

Chapter 18

Robustness of IRT Observed-Score Equating

C.A.W. Glas and Anton A. Béguin

18.1 Introduction

One of the most important aspects of equating, scaling, and linking is whether the models used are appropriate. In this chapter, some heuristic methods and a more formal model test for the evaluation of the robustness of the procedures used for equating, scaling, and linking are presented. The methods are outlined in the framework of item response theory (IRT) observed score (OS) equating of number-correct (NC) scores (Kolen & Brennan, 1995; Zeng & Kolen, 1995). The methods for the evaluation of IRT-OS-NC equating will be demonstrated using concurrent estimation of the parameters of the one-parameter logistic (1PL) model and the three-parameter logistic (3PL) model. In this procedure the parameters are estimated on a common scale by using all available data simultaneously. The data used are from the national school-leaving examinations at the end of secondary education in the Netherlands. To put the presentation into perspective, the application will be presented first.

18.2 Equating of School-Leaving Examinations

Although much attention is given to producing equivalent school-leaving examinations for secondary education from year to year, research has shown (see the Inspectorate 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, the following equating procedure was developed for setting the cut-off scores of the examinations. For all examinations participating

C.A.W. Glas (✉) and A.A. Béguin
University of Twente, P.O. Box 217, 7500, AE Enschede, Netherlands
e-mail: C.A.W.Glas@gw.utwente.nl

in the procedure, the committee for the examinations in secondary education chose a reference examination where the quality and the difficulty of the items were such that the cut-off score presented a suitable reference point. The cut-off scores of new examinations are 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. Another problem is that the examinations have no overlapping items. These problems are overcome by collecting additional data to create a link between the data from the two examinations. These additional data are collected in so-called *linking groups* which are sampled from 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.

Figure 18.1 displays the data collection design for equating the 1992 English language comprehension examinations at the higher general secondary education (HAVO) level to the analogous 1998 examination. The 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 where data are available, and the blank areas are unobserved.

Both examinations consisted of 50 dichotomously scored items. So the total number of items in the design was 100. Further, it can be seen that the design contains five linking groups and the design is such that the linking groups cover all items of the two examinations. Every linking group responded to a test with a test length between 18 and 22 items, and every item in the design was presented to exactly one of the linking groups.

Based on the data of the two examinations, one could directly apply equipercentile equating using the two OS distributions of the two examinations. This, however, would be based on the assumption that either the ability level of the two examinations populations (i.e., the 1992 and the 1998 population) or the difficulty level of the two examinations (i.e., the 1992 examination and the 1998 examination) had not changed. In practice, this assumption may not be tenable. The purpose of the linking groups is to collect additional information that makes it possible to estimate differences in ability levels and differences in difficulty levels using

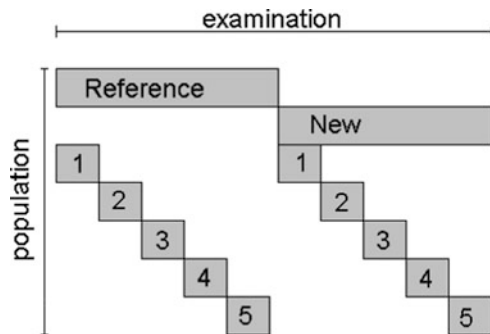


Fig. 18.1 Item administration design for equating examinations. The area labeled “Reference” pertains to the 1992 examination; the area labeled “New” pertains to the 1998 examination

concurrent marginal maximum likelihood (MML) estimation of an IRT model. This procedure will be outlined in detail below.

In concurrent MML estimation, the item and population parameters are estimated concurrently on a common scale. So, contrary to procedures where separate sets of parameters are estimated for different groups of respondents that are subsequently linked to a common scale, in concurrent estimation the model has only one set of item parameter estimates to describe all response data. That is, it is assumed that each item has the same difficulty parameters in each of the groups in which the item was administered (see, for instance, von Davier & von Davier, 2004). Although simulation studies (Hanson & Béguin, 2002; Kim & Cohen, 2002) have shown that concurrent calibration leads to better parameter recovery, Kolen and Brennan (1995) argued that separate calibration is preferred because it gives a check on the unidimensionality of the model. In the present chapter, examples of testing the model using concurrent MML estimates are given.

Using IRT creates some freedom in designing the data collection. For instance, the proficiency level of the linking groups and the examination populations need not be exactly the same; in the MML estimation procedure outlined below, every group in the design has its own ability distribution. On the other hand, the assumption underlying the procedure is that the responses of the linking groups 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 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 students as school tests with consequences for their final appraisal. One of the procedures for testing model fit proposed below will focus on the quality of the responses of the linking groups.

18.3 IRT Models

In this chapter, only examinations with dichotomously scored items are discussed. Consider an equating design with I items and N persons. Let item administration variable d_{in} be defined as

$$d_{ni} = \begin{cases} 1 & \text{if item } i, \text{ is presented to person } n, \\ 0 & \text{otherwise,} \end{cases} \quad (18.1)$$

for $i = 1, \dots, I$ and $n = 1, \dots, N$, and let the response of person n to the item i be represented by an stochastic variable X_{ni} , with realization x_{ni} . If $d_{ni} = 1$, x_{ni} is defined by

$$x_{ni} = \begin{cases} 1 & \text{if the response of person } n \text{ to item } i \text{ is correct,} \\ 0 & \text{otherwise} \end{cases} \quad (18.2)$$

