Chapter 11
Reliability Issues in High-Stakes Educational Tests

Cees A. W. Glas

Abstract High-stakes tests and examinations often give rise to rather specific measurement problems. Though nowadays item response theory (IRT) has become the standard theoretical framework for educational measurement, in practice, number-correct scores are still prominent in the definition of standards and norms. Therefore, in this chapter methods are developed for relating standards on the number-correct scale to standards on the latent IRT scale. Further, this chapter focuses on two related issues. The first issue is estimating the size of standard errors when equating older versions of a test to the current version. The second issue is estimating the local reliability of number-correct scores and the extra error variance introduced through number-correct scoring rather than using IRT proficiency estimates. It is shown that the first issue can be solved in the framework of maximum a posteriori (MAP) estimation, while the second issue can be solved in the framework of expected a posteriori (EAP) estimation. The examples that are given are derived from simulations studies carried out for linking the nation-wide tests at the end of primary education in the Netherlands.

11.1 Outline of the Problem

The problem addressed here is that the standard scoring rule in much educational measurement, that is, the number-correct score, is not the same one as the optimal scoring rule that is derived from the IRT model that fits the data. In this chapter, a method is outlined for how to evaluate the consequences of this discrepancy for an important inference that is often made using IRT, that is, the consequences for test equating. To explain this further, we first introduce an IRT model and outline the principle of test equating.

The IRT models used in this chapter are the one-, two- and three-parameter Logistic models. The data are responses of students labeled with an index $n = 1, \ldots, N$ to

C. A. W. Glas (✉)
University of Twente, Enschede, The Netherlands
e-mail: c.a.w.glas@utwente.nl

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_11

items labeled with an index $i = 1, \ldots, K$. To indicate whether a response is available, we define a variable

$$d_{ni} = \begin{cases} 1 \text{ if a response of student } n \text{ to item } i \text{ is available} \\ 0 \text{ if this is not the case.} \end{cases} \tag{11.1}$$

The responses will be coded by a stochastic variable Y_{ni}. In the sequel, upper-case characters will denote stochastic variables and lower-case characters will denote realizations. In the present case, there are two possible realizations, defined by

$$y_{ni} = \begin{cases} 1 \text{ if } d_{ni} = 1 \text{ and student } n \text{ gave a correct response to item } i \\ 0 \text{ if } d_{ni} = 1 \text{ and student } n \text{ did not give a correct response to item } i \\ c \text{ if } d_{ni} = 0, \text{ where } c \text{ is an arbitrary constant unequal 0 or 1.} \end{cases} \tag{11.2}$$

Define the logistic function $\Psi(.)$ as:

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

In the 3-parameter logistic model (3PLM, Birnbaum 1968) the probability of a correct response depends on three item parameters, a_i, b_i and c_i which are called the discrimination, difficulty and guessing parameter, respectively. The parameter θ_n is the latent proficiency parameter of student n. The model is given by

$$\begin{aligned} P_i(\theta_n) &= c_i + (1 - c_i) + \Psi(a_i(\theta_n - b_i)) \\ &= c_i + (1 - c_i)\frac{\exp(a_i(\theta_n - b_i))}{1 + \exp(a_i(\theta_n - b_i))}. \end{aligned} \tag{11.3}$$

The 2-parameter logistic model (2PLM, Birnbaum 1968) follows by setting the guessing parameter equal to zero, so by introducing the constraint $c_i = 0$. The 1-parameter logistic model (1PLM, Rasch 1960) follows by introducing the additional constraint $a_i = 1$.

