beta,\mu,\sigma) + \gamma(r,\beta)\int \psi_{ij}(\theta)\exp(r\theta)P_0(\theta,\beta)g(\theta|\mu,\sigma)d\theta \tag{14}$$

$$= -E(P_{rij}|\beta,\mu,\sigma) + E(P_r|\beta,\mu,\sigma)E(\psi_{ij}(\theta)|r,\beta,\mu,\sigma)$$

here $E(P_{rij}|\beta,\mu,\sigma)$ is the expected proportion of respondents scoring in category j of item i and obtaining a sum score r. The derivatives of (8) with respect to the population parameters are given by

$$\frac{\partial E(P_r|\beta,\mu,\sigma)}{\partial\mu} = \gamma(r,\beta)\int\left(\frac{\theta-\mu}{\sigma^2}\right)\exp(r\theta)P_0(\theta,\beta)g(\theta\,|\,\mu,\sigma)d\theta \tag{15}$$

and

$$\frac{\partial E(P_r|\beta,\mu,\sigma)}{\partial\sigma} = \gamma(r,\beta)\int\left(\frac{(\theta-\mu)^2-\sigma^2}{\sigma^3}\right)\exp(r\theta)P_0(\theta,\beta)g(\theta\,|\,\mu,\sigma)d\theta. \tag{16}$$

The covariance matrix of the score distribution can now be computed using (14), (15) and (16) as expressions for $\partial f/\partial\lambda$; the expression for the covariance matrix of the parameter estimates Σ_λ for the GPCM are given by Glas (1997, also see Glas & Verhelst, 1989).

The covariance matrix for the cumulative score distribution, say Σ_c, can now be derived from the covariance matrix for the score distribution Σ_f by noticing that the latter is a linear function F of the former, and Σ_c is derived by pre-multiplying Σ_f by F and post-multiplying it by F'. For instance, the covariance matrix of two cumulative distributions of two tests with 2 score points each is given by

$$\Sigma_c = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \Sigma_f \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{17}$$

Also confidence intervals for the estimates of the mean and the variance of the score distributions can be computed in this way, for instance, the estimate for the mean is based on the linear combination

$$\sum_r rE(P_r|\beta,\mu,\sigma), \tag{18}$$

and its standard error can be computed by pre-multiplying Σ_f by the row vector (0,1,...,r,...,R) and post-multiplying it by the transpose of this row vector. The expected second central moment and the variance of the score distribution can be computed in a similar vain. The derivation for the NRM is a straightforward generalization of the procedure for the GPCM. So the equivalent of (12) is now given by

$$\frac{\partial\gamma(r,\delta(\theta))}{\partial\alpha_{ij}} = \theta\exp(\alpha_{ij}\theta-\beta_{ij})\gamma(r-j,\delta(\theta)^{(i)}) \tag{19}$$

and

$$\frac{\partial\gamma(r,\delta(\theta))}{\partial\beta_{ij}} = -\exp(\alpha_{ij}\theta-\beta_{ij})\gamma(r-j,\delta(\theta)^{(i)}) \tag{20}$$

and the equivalent of (13) is

$$\frac{\partial P_0(\theta,\alpha,\beta)}{\partial\alpha_{ij}} = \theta\psi_{ij}(\theta)P_0(\theta,\alpha,\beta) \tag{21}$$

and

$$\frac{\partial P_0(\theta,\alpha,\beta)}{\partial\beta_{ij}} = \psi_{ij}(\theta)P_0(\theta,\alpha,\beta). \tag{22}$$

These expressions can be used for deriving the first order derivatives of (10) with respect to the item parameters. The first order derivatives of (10) with respect to the population parameters resemble (15) and (16), except that the combinatorial function is defined locally on ability as in (10) and should be placed after the integration sign. Again, the delta method can be used for computing confidence intervals for one or more expected score distributions by combining these expressions for the first order derivatives with the expressions for the asymptotic covariance matrix derived by Glas (1997).