If $d_{ni} = 0$, then x_{ni} is equal to an arbitrary constant

Two approaches to modeling the responses are used: the 1PL model (Rasch, 1960) and the 3PL model (Birnbaum, 1968; Lord, 1980). In the 1PL model, the probability of a correct response of person n on item i is given by

$$P(X_{ni} = 1 | \theta_n, \beta_i) = \frac{\exp(\theta_n - \beta_i)}{1 + \exp(\theta_n - \beta_i)}, \quad (18.3)$$

where β_i is the item parameter of item i . A computational advantage of this model is that the respondent's NC score is a minimal sufficient statistic for the ability parameter. The alternative model, the 3PL model, is a more general model, where for each item two additional item parameters are introduced. First, the ability parameter θ is multiplied by an item parameter α_i , which is commonly referred to as the discrimination parameter. Second, the model is extended with a guessing parameter γ_i which allows for describing guessing behavior. In this model, the probability of a correct response of person n on item i , is

$$P(X_{ni} = 1 | \theta_n, \alpha_i, \beta_i, \gamma_i) = \psi_i(\theta_n) = \gamma_i + (1 - \gamma_i) \frac{\exp(\alpha_i \theta_n - \beta_i)}{1 + \exp(\alpha_i \theta_n - \beta_i)}. \quad (18.4)$$

Note that if $\gamma_i = 0$ and $\alpha_i = 1$, $\psi_i(\theta_n)$ specializes to the response probability in the 1PL model.

The advantage of the 3PL over the 1PL model is that guessing of responses can be taken into account. The examinations used here as an example contain multiple-choice items so the 3PL model may fit better than the 1PL model. An advantage of the 1PL model over the 3PL model is that with small sample sizes it may result in more stable parameter estimates (Lord, 1983).

To estimate the item parameters, a multiple-group MML estimation procedure is used. Let the elements of the vector of item parameters, ξ , be defined as $\xi_i = (\beta_i)$ for the 1PL model and as $\xi_i = (\alpha_i, \beta_i, \gamma_i)$ for the 3PL model. The probability of observing response pattern \mathbf{X}_n , conditional on the item administration vector \mathbf{d}_n , is given by

$$p(\mathbf{x}_n | \mathbf{d}_n, \theta_n, \xi) = \prod_{i=1}^I \left(\psi_i(\theta_n)^{x_{ni}} (1 - \psi_i(\theta_n))^{1-x_{ni}} \right)^{d_{ni}} \quad (18.5)$$

with $\psi_i(\theta_n)$ as defined in 18.4. Notice that through d_{ni} , the probability, $p(\mathbf{x}_n | \mathbf{d}_n, \theta_n, \xi)$ only depends on the parameters of the items actually administered to person n . The likelihood in Equation 18.5 implies the usual assumption of local independence; that is, it is assumed that the responses are independent given θ_n . Further, throughout this chapter we will make the assumption of independence between respondents and ignorable missing data. The latter holds because the design vectors \mathbf{d}_n are fixed.

Table 18.1 Cumulative Percentages of the Reference (1992) and New (1998) Population on the Reference and New Examination

Score	Reference population		New population	
	Reference exam	New exam	Reference exam	New exam
25	19.8	14.5	12.1	15.7
26	23.6	17.5	14.7	18.7
27	28.0	20.7	17.8	22.0
28	31.8	24.5	21.2	25.7
29	35.9	28.6	25.1	29.7
30	41.0	33.1	29.4	34.0
31	45.6	37.9	34.0	38.6
32	50.4	43.0	39.0	43.4
Mean	32.3	33.2	33.9	33.2
SD	7.5	7.0	6.8	7.3
SE (Mean)		0.41		0.39
SE (SD)		0.13		0.10

In the concurrent MML estimation procedure, it will be assumed that every group in the design is sampled from a specific ability distribution. So, for instance, the data in the design in Figure 18.1 are evaluated using seven ability distributions; that is, one distribution for the examinees administered the reference examinations, one for the examinees administered the new examination, and five were for the linking groups. Let the ability parameters of the respondents of group b have a normal distribution with density $g(\theta|\mu_b, \sigma_b)$. More specifically, the ability parameter of a random respondent n has a normal distribution with density $g(\theta_n|\mu_{b(n)}, \sigma_{b(n)})$, where $b(n)$ denotes the population to which person n belongs. The probability of observing response pattern \mathbf{x}_n , given \mathbf{d}_n , as a function of the item and population parameters is

$$p(\mathbf{x}_n|\mathbf{d}_n, \boldsymbol{\xi}, \mu_{b(n)}, \sigma_{b(n)}) = \int p(\mathbf{x}_n|\mathbf{d}_n, \theta_n, \boldsymbol{\xi})g(\theta_n|\mu_{b(n)}, \sigma_{b(n)})d\theta_n, \tag{18.6}$$

where $p(\mathbf{x}_n|\mathbf{d}_n, \theta_n, \boldsymbol{\xi})$ is defined in Equation 18.5. MML estimation boils down to maximizing the log-likelihood

$$L(\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_n \log p(\mathbf{x}_n|\mathbf{d}_n, \boldsymbol{\xi}, \mu_{b(n)}, \sigma_{b(n)}), \tag{18.7}$$

with respect to all item parameters $\boldsymbol{\xi}$ and all population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

All item and population parameters can be concurrently estimated on a common scale using readily available software, for instance with BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Béguin (2000) and Béguin and Glas (2001) presented an alternative Bayesian approach to estimation of IRT models in the framework of IRT-OS-NC equating, but that approach is beyond the scope of this chapter.

18.4 IRT-OS-NC Equating

Once the parameters of the IRT model have been estimated, the next step is equipercentile equating carried out as if respondents of one population had taken both examinations. Consider the example in Table 18.1. The example was computed using the data of the 1992 and 1998 English language comprehension examinations at HAVO level introduced above. The second and fourth column of Table 18.1 display the cumulative relative frequencies for the reference and new examination as observed in 1992 and 1998, respectively. These two cumulative distributions were based on the actual OS distributions obtained from the two examinations. These distributions are displayed in Figure 18.2. Figure 18.2 also contains estimates of two score distributions that were not observed: the score distribution on the 1992 examination if it had been administered to the 1998 population and the score distribution on the 1998 examination if it had been administered to the 1992 population. The estimates of these two distributions are based on the concurrent MML estimates of the item and population parameters. These two estimated score distributions were used to compute the cumulative distributions in the third and fifth column of Table 18.1. The estimation method will be discussed in the next section. Essentially, the distribution of the reference population (1992) on the new (1998) examination is computed using the parameters of the items of the 1998 examination and the population parameters of the 1992 population.