Note that in the application of the models in high-stakes situations, the number of proficiency parameters θ_n can become very large. Besides the practical problem of computing estimates of all model parameters concurrently, this also leads to theoretical problems related to the consistency of the estimates (see, Neyman and Scott 1948; Kiefer and Wolfowitz 1956). Therefore, it is usually assumed that the proficiency parameters are drawn from one or more normal proficiency distributions, indexed $g = 1, \ldots, G$, which are often also referred to as population distributions. That is, θ_n has the density function

$$g(\theta_n; \mu_{g(n)}, \sigma^2_{g(n)}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\theta_n - \mu_{n(g)})^2}{\sigma^2}\right), \tag{11.4}$$

Table 11.1 Example of proficiency estimates and their standard errors on two linked tests

Score	Test A				Test B			
	Freq	Prob	θ	$Se(\theta)$	Freq	Prob	θ	$Se(\theta)$
0	156	0.03	−1.91	0.54	6	0.00	−2.40	0.36
1	504	0.13	−1.34	0.50	16	0.00	−2.03	0.34
2	1055	0.34	−0.81	0.47	52	0.02	−1.68	0.33
3	1077	0.56	−0.38	0.45	122	0.04	−1.31	0.33
4	839	0.73	0.02	0.42	261	0.09	−0.96	0.34
5	658	0.86	0.41	0.40	516	0.20	−0.57	0.36
6	367	0.93	0.78	0.37	956	0.39	−0.17	0.37
7	194	0.97	1.15	0.37	1194	0.63	0.25	0.38
8	102	0.99	1.51	0.38	978	0.82	0.71	0.40
9	42	1.00	1.87	0.39	638	0.95	1.19	0.42
10	6	1.00	2.22	0.41	261	1.00	1.73	0.46

where *g(n)* is the population to which student *n* belongs.

Test equating relates the scores on one test to the scores on another test. Consider a simulated example based on the estimates displayed in Table 11.1. The estimates emanate from two tests. A sample of 5000 students of a population A was given a test A consisting of the items $i = 1,\ldots,10$, while a sample of 5000 other students of a population B was given a test B consisting of the items $i = 6,\ldots,15$. So the anchor between the two tests, that is, the overlap between the two tests, consists of 5 items. The anchor supports the creation of a common scale for all parameter estimates. The responses were generated with the 2PLM. The difficulties of the two tests differed: test A had a mean difficulty parameter, $\overline{b}_A$, of 0.68, while the difficulty level of test B, $\overline{b}_B$, was equal to −0.92. The mean of the proficiency parameters θ_n of sample A, μ_A was equal to −0.25, while the mean of the proficiency parameters of sample B, μ_B was equal to 0.25. The variances of the proficiency parameters and the mean of the discrimination parameters were all equal to one.

Suppose that test A has a cutoff score of 4, where 4 is the highest number-correct score that results in failing the test. In the fourth column of Table 11.1, the column labeled θ, it can be seen that the associated estimate on the latent θ-scale is 0.02. We chose this point as a latent-cutoff point, that is, $\theta_0 = 0.02$. If the Rasch model would hold for these data, the number-correct score would be the sufficient statistic for θ. In the 2PLM, the relation between a number-correct score and a θ-estimate is more complicated; this will be returned to below. Through searching for number-correct scores on Test B with θ-estimates closest to the latent cutoff point, we find that a cutoff score 6 on Test B best matches a cutoff score 4 on Test A. This conclusion is consistent with the fact the average difficulty of test A was higher than the average difficulty of test B. On the other hand, the sample administered test B was more proficient than the sample of test A. The columns labeled "Freq" and "Prob" give the frequency distributions of the number-correct scores and the associated cumulative

proportions, respectively. Note that 73% of sample A failed their test, while 39% of sample B failed theirs. Again, this is as expected.

The next question of interest is the reliability of the equating procedure. This can be translated into the question how precise the two cutoff scores can be distinguished. If we denote the cutoff scores by S_A and S_B, and denote the estimates of the positions on the latent scale associated with these two cutoff points by $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$, then $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ can be used as a measure of the precision with which we can distinguish the two scores. The estimates $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$ are not independent. Firstly, they both depend on the same linked data set and, secondly, they both depend on a concurrent estimate of all item-parameters, a_i, b_i and c_i, and (functions of) all latent proficiency parameters θ_n. Therefore, the standard error $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ cannot be merely computed as the square root of $Var(\hat{\theta}_{SA} - \hat{\theta}_{SB}) = Var(\hat{\theta}_{SA}) + Var(\hat{\theta}_{SB})$, but the covariance of the estimates must be taken into account also. The method to achieve this is outlined below, after the outline of a statistical framework and considering the problem of test scoring with number-correct scores when these are not sufficient statistics.