As an alternative for the delta method, the bootstrap method (Efron, 1979, Efron & Gong, 1983) will be considered. The bootstrapping method proceeds by repeated re-sampling with replacement from the original data. The sample size of these re-samples is the same as the size of the original sample and the probability of being sampled is the same for all response patterns in the original sample. By estimating the model parameters on every re-sample the standard error of the estimator can be evaluated. For the present application standard errors for the estimated frequency distributions under the GPCM and the NRM were computed using both the bootstrap and the delta method. To avoid cumbersome tables, only the results of a subset from an actual data set will be used, the data consist of 10 items from the English language proficiency examination on Havo-level in 1992 and 10 items from the 1995 examination. Score distributions were computed on these two examinations for the 1995 population. Because only one linking group made the items studied here, the design was curtailed to the two examination populations with 2039 and 2003 candidates, respectively, and one linking group consisting of 175 candidates. In Table 5 an example of one of the estimated score distributions is shown, the example concerns an estimate of the distribution of the 1995 population on the 1992 test using the GPCM.

Insert Table 5 about here

The columns two and three contain the estimated score distribution and the cumulative distribution, the next two columns contain their standard errors estimated applying the delta method, respectively. Next, the bootstrapped estimates of these four estimates are given. Finally, in the two bottom lines of the table the mean, the standard deviation and their respective standard errors are given. The bootstrapped estimates were computed using 400 replications. It can be seen that the bootstrapped estimates of the standard errors are generally smaller than the ones computed using the delta method. This result is typically for all analyses that were carried out. Because the number of parameters estimated in the NRM is larger than the number of parameters estimated in the GPCM, the standard errors in the NRM are slightly smaller: for instance, the standard error of the mean computed using the delta method dropped from .15 to .12. Other estimates showed a comparable tendency. For both models and both estimation procedures, the computed standard errors dropped dramatically when the score distribution was estimated on the test the candidates actually made. For instance the standard error of the mean using the delta method was computed as .05, so markedly smaller than the standard error for the mean of the test not actually made by the candidates. This also held for the estimates of the score distribution, for instance the standard error of the estimate of the proportion of candidates with score 5 dropped from 1.03 to .25. Of course, this is as expected, since the data provide more information on the test made than on the test that was not made.

The final remark of this section concerns the practical implications of these results. Firstly, the estimates issued from the delta method are generally more conservative, so they must be preferred over the bootstrapped estimates. For the GPCM computing bootstrapped estimates offers little problems because the estimation procedure is both fast and robust. For the NRM this is less the case, in fact, repeated parameter estimation may be quite prohibitive for very large tests. However, for the NRM also the delta method seems to be running into trouble every once in a while, but in these cases replacing the observed information matrix by the expected information matrix usually solves the problem. Summing up, the delta method must be preferred..

Evaluating Model Fit

In this last section a procedure for evaluating model fit in the framework of IRT-OS-NC equating will be discussed. Of course, there are many possible sources of model violations, and many test statistics have been proposed for evaluating model fit, which are quite relevant in the present context (see, Andersen, 1973, Martin Löf, 1973, Glas, 1988, 1997, Glas & Verhelst, 1989, 1995, Molenaar, 1983, and Mislevy & Bock, 1990). Besides the model violations covered by these statistics, in the present application there is one special violation that deserves special attention: the question whether the data from the linking groups are suited for performing the equating of the examinations. Therefore, the focus of the present section will be on the stability of the estimated score distributions if different linking groups are used. The idea is to cross-validate the procedure using independent replications sampled from the original data. This is accomplished by partitioning the data of both examinations into G data sets. To every one of these data sets, the data of one or more linking groups are added, but the data sets will have no linking groups in common. So summing up, each data set consists of a sample from the data of both the examinations and of one ore more linking groups. In this way, the equating procedure can be carried out in G independent samples. The stability of the procedure will be evaluated in two ways: firstly by computing equivalent scores as was done above and evaluating whether the two equating functions produce similar results, and, secondly, by performing a Wald test. The Wald test will be explained first.