The third column of Table 18.1 contains a part of the cumulative score distribution of the reference population on the new examination. This cumulative score distribution is displayed in Figure 18.2b, together with a confidence interval and the observed cumulative distribution produced by the reference sample. The computation of confidence intervals will be returned to below. The cut-off score for the new examination is set in such a way that the expected percentage of respondents

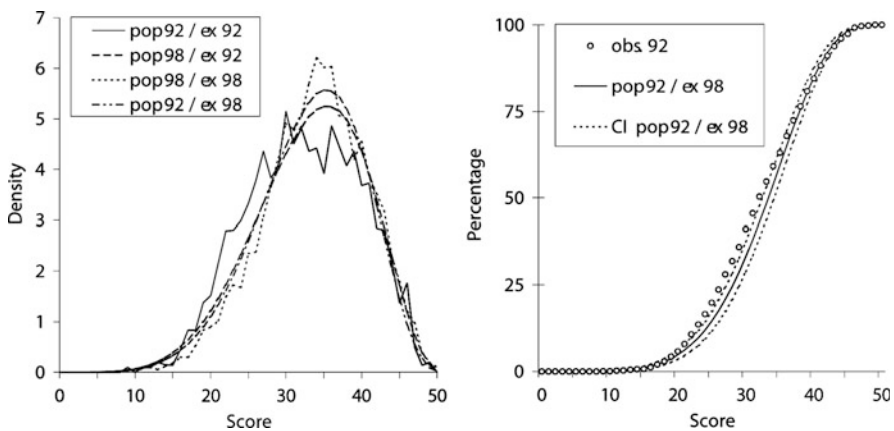


Fig. 18.2 Observed and expected score distributions

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 in Table 18.1, the cut-off score of the reference examination was 27; as a result, 28.0% failed the examination. The new cut-off score was set to 29, because at this cut-off score the percentage of the reference population failing the new examination was approximately equal to 28.0, which is the percentage failing the reference examination. Obviously, the new examination was easier. This is also reflected in the mean score of the two examinations displayed at the bottom of Table 18.1. The old and the new cut-off scores are marked by a straight line under the percentages. It can be seen that 25.1% of students in the new population failed the new examination, suggesting that the new population is more proficient than the reference population. This is also reflected in the mean scores of the two populations.

The procedure has two interesting aspects. First, the score distributions on the reference examination for the reference population can be obtained in two different ways: It can be computed directly from the data, as done above, or it can be estimated based on the IRT model. Analogously, the score distribution on the new examination for the new population also can be obtained in two different ways: using the actual examination data with the associated score distribution and using an estimate of this score distribution under an IRT model. The expected score distribution estimated under an IRT model will be referred to as *expected score distribution*, whereas the distribution determined directly from the data will be referred to as *OS distribution*. The difference between the observed and expected frequencies can be the basis for a Pearson-type test statistic for testing the model fit in either the reference-examination data set or the new-examination data set (see Glas & Verhelst, 1989, the R_o -statistic). In the example of equating examinations English HAVO, Figure 18.3 contains only expected-score distributions, whereas Figure 18.2 contains OS distributions for the reference population on the

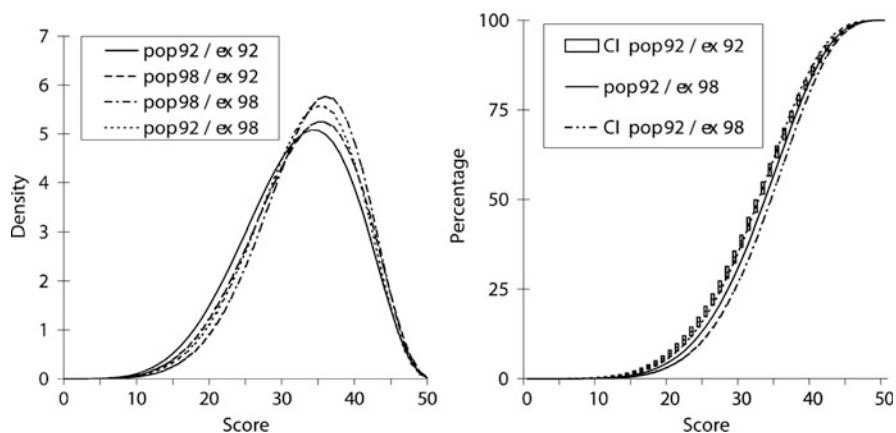


Fig. 18.3 Expected score distributions

reference examination and for the new population on the new examination. Consequently, equating could be performed in two different ways, based on the curves of Figure 18.3 or on the curves of Figure 18.2.

Second, the cut-off scores of the two examinations could be equated based on the score distributions of the new population or based on the score distributions of the reference population. If the IRT model fits, a choice between these possibilities should not influence the outcome of the equating procedure. This provides a basis for the evaluation of the appropriateness of the procedure, which we return to below, together with a comparison of the results obtained using the 1PL and 3PL models.