11.2 Preliminaries

Nowadays, marginal maximum likelihood (MML, see, Bock and Aitkin 1981) and fully Bayesian estimation (Albert 1992; Johnson and Albert 1999) are the prominent frameworks for estimating IRT models. Mislevy (1986, also see, Glas 1999) point out that they are closely related, because MML estimation is easily generalized to Bayes modal estimation, an estimation method that seeks the mode of the posterior distribution rather than the mode of the likelihood function. In this chapter, we adopt the MML and Bayes modal framework. In this framework, it is assumed that the θ-parameters are drawn from a common distribution, say, a population proficiency distribution as defined in Formula (11.4). Estimates of the item parameters and the parameters of the population proficiency distribution are obtained by maximizing a likelihood function that is marginalized with respect to the θ-parameters.

An important tool for deriving the estimation equations is Fisher's identity (Efron 1977; Louis 1982). For this identity, we distinguish N independent observations y_n and unobserved data z_n. The identity states that the first order derivatives of the parameters of interest δ with respect to the log-likelihood function $L(.)$ are given by

$$\frac{\partial L(\delta)}{\partial \delta} = \sum_{n=1}^{N} E_{z|y}(\nabla_n(\delta)|\, y_n) = \sum_{n=1}^{N} \int \ldots \int \left[\frac{\log p(y_n, z_n; \delta)}{\partial \delta}\right] p(z_n|y_n; \delta) dz_n, \tag{11.5}$$

where $p(y_n, z_n; \delta)$ is the likelihood if z_n would be observed, $\nabla_n(\delta)$ is the first-order derivative of its logarithm, and $p(z_n|y_n; \delta)$ is the posterior distribution of the unobserved data given the observations.

Bock and Aitkin (1981) consider the θ-parameters as unobserved data and use the EM-algorithm (Dempster et al. 1977) for maximum likelihood estimation from incomplete data to obtain estimates of the item and population parameters. In this framework, Glas (1999, 2016) uses Fisher's identity to derive estimation and testing procedures for a broad class of IRT models.

Standard errors can be obtained as the square roots of the covariance matrix of the estimates $Cov(\hat{\delta}, \hat{\delta})$ which can be obtained by inverting the observed Fisher information matrix, say, $Cov(\hat{\delta}, \hat{\delta}) = I(\hat{\delta}, \hat{\delta})^{-1}$. Louis (1982) shows that this matrix is given by

$$I(\delta, \delta) = -\frac{\partial^2 L(\delta)}{\partial \delta \, \partial \delta^t} = -\sum_{n=1}^{N} E_{z|y}\big(\nabla_n(\delta, \delta^t)\big|\, y_n\big) - Cov_{z|y}\big(\nabla_n(\delta)\nabla_n(\delta)^t\big|\, y_n\big), \tag{11.6}$$

where $\nabla_n(\delta, \delta^t)$ stands for the second-order derivatives of $\log p(y_n, z_n; \delta)$ with respect to δ. Evaluated at the MML estimates, the information matrix can be approximated by

$$I(\hat{\delta}, \hat{\delta}) \approx \sum_{n=1}^{N} E_{z|y}\big(\nabla_n(\delta)\nabla_n(\delta)^t\big|\, y_n\big) \tag{11.7}$$

(see Mislevy 1986). In the next sections, this framework will be applied to the issues addressed in this chapter: the reliability of tests scored with number-correct scores and to equating errors.