Glas and Verhelst (1995) have pointed out that in the framework of IRT, the Wald test (Wald, 1943) can be used for testing whether some IRT model holds in meaningful subgroups of the sample of respondents. In this section, the Wald test will be used to evaluate the null hypothesis that the expected score distributions on which the equating procedure is based are constant over subgroups against the alternative that they are not. This principle applies to G sub-groups, but only the case of two subgroups will be elaborated here, the generalization to more subgroups is straightforward. Let the model parameters for the g-th subgroup be denoted λ_g, $g = 1,2$. These parameters are estimated in the two subgroups separately. Above a vector $f(\lambda)$ with elements $E(P_r|\alpha,\beta,\mu,\sigma)$ for one or more score distributions was defined. Here this definition will be altered in the sense that for every distribution at least one proportion P_r will be deleted. In the sequel it will become clear that this has to do with the restriction that the proportions P_r sum to one, i.e. $\sum_r P_r = 1$, which results in covariance matrices of incomplete rank.
In the examples below, more scores will deleted because their expected proportions are either zero or very small, for data emanating from examinations this especially happens in the low score regions. Let $\boldsymbol{f_g}$ $(\lambda_{\boldsymbol{g}})$ be one or more

distributions computed via group G. Further, let $\lambda = (\lambda_1', \lambda_2')'$ and consider the difference

$$h(\lambda) = f_1(\lambda_1) - f_2(\lambda_2), \tag{23}$$

that is, $h(\lambda)$ is the difference between one or more score distributions computed using independent samples of examination candidates and different and independent linking groups. Under the null hypothesis $h(\lambda) = \mathbf{0}$, that is, in the population the score distributions are equal. Since the responses of the two subgroups are independent, it follows that the variance-covariance matrix of the ML estimator of $(f_1(\lambda_1)', f_2(\lambda_2)')$ is given by

$$\Sigma_{f_1,f_2} = \begin{pmatrix} \Sigma_{f_1} & 0 \\ 0 & \Sigma_{f_2} \end{pmatrix}, \tag{24}$$

where the matrices Σ_{f_g}, $g = 1,2$ are computed using (11). For this application, the Wald test statistic is given by the quadratic form

$$W = h(\lambda)'[\Sigma_{f_1} + \Sigma_{f_2}]^{-1} h(\lambda);$$

if W is evaluated using ML-estimates, under mild regularity assumptions, it is asymptotically chi-square distributed with degrees of freedom equal to the number of elements of $h(\lambda)$ (Wald, 1943).

Insert Table 6 about here

Some results of the test are given in Table 6. The tests pertain to estimated score distributions on the reference examination. To test the stability of the score distribution, the samples of respondents of the examinations were divided into four subgroups of approximately equal sample size. Next, four data sets were assembled, each one consisting of the data of one linking group, the data of one of the four subgroups from the reference examination and the data of one of the four subgroups from the new examination. So the design for these four new data sets is similar to the design depicted in Figure 1, except that in the prevailing case only one linking group is present. In this way four data sets were constructed, for each data set the item- and population parameters of the GPCM were estimated, all relevant distributions were estimated by computing their expected values and the equating procedure was conducted. Finally, four Wald statistics were computed. Consider Table 6. The first column concerns the hypothesis that there is no difference between the estimated distributions of the reference population on the reference examination in the setup where the first linking group provided the link and the setup where this link was forged by the second linking group. The next column pertains to a similar hypothesis concerning the third and fourth linking group. The last two columns contain the result for a similar hypothesis concerning the estimated distributions of the new population on the reference examination. For all six examination topics, the score distribution considered ranged from 21 to 40, that is, 20 of the 50 possible score points were considered. This results in four Wald statistics with 20 degrees of freedom each, realizations with a significance probability less than 0.01 are marked with a double asterisk. It can be seen that model fit is not overwhelmingly good: 12 out of 24 tests are significant at the 0.01 level. However, there seem to be differences between the various topics, for instance, French at HAVO-level seems to fit quite well. This was corroborated further by a procedure were equivalent scores were computed for a partition of the data into five different sub-samples, each one with its own linking group. Consider Table 7. For six topics four scores on the reference test were considered. For each of the five sub-samples, these four scores were equated to scores on the new examination via the reference population.