18.5 Some Computational Aspects of IRT-OS-NC Equating

After the parameters of the IRT model have been estimated on a common scale, the estimates of the frequency distributions are computed as follows. Label the populations in the design displayed in Figure 18.1 as $b = new, reference, 1, \dots, 5$. In this design, every population is associated with a specific design vector, d_b , that indicates which items were administered to the sample of examinees from population b . Let d_{ref} and d_{new} be the design vector of the examinees from the reference and new examination, respectively. The probability of obtaining a score r on the reference examination for the examinees in population b ($b = new, reference, 1, \dots, 5$) is denoted by $p(r; ref, b)$. Given the IRT item and population parameters, this probability is given by

$$p(r; ref, b) = \sum_{\{\mathbf{x}|r, \mathbf{d}_{ref}\}} \int p(\mathbf{x}|\mathbf{d}_{ref}, \theta, \boldsymbol{\xi}) g(\theta|\mu_b, \sigma_b) d\theta, \quad (18.8)$$

where $\{\mathbf{x}|r, \mathbf{d}_{ref}\}$ stands for the set of all possible response patterns on the reference examination resulting in a score r and $p(\mathbf{x}|\mathbf{d}_{ref}, \theta, \boldsymbol{\xi})$ is the probability of a response pattern on the reference examination given θ computed as defined in Equation 18.5. In the same manner, one can compute the probability of students in population b ($b = new, reference, 1, \dots, 5$) obtaining a score r on the new examination using

$$p(r; new, b) = \sum_{\{\mathbf{x}|r, \mathbf{d}_{new}\}} \int p(\mathbf{x}|\mathbf{d}_{new}, \theta, \boldsymbol{\xi}) g(\theta|\mu_b, \sigma_b) d\theta, \quad (18.9)$$

where $\{\mathbf{x}|r, \mathbf{d}_{new}\}$ stands for the set of all possible response patterns on the new examination resulting in a score r .

Computation of the distributions defined by Equations 18.8 and 18.9 involves summing over the set of all possible response patterns \mathbf{x} on some examination resulting in a score r , $\{\mathbf{x}|r, \mathbf{d}\}$, where \mathbf{d} is a design vector. For a given proficiency θ the score distribution is a compound binomial distribution (Kolen & Brennan,

1995). A recursion formula (Lord & Wingersky, 1984) can be used to compute the score distribution of respondents of a given ability. Define $f_k(r|\theta_i)$ as the probability of a NC score r over the first k items, given the ability θ . Define $f_1(r = 0|\theta) = (1 - p_1)$ as the probability of earning a score of 0 on the first item and $f_1(r = 1|\theta) = p_1$ as the probability of earning a score 1 on the first item. For $k > 1$, the recursion formula is given by

$$f_k(r|\theta) = \begin{cases} f_{k-1}(r|\theta)(1 - p_k) & \text{if } r = 0, \\ f_{k-1}(r|\theta)(1 - p_k) + f_{k-1}(r - 1|\theta)p_k & \text{if } 0 < r < k, \\ f_{k-1}(r - 1|\theta)p_k & \text{if } r = k. \end{cases} \quad (18.10)$$

The expected score distribution for a population can be obtained from Equations 18.8 and 18.9 by changing the order of summation and integration. The integration over a normal distribution can be evaluated using Gauss-Hermite quadrature (Abramowitz & Stegun, 1972). In the examples presented in this chapter, 180 quadrature points were used for every one of the seven ability distributions (new, reference and linking groups 1, . . . , 5) involved. At each of the quadrature points the result of the summation is obtained using the recursion formula shown in Equation 18.10.

18.6 Evaluation of the Results of the Equating Procedure

The cut-off scores of six examinations in language comprehension administered in 1998 were equated to the cut-off scores of reference examinations administered in 1992. The subjects of the examinations are listed under the heading “Subject” in Table 18.2. The examinations were administered at two levels: subjects labeled “D” in Table 18.2 were at the intermediate general secondary education level (MAVO-D-level), and subjects labeled “H” were at HAVO level. All examinations consisted of 50 dichotomous items. All designs were as depicted in Figure 18.1.

In this section, it is investigated whether the 1PL and 3PL models produced similar results. Two other aspects of the procedure will be compared. First, two

Table 18.2 Data Overview

Subject	Reference exam			New exam			Link
	N_{Ref}	Mean	SD	N_{New}	Mean	SD	N_{link}
English D	1,693	35.16	6.92	4,000	34.42	7.47	1,101
English H	2,039	32.32	7.45	4,000	33.89	6.76	673
German D	2,021	34.00	6.28	4,000	29.72	6.45	803
German H	2,129	34.51	5.59	4,000	32.19	6.57	327
French D	2,097	32.28	7.23	4,000	30.13	7.06	750
French H	2,144	35.72	6.80	4,000	31.48	6.69	454

Note: Mean = mean observed frequency distribution; SD = standard deviation observed frequency distribution

versions of the equating procedure are compared, one version where all relevant distributions are expected score distributions, and one version where the available OS distributions are used, that is, the score distribution observed in 1992 and the score distribution observed in 1998. Second, results obtained using either the reference or the new population as the basis for equating the examinations are compared.

In Table 18.3, the results of the equating procedure are given for the version of the procedure where only expected score distributions are used. These score distributions were obtained for the reference population and for the new population. In Table 18.3, equivalent scores on the new examinations are presented for

Table 18.3 Results of the Equating Procedure With Expected Score Distributions

Subject	$r^{(b)}$	ϕ_R^1	ϕ_R^3	ϕ_{RR}^{13}	ϕ_N^1	ϕ_N^3	Δ_{NN}^{13}	Δ_{RN}^{11}	Δ_{RN}^{33}
German D	25	23	23	0	23	23	0	0	0
	30	28	27	1	28	27	1	0	0
	31	29	28	1	29	28	1	0	0
	35	33	32	1	33	32	1	0	0
	40	39	37	2	39	37	2	0	0
German H	25	26	25	1	26	25	1	0	0
	30	31	30	1	31	30	1	0	0
	35	36	36	0	36	36	0	0	0
	40	41	41	0	41	41	0	0	0
English D	25	23	25	-2	23	25	-2	0	0
	28	26	27	-1	26	28	-2	0	-1
	30	29	29	0	29	30	-1	0	-1
	35	34	35	-1	34	35	-1	0	0
	40	40	40	0	40	40	0	0	0
English H	25	26	26	0	26	26	0	0	0
	27	28	28	0	28	28	0	0	0
	30	31	31	0	31	31	0	0	0
	35	36	36	0	36	36	0	0	0
	40	40	40	0	40	40	0	0	0
French D	25	27	27	0	27	27	0	0	0
	30	32	32	0	32	32	0	0	0
	35	37	37	0	37	37	0	0	0
	40	41	42	-1	41	42	-1	0	0
French H	25	23	23	0	23	23	0	0	0
	30	28	28	0	28	28	0	0	0
	35	34	33	1	34	33	1	0	0
	40	39	38	1	39	38	1	0	0
Abs. sum				14			16	0	2

r : number-correct score (boldface: cut-off score)