11.3 MAP Proficiency Estimates Based on Number-Correct Scores

Glas (1999, 2016) shows how the estimation equations for the item and population parameters of a broad class of IRT models can be derived using Fisher's identity. This identity can also be applied to derive an estimation equation for a proficiency estimate based on a number-correct score

$$s = \sum_{i=1}^{k} d_i y_i, \tag{11.8}$$

with d_i and y_i as defined in (11.1) and (11.2) dropping the subscript n. The application of Fisher's identity is based on viewing a response pattern as unobserved and the number-correct score as observed. Define $L_s(\theta)$ as the product of the normal prior distribution $g(\theta; \lambda)$ with $\lambda = (\mu, \sigma^2)$ and the probability of a number-correct score s given θ. Define $\{y|s\}$ as the set of all response patterns resulting in a number correct

score s. Then the probability of a number-correct score s given θ is equal to the sum over $\{y|s\}$ of the probabilities of response patters $P(y|\theta, \beta)$ given item parameters β and proficiency parameters θ. Application of Fisher's identity results in a first order derivative

$$\frac{\partial L_s(\theta)}{\partial \theta} = E_{y|s}(\nabla(\theta)|\, s, \beta) = \frac{\sum_{\{y|s\}} \left[\frac{\partial \log P(y,\theta;\beta,\lambda)}{\partial \theta}\right] P(y|\theta, \beta)}{\sum_{\{y|s\}} P(y|\theta, \beta)}. \tag{11.9}$$

Equating this expression to zero gives the expression for the MAP estimate. Computation of the summation over $\{y|s\}$ can be done using the recursive algorithm by Lord and Wingersky (1984). The algorithm is also used by Orlando and Thissen (2000) for the computation of expected a-posteriori estimates of θ given a number-correct score s.

Note that in expression (11.9), the prior $g(\theta; \lambda)$ cancels in the posterior, so $p(y|s; \theta, \beta, \lambda) \equiv p(y|s; \theta, \beta)$.

As an example, consider the 2PLM, given by expression (11.3) with $c_i = 0$. The probability of a response pattern becomes

$$\begin{aligned} L_s(\theta) = &\sum_{\{y|s\}} \log P(y, \theta; \beta, \lambda) = \log g(\theta; \mu, \sigma^2) \\ &+ \sum_{\{y|s\}} \sum_{i=1}^{K} \log\big(P_i(\theta)^{d_i y_i}(1 - P_i(\theta))^{d_i(1-y_i)}\big), \end{aligned} \tag{11.10}$$

and

$$\frac{\partial L_s(\theta)}{\partial \theta} = \frac{\mu - \theta}{\sigma^2} + \sum_{\{y|s\}} \sum_{i=1}^{K} (d_i a_i (y_i - P_i(\theta)))\, p(y|s; \theta, \beta)\,. \tag{11.11}$$

The estimation equation can be solved by either the Newton-Raphson algorithm, or by the EM algorithm. Standard errors can be based on observed information as defined in expression (11.7). One way of estimating θ and computing the standard errors is to impute the item parameters as known constants. However, when we want to compare the estimated proficiencies obtained for two tests through their difference, say, $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$, we explicitly need to take the precision of the estimates of all item and population parameters into account. How this is accomplished is outlined in the next section.

11.4 Equating Error

Suppose θ_0 is a cutoff point on the latent scale and we want to impose this cutoff point on several test versions. Further, we want to estimate the reliability of the created link. Three procedures for the computation of equating errors will be discussed, using some possible data collection designs displayed in Fig. 11.1.