Insert Table 7 about here

In the columns labeled "L1" to "L5", the resulting scores on the new test are shown. These new scores seem to fluctuate quite a bit, but it must be kept in mind that every one of these scores was computed using only a fifth of the original sample size, so the precision has suffered considerably. In the column labeled "Total", the sum of the absolute differences between all pairs of new scores is displayed. Since there are five new scores for every original score, there are ten such pairs. So, for instance, the mean absolute difference between the new scores associated with the original score 20 on the D-level examination in German is 4.8 score points. An interesting question in this context is how this result must be interpreted given the small sample sizes in the sub-groups. To shed some light on this question, the following procedure was followed. For every examination, new data sets were generated using the parameter estimates obtained on the original complete data sets, that is, the data sets described in Table 2. So these new generated data sets conformed the null-hypothesis of the GPCM. Next, for every data set, the procedure of equating the two examinations via the reference population in the five sub-samples was conducted. For every examination this procedure was replicated 100 times. In this manner, the distribution of the sum of the absolute differences of new scores under the null-hypothesis that the GPCM (with true parameters as estimated) holds, could be approximated and the approximated significance probability of the realization using the real data could be determined. The mean sum of absolute differences over the 100 replications and the significance probability of the real data realization are given in the last two columns of Table 7. It can be seen that the overall model fit is not very good, however, also here French at HAVO-level stands out as well fitting, while also German at HAVO-level shows acceptable model fit.

Conclusions

In the present paper, the technique of IRT-OS-NC equating introduced by Zeng and Kolen (1995) was adapted to a situation were both differences in proficiency level of various populations of respondents and differences between the difficulty of measurement instruments are meaningful and important variables that have to be accounted for. Further, methods for computing standard errors and evaluating the appropriateness of the equating method were suggested. The feasibility of the procedure in a practical situation was shown using an application in a real examination situation. In the present application, the differences between the results obtained by the GPCM and the NRM were not very striking. However, the present study did not include systematic simulations of other conceivable testing arrangements, so there is no evidence that this result also holds for other applications. Overall model fit was not very satisfactory, only one of the examination topics fitted well, while a second topic fitted acceptably. Therefore, further research must be done on adapting IRT-OS-NC equating to multi-dimensional IRT models, such as the multi-dimensional Rasch model by Glas (1992) and by Adams and Wilson (1995) and the Testfact model by Bock, Gibbons and Muraki (1985). Finally, it must be stressed that equity of testing is only relative in case that the scoring rule of the test is different from the sufficient statistic for ability or from some other IRT-based measure of ability, both derived from the IRT model that fits the data. Generally, scoring a test using IRT-based statistics or measures is to be preferred above adopting a scoring rule and then using IRT-OS-NC equating for rendering the scores comparable. However, the scoring rule is often beyond the control of the psychometrician, and in these cases IRT-OS-NC equating serves an important purpose.

References

Abramowitz, M. & Stegun, I.A. (1974). *Handbook of Mathematical Functions*. New York: Dover Publications.

Adams, R.J., & Wilson, M.R. (1995). Formulating the Rasch model as a mixed coefficients multinomial logit model. In G.Engelhard and M. Wilson, (Eds.), *Objective measurement: theory into practice, Vol. 3,* Nordwood, NJ: Ablex Publishing Corporation.

Andersen, E.B. (1973). A goodness of for test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Addison-Wesley: Reading (Mass.).

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* MIT press: Cambridge (Mass.).

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

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

Bock, R.D., Gibbons, R.D. & Muraki, E. (1985). Full-information factor analysis. *Applied Psychological Measurement 12,* 261-280.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7,* 1-26.

Efron B. & Gong, G. (1982). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37,* 36-49.

Fischer, G.H. (1974). *Einführung in die Theorie Psychologischer Tests* [Introduction to the theory of psychological tests]. Bern: Huber.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C.A.W. (1997) Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, to appear.*

Glas, C.A.W., & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Inspection of Secondary Education in the Netherlands] (1992). *Examens op Punten Getoetst [Evaluation of Examinations],* 's Gravenhage: Inspectie van het Voortgezet Onderwijs.

Liou, M. (1995). More on the computation of higher order derivatives of the elementary symmetric functions in the Rasch model. *Applied Psychological Measurement, 18,* 53-62.

Kolen, M.J. & Brennan, R.L. (1995). *Test Equating.* New York: Springer.

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

Martin Löf, P. (1973). *Statistika Modeller. Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973.* Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R.J. & Bock, R.D. (1990). *PC-BILOG. Item analysis and test scoring with binary logistic models.* Scientific Software: Mooresville.

Molenaar, I.W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika, 48,* 49-72.

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

Rasch, G. (1961). *On the general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 321-333. Berkeley: University of California Press.