ϕ_R^1 : equated scores using 1PLM and Reference population

ϕ_R^3 : equated scores using 3PLM and Reference population

ϕ_N^1 : equated scores using 1PLM and New population

ϕ_N^3 : equated scores using 3PLM and New population

Δ_{RR}^{13} : difference $\phi_R^1 - \phi_R^3$; Δ_{NN}^{13} : difference $\phi_N^1 - \phi_N^3$

Δ_{RN}^{11} : difference $\phi_R^1 - \phi_N^1$; Δ_{RN}^{33} : difference $\phi_R^3 - \phi_N^3$

a number of different score points on the reference examination. The score points listed in Table 18.3 are the actual cut-off score and the score points, $r=25,30,35,40$ on the reference examination. In Table 18.3, the results pertaining to the actual cut-off scores are printed in boldface characters. The results obtained using score distributions based on the reference population are listed in Columns 3–5, and the results obtained using score distributions based on the new population are listed in Columns 6–8. The scores on the new examination associated with the scores of the reference examination computed using the 1PL model are given in Column 3. For instance, a score of 25 on the reference examination Reading Comprehension in German at MAVO-D level is equated to a score of 23 on the new examination, and a score of 30 on the reference examination is equated to a score of 28 on the new examination. In the Column 4, the scores obtained under the 3PL model are given. For this case, a score of 30 on the reference examination is equated to a score of 27 on the new one. Notice that for a score of 30 the results for the 1PL and the 3PL models differ by 1 score point. Column 5 contains the difference between the new scores obtained via the 1PL and 3PL models. For convenience, the sum of the absolute values of these differences is given at the bottom line of the table. So, for the 27 scores equated here, the absolute difference in equated score points computed using the 1PL and the 3PL models equals 14 score points, and the absolute difference between equated scores is never more than 2 points. The three following columns contain information comparable to the three previous ones, but the scores on the new examination were computed using score distributions for the new population. Notice that the results obtained using the reference and the new population are much alike. This is corroborated in the two last columns that contain the differences in results obtained using score distributions based on either the reference or new population. The column labeled “ Δ_{RN}^{11} ” shows the differences for the 1PL model and column labeled “ Δ_{RN}^{33} ” shows the differences for the 3PL model.

Two conclusions can be drawn from Table 18.3. First, the 1PL and 3PL models produce quite similar results. On average there was less than 1 score point difference, with differences never larger than 2 score points. The second conclusion is that using either the reference or new population for determining the difference between the examination made little difference. The bottom of Table 18.3 shows that the sum of the absolute values of the differences are 0 and 2 score points.

As already mentioned, the procedure can be carried out in two manners: one where all relevant score distributions are estimated using the IRT model and one where the available OS distributions of the two examinations are used. The above results used the former approach. Results of application of the second approach are given in Table 18.4.

The format of Table 18.4 is the same as the format of Table 18.3. The indices in Tables 18.3 and 18.4 (explained at the bottom of the two tables) are defined exactly the same, only they are computed using two alternative methods. In Table 18.3 expected score distributions are used, and in Table 18.4 the available OS distributions on the examination are used. The latter approach produced results that were far less satisfactory. For the 1PL model, the summed differences between using the reference and the new population, Δ_{RN}^{11} , rose from 0 to 11 score points. For the

Table 18.4 Results of the Equating Procedure With Observed Score Distributions

	$r^{(b)}$	ϕ_R^1	ϕ_R^3	ϕ_{RR}^{13}	ϕ_N^1	ϕ_N^3	Δ_{NN}^{13}	Δ_{RN}^{11}	Δ_{RN}^{13}
German D	25	23	23	0	23	22	1	0	1
	30	28	27	1	28	27	1	0	0
	31	29	28	1	29	28	1	0	0
	35	33	32	1	33	32	1	0	0
	40	39	36	3	39	37	2	0	-1
German H	25	25	25	0	26	24	2	-1	1
	30	31	30	1	31	30	1	0	0
	35	36	36	0	36	36	0	0	0
	40	41	41	0	42	41	1	-1	0
English D	25	23	25	-2	23	24	-1	0	1
	28	27	28	-1	27	27	0	0	1
	30	29	30	-1	28	29	-1	1	1
	35	34	35	-1	34	35	-1	0	0
	40	40	39	1	39	40	-1	1	-1
English H	25	27	27	0	27	26	1	0	1
	27	29	29	0	28	28	0	1	1
	30	32	31	1	31	31	0	1	0
	35	36	35	1	35	35	0	1	0
	40	40	40	0	40	40	0	0	0
French D	25	27	28	-1	27	27	0	0	1
	30	33	33	0	31	33	-2	2	0
	35	37	37	0	37	37	0	0	0
	40	41	41	0	42	42	0	-1	-1
French H	25	23	23	0	23	22	1	0	1
	30	28	28	0	28	27	1	0	1
	35	34	33	1	33	32	1	1	1
	40	39	38	1	39	38	1	0	0
Abs. sum				18			21	11	14

r : number-correct score (boldface: cut-off score)