To introduce the first method, consider the design displayed in Fig. 11.1a. In this design, students were administered both test versions, that is, Version A and Version B. The first measure for the strength of the link is based on the standard error of the difference between the average difficulties of the two versions, say, $Se(\overline{b}_A - \overline{b}_B)$, where $\overline{b}_A$ is the estimate of the mean difficulty of Version A and $\overline{b}_B$ the estimate of the mean difficulty of Version B. The strength of the link is mainly determined by the number of students, but also by the number of item parameters making up the two means. Since the estimates are on a latent scale that is subject to linear transformations, we standardize the standard error with the standard deviation of the proficiency distribution. This leads to the definition of the index

$$\text{Equating Error} = \frac{Se(\overline{b}_A - \overline{b}_B)}{Sd(\theta)}. \tag{11.12}$$

The standard error can be computed as the square root of $Var(\overline{b}_A - \overline{b}_B)$, which can be computed by pre- and post-multiplying the covariance matrix by a vector of weights, that is, $\boldsymbol{w}^t Cov(\hat{\delta}, \hat{\delta})\boldsymbol{w}$,

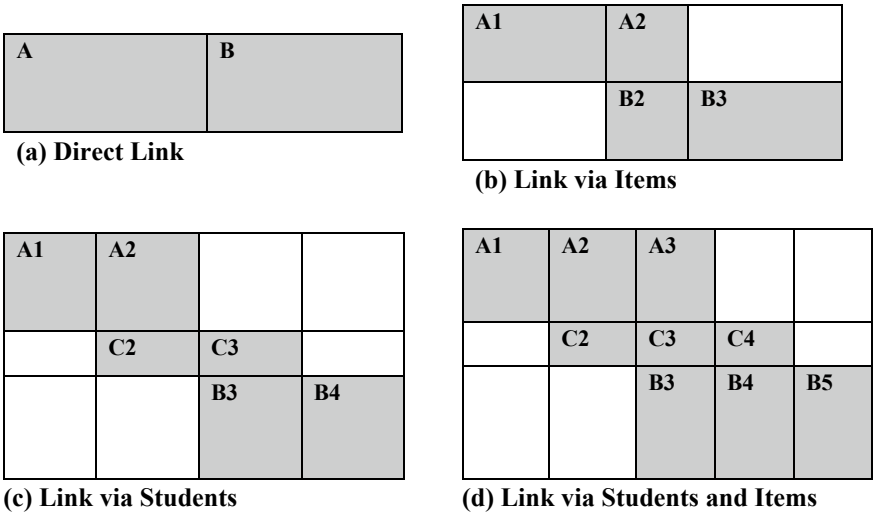

Fig. 11.1 Four designs for test equating

$$\text{where } \boldsymbol{w} \text{ has elements } \quad \mathrm{w}_j = \begin{cases} \frac{d_{iA}}{\Sigma_i d_{iA}} - \frac{d_{iB}}{\Sigma_i d_{iB}} & \text{if } j \text{ is related to } Cov(\hat{b}_i, \hat{b}_i) \\ 0 & \text{if this is not the case,} \end{cases} \tag{11.13}$$

where d_{iA} and d_{iB} are defined by expression (11.1), for a student administered test A and a student administered test B, respectively.

Figure 11.1b gives an example of equating two tests via common items (the so-called anchor). The test consisting of the items A1 and A2 is linked to the test consisting of the items B2 and B3, because A2 and B2 consist of the same items. The larger the anchor, the stronger the link. In this design it is usually assumed that the means of the two proficiency distributions are different. This leads to a second definition of an index for equating error, that is:

$$\text{Equating Error} = \frac{Se(\hat{\mu}_A - \hat{\mu}_B)}{Sd(\theta)}, \tag{11.14}$$

where $Sd(\theta)$ is a pooled estimate of the standard deviations of the proficiency distributions of the two populations. In Fig. 11.1c, the test consisting of parts A1 and A2 and the test consisting of the parts B3 and B4 have no items in common, but a link is forged by the students administered C2 and C3.