Verhelst, N.D., & Glas, C.A.W. (1995). The generalized one parameter model: OPLM. In: G.H.Fischer & I.W.Molenaar (eds.). *Rasch models: their foundations, recent developments and applications.* New York: Springer.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1993). *OPLM: computer program and manual.* Arnhem: Cito.

Verhelst, N.D., Glas, C.A.W. & van der Sluis, A. (1984). Estimation problems in the Rasch model: the basic symmetric functions. *Computational Statistics Quarterly, 1,* 245-262.

Verhelst, N.D. & Veldhuijzen, N.H. (1991). *A new algorithm for computing elementary symmetric functions and their first and second derivatives.* Measurement and Research Department reports, 91-1. Arnhem: Cito.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426-482.

Wilson, M. & Masters, G.N.(1993). The partial credit model and null categories. *Psychometrika, 58,* 85-99.

Zeng, L. & Kolen, M.J. (1995). An alternative approach for IRT observed-score equating of number correct scores. *Applied Psychological Measurement, 19,* 231-240.

Table 1. Cumulative Percentages of the Reference and New Population on the Reference and New Examination

Population	Reference		New	
Examination	Ref.	New	New	Ref.
Score	Cum. Perc.	Cum. Perc.	Cum. Perc.	Cum. Perc.
16	2.4	13.5	7.3	.3
17	3.9	14.7	10.3	.6
18	4.8	19.8	15.8	1.5
19	7.5	22.5	19.1	2.1
20	9.9	24.3	27.3	4.5
21	12.3	29.3	34.5	8.2
22	14.7	31.4	39.1	10.6
23	17.7	38.0	44.5	14.2
24	21.0	42.2	50.9	16.9
25	23.7	48.5	56.1	23.2
26	28.7	54.2	63.3	27.2
Mean	28.8	24.6	25.6	29.6
Std.	9.1	9.3	8.9	8.6

Table 2. Data Overview

Topic	Score Points	Reference R	Reference N	Reference Mean	Reference Std	New Examination R	New Examination N	New Examination Mean	New Examination Std	Link N
German D	50	50	2115	31.72	6.92	50	2021	34.00	6.28	1033
German H	50	50	2129	34.51	5.59	50	2015	32.08	6.27	607
English D	50	50	1693	35.14	6.91	50	2010	34.74	6.87	1137
English H	50	50	2039	32.32	7.45	50	2003	34.45	7.23	873
French D	50	50	1666	33.18	7.39	50	2097	32.28	7.23	1037
French H	50	50	2144	35.72	6.80	50	2138	34.02	7.21	428
Dutch D	90	39	1572	56.17	12.05	44	2266	59.01	9.82	701
Music D	50	50	335	30.25	6.43	50	370	34.54	6.38	387

Table 3. Results of the Equation Procedure

Topic	$r^{(b)}$	ϕ_R^1	ϕ_R^2	$\phi_R^1-\phi_F^2$	ϕ_N^1	ϕ_N^2	$\phi_N^1-\phi_N^2$	$\phi_R^1-\phi_N^1$	$\phi_R^2-\phi_N^2$
German D	20	24	25	-1	24	24	0	0	1
	25	29	30	-1	29	29	0	0	1
	30	34	34	0	34	34	0	0	0
	35	38	38	0	38	38	0	0	0
German H	20	18	19	-1	18	19	-1	0	0
	25	24	24	0	24	24	0	0	0
	30	29	29	0	29	29	0	0	0
	35	34	34	0	34	34	0	0	0
English D	20	19	21	-2	19	21	-2	0	0
	25	24	26	-2	24	26	-2	0	0
	30	30	30	0	30	30	0	0	0
	35	35	35	0	35	35	0	0	0
English H	20	21	21	0	21	21	0	0	0
	25	26	26	0	26	26	0	0	0
	30	31	31	0	31	31	0	0	0
	35	36	36	0	36	36	0	0	0
French D	20	21	22	-1	21	22	-1	0	0
	25	26	26	0	26	26	0	0	0
	30	31	31	0	31	31	0	0	0
	35	36	37	-1	36	36	0	0	1
French H	20	19	19	0	19	19	0	0	0
	25	24	24	0	24	24	0	0	0
	30	28	29	-1	28	29	-1	0	0
	35	34	34	0	34	34	0	0	0
Dutch D	45	47	47	0	47	47	0	0	0
	55	56	56	0	56	55	1	0	1
	65	65	64	1	65	64	1	0	0
	75	74	73	1	74	73	1	0	0
Music D	20	23	23	0	23	23	0	0	0
	25	28	28	0	28	28	0	0	0
	30	33	33	0	33	33	0	0	0
	35	38	37	1	38	37	1	0	0
Abs. sum				13			11	0	4