ϕ_R^1 : equated scores using 1PLM and Reference population

ϕ_R^3 : equated scores using 3PLM and Reference population

ϕ_N^1 : equated scores using 1PLM and New population

ϕ_N^3 : equated scores using 3PLM and New population

Δ_{RR}^{13} : difference $\phi_R^1 - \phi_R^3$; Δ_{NN}^{13} : difference $\phi_N^1 - \phi_N^3$

Δ_{RN}^{13} : difference $\phi_R^1 - \phi_N^1$; Δ_{RN}^3 : difference $\phi_R^3 - \phi_N^3$

3PL model, this difference, Δ_{RN}^{33} , rose from 2 to 14 points. In other words, the requirement of an equating function that is invariant over populations (Petersen, Kolen, & Hoover, 1989) is better met when using frequencies that are all based on the IRT model. An explanation for this outcome is that results of equipercntile equating are vulnerable to fluctuations due to sampling error in the OS distributions. To overcome this problem it is common practice to use smoothing techniques to reduce the fluctuation in the score distributions. The expected score distribution computed using IRT models can be considered as a smoothed score distribution. Therefore, IRT equating using only expected score distributions reduces

fluctuations in the results of equating and, consequently, improves the invariance of equating over populations.

18.7 Confidence Intervals for Score Distributions

When a practitioner must set a cut-off score on an examination that is equivalent to the cut-off on some reference examination, the first question that comes to mind is about the reliability of the equating function. In the example in Table 18.1, a cut-off score of 27 on the reference examination is equated with a cut-off score 29 on the new examination upon observing that the observed 28.0% in the second column is closest to the 28.7% in the third column. To what extent are these percentages reliable? In Table 18.5, 90% confidence intervals are given for the estimated percentages on which equating is based. Their computation will be explained below. Consider the information on the English reading comprehension examination, which was also used for producing Table 18.1. In the boldface row labeled “English H,” information is given on the results of the reference population taking the reference examination. This row contains the observed percentage of students

Table 18.5 90% Confidence Intervals for Cumulative Percentages

Subject	Score	Observed %	Lower bound %	Upper bound %	Observed-Expected	Z
German D	31	33.1	31.6	33.0 34.4	0.0	0.06
	28		24.2	27.9 31.5		
	29		29.1	32.7 36.4		
	30		34.4	37.9 41.5		
German H	30	23.4	22.0	23.4 24.8	0.0	0.04
	30		14.8	18.4 21.9		
	31		19.3	22.7 26.1		
	32		24.4	27.6 30.7		
English D	28	18.1	16.3	17.5 18.6	0.6	0.84
	25		11.1	13.6 16.1		
	26		13.5	16.1 18.7		
	27		16.2	19.0 21.7		
English H	27	28.0	24.4	25.7 27.0	0.3	2.92
	27		17.0	20.7 24.4		
	28		20.7	24.5 28.3		
	29		24.7	28.6 32.4		
French D	25	18.8	17.5	18.6 19.7	0.2	0.29
	26		11.8	14.7 17.6		
	27		14.7	17.8 20.9		
	28		18.0	21.2 24.4		
French H	30	22.4	20.6	21.8 23.0	0.6	0.84
	27		13.1	17.5 22.0		
	28		16.0	20.7 25.3		
	29		19.4	24.2 29.0		

Note: Obs. % = observed cumulative percentage; Expected % = expected cumulative percentage under the one-parameter-logistic model. Lower and upper bounds are of 90% confidence interval of expected %. Z: normalized difference Observed – Expected

scoring 27 points or less, the cumulative percentage under the 1PL model, the lower and upper bound of the 90% confidence interval for this percentage, and the difference and normalized difference between the observed cumulative percentage and the cumulative percentage under the 1PL model. The normalized difference was computed by dividing the difference by its standard error. This normalized difference can be seen as a very crude measure of model fit. Together with the plots of the frequency distributions given in Figures 18.1 and 18.2, these differences give a first indication of how well the model applied.

Continuing the example labeled “English H” in Table 18.5, in the three rows under the boldface row, for three scores the estimates of the cumulative percentages for the reference population on the new examination and their confidence intervals are given. These three scores are chosen in such a way that the middle score is the new cut-off score if equating is performed using only expected score distributions. The other two scores can be considered as possible alternative cut-off scores. For instance, in the “English H” example in Table 18.5, the observed cumulative percentage 28% is located within the confidence band related to score 29, while it is near the upper confidence bound related to score 28. If the observed percentage of the cut-off score is replaced by an expected percentage, the confidence band of this estimate, which is given in the boldface row, also comes into play.

But the basic question essentially remains the same: Are the estimates precise enough to justify equating an old cut-off score to a unique new cut-off score? Or are the random fluctuations such that several new cut-off scores are plausible? Summarizing the results of Table 18.5, the exams German D and German H each has only one plausible cut-off score of 29 and 31, respectively. Notice that the cumulative percentages of 33.0% and 23.4% are well outside the confidence bands of the scores directly above and below the chosen cut-off score. For the exams French D, English D, and English H, two cut-off scores could be considered plausible. For the exams in English it also made a difference whether equating was performed using only expected score distributions or using available OS distributions. Finally, for the examination French H, the confidence interval of Score 27, 28 and 29 contained the percentage 21.8. So using the cumulative percentage of examinees under the cut-off score on the reference examination estimated under the IRT model, all three scores could be considered plausible values for the cut-off score in the new examination.

18.8 Computation of Confidence Intervals for Score Distributions

In this section, the parametric and nonparametric bootstrap method (Efron, 1979, 1982; Efron & Tibshirani, 1993) will be introduced as a method for computing confidence intervals of the expected score distributions. The nonparametric bootstrap proceeds by resampling with replacement from the data. The sample size is the same as the size of the original sample, and the probability of an element being

sampled is the same for all response patterns in the original sample. By estimating the parameters of the IRT model on every sample, the standard error of the estimator of the computed score distribution can be evaluated. In the parametric bootstrap, new values for the parameters are drawn based on the parameter estimates and estimated inverse information matrix. Using these repeated draws, the score distributions can be computed and their standard errors can be evaluated by assessing the variance over repeated draws.