Again, the standard error can be computed as the square root of the associated variance, which can be computed by pre- and post-multiplying the covariance matrix of the parameter estimates by a vector of weights, that is, $\boldsymbol{w}^t Cov(\hat{\delta}, \hat{\delta})\boldsymbol{w}$, where $\boldsymbol{w}$ has elements

$$w_j = \begin{cases} 1 & \text{if } j \text{ is related to } Cov(\hat{\mu}_A, \hat{\mu}_A) \\ -1 & \text{if } j \text{ is related to } Cov(\hat{\mu}_B, \hat{\mu}_B) \\ 0 & \text{if this is not the case.} \end{cases} \tag{11.15}$$

A third method to assess a equating error is based on the position of the cutoff point on the latent scale. This approach gives a more precise estimate of the equating error of the cutoff point, but below it becomes clear that it is somewhat more complicated to compute. Suppose θ_0 is the cutoff point on the latent scale. On both tests, we choose an observed cutoff score, say S_A and S_B, that are associated with the same (mean) proficiency level θ_0. Then an equating error index can be defined as

$$\text{Equating Error} = \frac{Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})}{Sd(\theta)} \tag{11.16}$$

where $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$ are the estimates of the positions on the latent scale with the two observed cutoffs.

To define this standard error, we augment the log-likelihood given the observed data with two observations, one for each of the sum scores S_A and S_B. So the complete likelihood becomes $L(\delta, \theta) = L(\delta) + L_s(\theta)$, and the information matrix becomes

$$I(\delta,\theta) \approx E_\theta \left(\begin{array}{ccc|c} \nabla(\delta)\nabla(\delta)^t & \nabla(\delta)d(\theta_{SA})^t & \nabla(\delta)d(\theta_{SB})^t & \\ \nabla(\theta_{SA})\nabla(\delta)^t & \nabla(\theta_{SA})\nabla(\theta_{SA})^t & 0 & y \\ \nabla(\theta_{SB})\nabla(\delta)^t & 0 & \nabla(\theta_{SB})\nabla(\theta_{SB})^t & \end{array} \right). \quad (11.17)$$

As above, the standard error of the difference between $\hat{\theta}_{SA}$ and $\hat{\theta}_{SB}$ can be computed as the square root of the associated variance, which can be computed by pre- and post-multiplying the covariance matrix by a vector of weights, that is, $\boldsymbol{w}^t Cov(\hat{\delta}, \hat{\delta})\boldsymbol{w}$. In this case, the vector $\boldsymbol{w}$ has elements

$$w_j = \begin{cases} 1 & \text{if } j \text{ is related to } Cov(\hat{\theta}_{SA}, \hat{\theta}_{SA}) \\ -1 & \text{if } j \text{ is related to } Cov(\hat{\theta}_{SB}, \hat{\theta}_{SB}) \\ 0 & \text{if this is not the case.} \end{cases} \quad (11.18)$$

Examples will be given below.

EAP estimates and another approach to the reliability of number-correct scores.

In test theory we distinguish between global reliability and local reliability. Global reliability is related to the precision with which we can distinguish two randomly drawn students from some well-defined population, while local reliability relates to the precision given a specific test score. We discuss these two concepts in the framework of IRT in turn.

One of the ways in which global reliability can be defined is as the ratio of the true variance relative to the total variance. For the framework of IRT, consider the variance decomposition

$$var(\theta) = var[E(\theta|\boldsymbol{y})] + E[var(\theta|\boldsymbol{y})], \quad (11.19)$$

where $\boldsymbol{y}$ is an observed response pattern, $var(\theta)$ is the population variance of the latent variable, $var[E(\theta|\boldsymbol{y})]$ is the posterior variance of the expected person parameters (say, the EAP estimates of θ). So this EAP estimate is the error variance averaged over the values that can be observed weighted with their probability of their occurrence under the model. Further, $E[var(\theta|\boldsymbol{y})]$ is the expected posterior variance of the EAP estimate. Then reliability is given by the ratio

$$\rho = \frac{var[E(\theta|\boldsymbol{y})]}{var(\theta)} = 1 - \frac{E[var(\theta|\boldsymbol{y})]}{var(\theta)} \quad (11.20)$$

(See, Bechger et al. 2003). The middle expression in (11.20) is the variance of the estimates of the person parameters relative to the 'true' variance, and the right-hand expression in (11.15) is one minus the average variance of the estimates of the student parameters, say, the error variance, relative to the 'true' variance.