Table 4. Differences between Equation Functions

Topic	$r^{(b)}$	$\omega_R^1-\omega_R^2$	$\omega_N^1-\omega_N^2$	$\omega_R^1-\omega_N^1$	$\omega_R^2-\omega_N^2$	$\omega_R^1-\phi_R^1$	$\omega_R^2-\phi_R^2$	$\omega_N^1-\phi_N^1$	$\omega_N^2-\phi_N^2$
German D	20	0	0	-1	-1	0	-1	1	1
	25	-1	0	0	1	0	0	0	0
	30	0	0	1	1	0	0	-1	-1
	35	-1	0	0	1	0	1	0	0
German H	20	1	-1	-1	-3	0	-2	1	1
	25	0	0	-1	-1	-1	-1	0	0
	30	0	0	0	0	0	0	0	0
	35	-1	0	0	1	0	1	0	0
English D	20	0	-2	0	-2	0	-2	0	0
	25	-1	-2	0	-1	0	-1	0	0
	30	-1	0	0	1	0	1	0	0
	35	0	0	0	0	0	0	0	0
English H	20	1	0	-1	-2	-1	-2	0	0
	25	0	0	0	0	0	0	0	0
	30	0	1	1	2	1	1	0	-1
	35	0	0	0	0	0	0	0	0
French D	20	0	-1	-1	-2	0	-1	1	1
	25	0	0	0	0	0	0	0	0
	30	-1	1	0	2	0	1	0	-1
	35	0	0	1	1	1	0	0	0
French H	20	0	0	-2	-2	-1	-1	1	1
	25	0	0	-1	-1	-1	-1	0	0
	30	0	0	1	1	1	0	0	-1
	35	0	0	1	1	0	0	-1	-1
Dutch D	45	0	0	0	0	0	0	0	0
	55	-1	1	-1	1	-1	0	0	0
	65	1	1	0	0	0	0	0	0
	75	2	1	1	0	1	0	0	0
Music D	20	1	0	-2	-3	-1	-2	1	1
	25	0	0	1	1	1	1	0	0
	30	0	0	2	2	1	1	-1	-1
	35	0	0	1	1	0	1	-1	0
Abs. sum		13	11	20	35	12	22	9	10

Table 5. Confidence Intervals using the Delta Method and the Bootstrap Method

r	DELTA METHOD E(P)	CUM	SE(E)	SE(CUM)	BOOTSTRAP METHOD, 400 REPLICATIONS E(P)	CUM	SE(E)	SE(CUM)
0	.01	.01	.00	.00	.01	.01	.00	.00
1	.11	.12	.04	.04	.11	.12	.03	.04
2	.53	.65	.14	.18	.56	.68	.13	.16
3	1.83	2.48	.39	.57	1.88	2.56	.33	.49
4	4.72	7.20	.74	1.30	4.79	7.35	.61	1.09
5	9.50	16.70	1.03	2.32	9.58	16.93	.83	1.90
6	15.49	32.19	.97	3.25	15.52	32.45	.77	2.62
7	20.63	52.82	.46	3.52	20.58	53.03	.40	2.81
8	21.98	74.80	.85	2.74	21.88	74.91	.71	2.18
9	17.33	92.13	1.56	1.19	17.24	92.15	1.24	.95
10	7.87	100.00	1.19	.00	7.85	100.00	.95	.00
	MEAN 7.21		STD 1.73		MEAN 7.20		STD 1.74	
	SE(MEAN) .15		SE(STD) .04		SE(MEAN) .18		SE(STD) .04	