Results of application of both bootstrap procedures are presented for the data from the English language proficiency examination on HAVO level in 1992 and 1998. These data were also used for producing the Figures 18.2 and 18.3 and Table 18.1. The confidence intervals presented in the two figures and the standard

Table 18.6 Confidence Intervals Using Bootstrap Procedures, English HAVO 1998, Population 1992

<i>r</i>	Nonparametric, 400 replications				Parametric, 400 replications			
	P_r	Cum.	SE(P_r)	SE(Cum.)	E(P_r)	Cum.	SE(P_r)	SE(Cum)
5	0.003	0.01	0.001	0.001	0.004	0.01	0.001	0.002
10	0.054	0.13	0.010	0.026	0.057	0.14	0.013	0.036
15	0.318	1.06	0.041	0.157	0.329	1.10	0.053	0.206
20	1.085	4.68	0.097	0.519	1.106	4.82	0.119	0.658
25	2.566	14.27	0.147	1.151	2.590	14.52	0.172	1.410
30	4.463	32.79	0.130	1.817	4.469	33.12	0.148	2.157
35	5.594	59.14	0.086	1.980	5.571	59.41	0.114	2.292
40	4.439	84.73	0.213	1.254	4.405	84.84	0.248	1.433
45	1.460	98.13	0.156	0.258	1.450	98.14	0.178	0.296
50	0.019	100.00	0.004	0.000	0.020	100.00	0.005	0.000
	Mean	33.25	<i>SD</i>	6.98	Mean	33.20	<i>SD</i>	7.01
	<i>SE</i>	0.35	<i>SE</i>	0.10	<i>SE</i>	0.41	<i>SE</i>	0.13

Note: *r* = number-correct score; P = estimated percentage; Cum. = cumulative percentage

Table 18.7 Confidence Intervals Using Bootstrap Procedures English HAVO 1998, Population 1998

<i>r</i>	Nonparametric, 400 replications				Parametric, 400 replications			
	P_r	Cum.	SE(P_r)	SE(Cum.)	E(P_r)	Cum.	SE(P_r)	SE(Cum)
5	0.002	0.00	0.000	0.001	0.002	0.00	0.000	0.001
10	0.038	0.09	0.004	0.011	0.038	0.09	0.004	0.012
15	0.244	0.78	0.015	0.062	0.244	0.78	0.018	0.070
20	0.901	3.73	0.030	0.183	0.901	3.73	0.038	0.218
25	2.288	12.07	0.037	0.345	2.290	12.08	0.049	0.437
30	4.257	29.33	0.050	0.465	4.261	29.35	0.052	0.616
35	5.702	55.51	0.063	0.526	5.705	55.55	0.065	0.651
40	4.839	82.63	0.054	0.447	4.835	82.67	0.070	0.481
45	1.707	97.76	0.065	0.130	1.700	97.77	0.067	0.130
50	0.024	100.00	0.003	0.000	0.024	100.00	0.003	0.000
	Mean	33.90	<i>SD</i>	6.84	Mean	33.90	<i>SD</i>	6.83
	<i>SE</i>	0.09	<i>SE</i>	0.07	<i>SE</i>	0.11	<i>SE</i>	0.07

Note: *r* = number-correct score; P = estimated percentage; Cum. = cumulative percentage

errors reported in the two bottom lines of Table 18.1 were computed using a parametric bootstrap procedure with 400 replications. Table 18.6 shows results for the parametric and nonparametric bootstrap estimate of the score distribution of the 1992 population on the 1998 examination using the 1PL model. Table 18.7 shows the results for the 1998 population on the 1998 examination using the 1PL model. For brevity, only the results for every fifth score point are presented. In the two bottom lines of the tables, the mean, the standard deviation, and their standard errors are given.

The standard errors in Table 18.7 are much smaller than the standard errors in Table 18.6. So, the computed standard errors dropped markedly when the score distribution was estimated on the test the candidates actually took. For instance, the standard error of the mean using the nonparametric bootstrap was 0.09, markedly smaller than 0.35, the standard error of for the mean on the test not taken by the candidates. The same results held for the estimated score distributions. For instance, the standard error of the estimate of the percentage of candidates with score 25 dropped from .147 to .037. This effect can also be seen in the two bottom lines of Table 18.1. For instance, for the reference population, the standard error of the mean was .16 for the reference examination and .38 for the new examination. This difference, of course, was as expected, since the data provide more information on the examination 1998 for population 1998 than for population 1992. Furthermore, the estimated standard error of the nonparametric bootstrap is a bit smaller than the estimate using the parametric bootstrap. This result is as expected since the parametric bootstrap accounts for an additional source of variance, that is, the uncertainty about the parameters. Therefore, in this context, the parametric bootstrap is preferred over the nonparametric bootstrap. A disadvantage of the parametric bootstrap is that it cannot be applied in problems where the number of parameters is such that the inverse of the information matrix cannot be precisely computed. An example is the 3PL model in the above design. With two examinations of 50 items each and seven population parameter distributions, the number of parameters is 312. In such cases, the nonparametric bootstrap is the only feasible alternative.

18.9 A Wald Test for IRT-OS-NC Equating

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; Glas, 1988, 1999; Glas & Verhelst, 1989, 1995; Molenaar, 1983; and Orlando & Thissen, 2000). Besides the model violations covered by these statistics, in the present application one specific violation deserves special attention: the question whether the data from the linking groups are suited for equating 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, that is, into G subsamples. To every one of these data sets, the data of one or more linking groups are added, but in such a way that the data sets will have no linking groups in common. Summing up, each data set consists of a sample from the data of both the examinations and of one or 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: (a) by computing equivalent scores as was done above and evaluating whether the two equating functions produce similar results and (b) by performing a Wald test. The Wald test will be explained first.