The generalization to number-correct scores s is straightforward. If the observations are restricted from $\boldsymbol{y}$ to s, a student's proficiency can be estimated by the EAP $E(\theta|s)$, that is, the posterior expectation of θ given s, and the precision of the estimate is given by the posterior variance $var(\theta|s)$. Then global reliability generalizes to

$$\rho_s = \frac{var[E(\theta|s)]}{var(\theta)} = \frac{\text{var}(\theta) - E[var(\theta|s)]}{var(\theta)}. \tag{11.21}$$

If the 1PLM holds, s is a sufficient statistic for θ. Therefore, it is easily verified that $E(\theta|s) \equiv E(\theta|\boldsymbol{y})$ and the expressions (11.20) and (11.21) are equivalent. In all other cases, computation of the posterior distribution involves a summation over all possible response patterns resulting in a number-correct score s, and, as already noticed above, this can be done using the recursive algorithm by Lord and Wingersky (1984).

If the 1PLM does not hold, there is variance in $E(\theta|\boldsymbol{y})$ conditional on s. This leads to the interesting question how much extra error variance is created by using s as the basis for estimating θ. That is, we are interested in the contribution of $Var(E(\theta|\boldsymbol{y})|s)$ to the total error variance, that is, to the posterior variance $Var(\theta|s)$. This contribution can be worked out by using an identity analogous to Expression (11.21), that is,

$$Var(\theta|s) = E(Var(\theta|\boldsymbol{y})|s) + Var(E(\theta|\boldsymbol{y})|s)). \tag{11.22}$$

Note that $E(Var(\theta|\boldsymbol{y})|s)$ is the squared measurement error given $\boldsymbol{y}$ averaged over the distribution of $\boldsymbol{y}$ given s, and $Var(E(\theta|\boldsymbol{y})|s))$ is the variance of the EAP estimates, also over the distribution of $\boldsymbol{y}$ given s. In the next section, examples of local reliability estimates will be given.

Examples of Reliability Estimates

In this section, two simulated examples are presented to show the kind of results that the local reliability indices presented above produce.

The first example is created by simulating 1000 response patterns on a 20-item test. The data were created with the 2PLM, with the θ-values drawn from a standard normal distribution. The 20 item parameters were the product of a set of four discrimination parameters $a = \{0.8, 0.9, 1.10, 1.20\}$ and five difficulty parameters $b = \{-1.0, -0.5, 0.0, 0.5, 1.0\}$. MML estimates (i.e., Bayes modal estimates) were computed with a standard normal distribution for the θ-values. The results are displayed in Table 11.2.

Note that the MAP estimates and the EAP estimates are very similar, as are their standard deviations displayed in the columns labeled $Sd_{MAP}(\theta|s)$ and $Sd_{EAP}(\theta|s)$. The last three columns give the variance decomposition as defined in Expression (11.22). It can be seen that $Var(E(\theta|\boldsymbol{y})|s)$ is relatively small compared to $E(Var(\theta|\boldsymbol{y})|s))$. So the potential bias in a student's proficiency estimate when using number-correct scores is much less than the inflation of the precision of the estimate. A final observation that can be made from this simulation study is that the global reliability when switching from scoring using the complete response patters to using the number-correct scores dropped from 0.788 to 0.786. So the loss in global reliability was negligible.