BEST COPY AVAILABLE

Table 6. Results of the Wald Test for Stability of Estimated Score Distributions

population Linking Groups Topic	reference 1 vs 2	 3 vs 4	new 1 vs 2	 3 vs 4
German D	97.9**	12.0	202.3**	180.0**
German H	156.5**	16.8	8.1	232.7**
English D	24.6	8.9	460.1**	19.5
English H	52.9**	8.1	239.8**	4.1
French D	120.3**	100.4**	547.6**	158.2**
French H	4.5	15.6	21.7	10.8

Table 7. Stability of Equating Functions in Sub-samples

Topic	r(b)	L1	L2	L3	L4	L5	Total	Expct	p-value
German D	20	16	23	21	15	14	48	15.5	.00
	25	20	28	27	21	19	50	14.5	.00
	30	26	32	32	27	24	44	13.1	.00
	35	31	37	37	33	29	44	11.4	.00
German H	20	16	19	17	21	17	24	15.2	.10
	25	22	24	22	26	22	20	12.4	.15
	30	27	29	27	31	28	20	10.3	.05
	35	33	34	32	36	33	18	9.5	.10
English D	20	20	26	18	19	20	34	14.1	.00
	25	24	31	23	24	25	34	12.5	.00
	30	29	35	28	29	30	30	10.3	.00
	35	34	39	33	34	34	24	8.8	.00
English H	20	21	26	19	18	23	40	12.8	.00
	25	26	31	24	23	28	40	12.0	.00
	30	31	36	29	28	32	38	10.0	.00
	35	36	40	34	33	37	34	9.2	.00
French D	20	18	13	19	16	23	46	13.2	.00
	25	24	18	24	20	27	44	13.7	.00
	30	29	22	29	25	32	48	13.4	.00
	35	35	28	34	29	36	44	12.7	.00
French H	20	21	20	18	18	19	16	16.0	.55
	25	26	25	23	24	24	14	15.4	.75
	30	31	30	29	29	29	10	12.8	.85
	35	36	35	34	34	34	10	10.7	.70

Figure Captions

Figure 1. Test Administration Design.

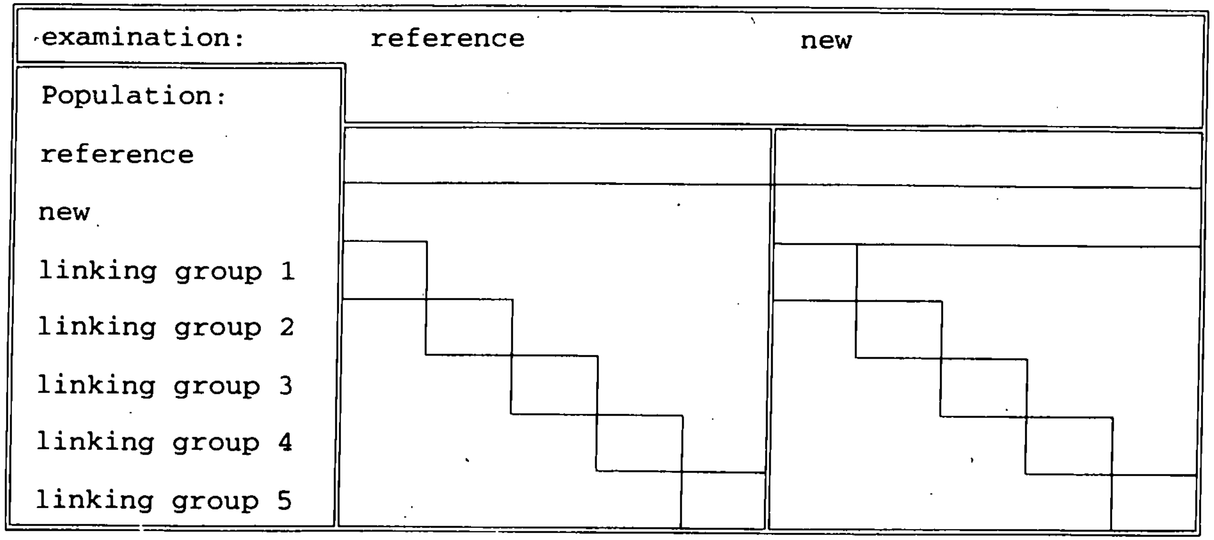

Figure 1 Test Administration Design

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der.Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.