Glas and Verhelst (1995) 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 subsamples of the complete 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 subsamples against the alternative that they are not. Let the parameters of the IRT model for the g -th subsample be denoted λ_g , $g \in \{1, 2, \dots, G\}$. Define a vector $\mathbf{f}(\lambda)$ with elements P_r , where P_r is the probability of obtaining a score r such as defined in Equations 18.8 and 18.9. In the example below, this will be the expected score distribution on the reference examination. Because of the restriction $\sum_r P_r = 1$, at least one proportion P_r is deleted. Let $\mathbf{f}(\lambda_g)$ be a distribution computed using the data of subsample g . Further, let $\hat{\lambda}_g$ and $\hat{\lambda}_{g'}$ be the parameter estimates in two subsamples g and g' , respectively. We will test the null hypothesis that the two score distributions are identical, that is,

$$\mathbf{h} = \mathbf{f}(\hat{\lambda}_g) - \mathbf{f}(\hat{\lambda}_{g'}) = 0. \quad (18.11)$$

The difference \mathbf{h} is estimated using independent samples of examination candidates and different and independent linking groups. Since the responses of the two subsamples are independent, the Wald test statistic is given by the quadratic form

$$W = \mathbf{h}'[\Sigma_g + \Sigma_{g'}]^{-1}\mathbf{h}, \quad (18.12)$$

where Σ_g and $\Sigma_{g'}$ are the covariance matrices of $\mathbf{f}(\lambda_g)$ and $\mathbf{f}(\lambda_{g'})$, respectively. W is asymptotically chi-square distributed with degrees of freedom equal to the number of elements of \mathbf{h} (Wald, 1943). For shorter tests, W can be evaluated using MML estimates and the covariance matrices can be explicitly computed. For longer tests and models with many parameters, such as the 3PL model, both the covariance matrices and the value of the test statistic W can be estimated using the nonparametric bootstrap method described above. This approach was followed in the present example.

Some results of the test are given in Table 18.8. 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

Table 18.8 Results of the Wald Test for Stability of Estimated Score Distributions, by Population Subsample

Subject	Reference		New	
	1 vs 2	3 vs 4	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

** $p < 0.01$

subsamples of approximately equal sample size. Next, four data sets were assembled, each consisting of the data of one linking group, the data of one of the four subsamples from the reference examination, and the data of one of the four subsamples from the new examination. The design for these four new data sets is similar to the design depicted in Figure 18.1, 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 3PL model 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 18.8. 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. In Table 18.8, the Wald tests with a significance probability less than .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 .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 where equivalent scores were computed for a partition of the data into five different subsamples, each one with its own linking group.

Consider Table 18.9. For six topics four scores on the reference test were considered. For each of the five subsamples, these four scores were equated to scores on the new examination via the reference population. The columns labeled “L1” to “L5” show the resulting scores on the new test. 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. The column labeled “Total” displays the sum of the absolute differences between all pairs of new scores. Since there are five new scores for every original score, there are 10 such pairs. So, for instance, the mean absolute difference

Table 18.9 Stability of Equating Functions in Subsamples

Topic	r(b)	L1	L2	L3	L4	L5	Total	Expected	p
German D	20	16	23	21	15	14	48	15.5	0.00
	25	20	28	27	21	19	50	14.5	0.00
	30	26	32	32	27	24	44	13.1	0.00
	35	31	37	37	33	29	44	11.4	0.00
German H	20	16	19	17	21	17	24	15.2	0.10
	25	22	24	22	26	22	20	12.4	0.15
	30	27	29	27	31	28	20	10.3	0.05
	35	33	34	32	36	33	18	9.5	0.10
English D	20	20	26	18	19	20	34	14.1	0.00
	25	24	31	23	24	25	34	12.5	0.00
	30	29	35	28	29	30	30	10.3	0.00
	35	34	39	33	34	34	24	8.8	0.00
English H	20	21	26	19	18	23	40	12.8	0.00
	25	26	31	24	23	28	40	12.0	0.00
	30	31	36	29	28	32	38	10.0	0.00
	35	36	40	34	33	37	34	9.2	0.00
French D	20	18	13	19	16	23	46	13.2	0.00
	25	24	18	24	20	27	44	13.7	0.00
	30	29	22	29	25	32	48	13.4	0.00
	35	35	28	34	29	36	44	12.7	0.00
French H	20	21	20	18	18	19	16	16.0	0.55
	25	26	25	23	24	24	14	15.4	0.75
	30	31	30	29	29	29	10	12.8	0.85
	35	36	35	34	34	34	10	10.7	0.70

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 subsamples. 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 18.2. These new generated data sets conformed the null hypothesis of the 3PL model. Next, for every data set, the procedure of equating the two examinations via the reference population in the five subsamples 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 3PL model (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 18.9. The overall model fit is not very good, however. Also, here French at HAVO-level stands out as well fitting, and German at HAVO-level shows acceptable model fit.

18.10 Conclusions

In this chapter, we proposed some heuristic methods and a more formal model test for the evaluation of the robustness of IRT-OS-NC equating, and we showed the feasibility of the methods in a practical situation using an application in a real examination situation. In the application, the differences between the results obtained using the 1PL and the 3PL models were not very striking. Overall model fit was not very satisfactory; only one of the examination topics fitted well, and a second topic fitted acceptably. The case presented here was further analyzed by Béguin (2000) and by Béguin and Glas (2001) using multidimensional IRT models (Bock, Gibbons, & Muraki, 1988) and a Bayesian approach to estimation. However, the methods presented in this chapter easily can be adapted to an MML framework for multidimensional IRT models; the main difference is replacing the normal ability distribution $g(\theta|\mu_b, \sigma_b)$ with a multivariate normal distribution $g(\theta|\mu_b, \Sigma_b)$.