It is expected that if the variability of the discrimination parameters is enlarged, $Var(E(\theta|\boldsymbol{y})|s)$ increases. The reason is that if the discrimination parameters are considered known, the weighted sum score $\Sigma_i d_i a_i y_i$ is a sufficient statistic for θ. If

Table 11.2 MAP and EAP estimates and their local reliability

Score	Freq	MAP (θ)	$Sd_{MAP}(\theta\|s)$	EAP(θ)	$Sd_{EAP}(\theta\|s)$	$Var(\theta\|s)$	$Var(E(\theta\|\mathbf{y})\|s)$	$E(Var(\theta\|\mathbf{y})\|s)$
0	1	−2.24	0.59	−2.26	0.60	0.36	0.00	0.36
1	15	−1.92	0.55	−1.93	0.56	0.32	0.00	0.31
2	17	−1.64	0.52	−1.67	0.53	0.29	0.00	0.28
3	32	−1.38	0.50	−1.37	0.51	0.26	0.01	0.25
4	46	−1.14	0.48	−1.17	0.49	0.24	0.00	0.23
5	51	−0.92	0.46	−0.93	0.47	0.22	0.00	0.22
6	64	−0.71	0.45	−0.72	0.46	0.21	0.00	0.21
7	59	−0.51	0.44	−0.53	0.45	0.20	0.01	0.20
8	79	−0.32	0.44	−0.34	0.44	0.19	0.01	0.19
9	85	−0.13	0.43	−0.14	0.44	0.19	0.00	0.19
10	89	0.06	0.44	0.07	0.44	0.20	0.01	0.19
11	73	0.25	0.44	0.25	0.44	0.19	0.00	0.19
12	88	0.44	0.44	0.44	0.44	0.19	0.01	0.19
13	61	0.63	0.44	0.65	0.45	0.20	0.01	0.20
14	73	0.83	0.45	0.84	0.46	0.21	0.00	0.21
15	64	1.04	0.46	1.06	0.47	0.22	0.01	0.22
16	37	1.26	0.48	1.28	0.49	0.24	0.00	0.23
17	30	1.50	0.50	1.50	0.51	0.26	0.01	0.25
18	19	1.75	0.52	1.78	0.53	0.28	0.00	0.28
19	12	2.04	0.55	2.08	0.57	0.32	0.00	0.32
20	5	2.36	0.59	2.38	0.60	0.36	0.00	0.36

all discrimination parameters are equal to 1.0, the 2PLM becomes the 1PLM, and then the number-correct score becomes a sufficient statistic. So the more variance in the discrimination parameters, the greater the violation of the 1PLM and the depreciation of the appropriateness of the scoring rule.

To investigate this effect, the discrimination parameters of the simulation were changed to parameters $a = \{0.40, 0.60, 1.40, 1.60\}$. The results are displayed in Table 11.3. It can be seen that the standard deviations in the columns labeled $Sd_{MAP}(\theta|s)$ and $Sd_{EAP}(\theta|s)$ blew up a bit, but the effect was not very large. Further, in the column labeled $Var(E(\theta|\boldsymbol{y})|s)$ the values clearly increased, while this is less the case in the column labeled $E(Var(\theta|\boldsymbol{y})|s))$. For instance, if we consider a number-correct score 10, we observe that the initial values 0.01 and 0.19 changed to 0.04 and 0.17. The net effect was a change in $Var(\theta|s)$ from 0.20 to 0.21. So the increase in variance of θ-estimates (that is, of expectations $E(\theta|\boldsymbol{y})$) was counterbalanced by an increase of the overall precision $Var(\theta|\boldsymbol{y})$.

11.5 Simulation Study of Equating Errors

In this section, two sets of simulation studies will be presented. The first study was based on the design displayed in Panel b of Fig. 11.1, which displays a design with a link via common items. The simulation was carried out to study the effect of the size of the anchor. The second set of simulations was based on the design of Panel c of Fig. 11.1, which displays a design with common students. These simulations were carried out to study the effect of the number of students in the anchor.

The studies were carried out using the 2PLM. To create realistic data, the item parameters were sampled from the pool of item parameters used in the final tests in primary education in the Netherlands. Also the means of proficiency distributions and cutoff scores were chosen to create a realistic representation of the targeted application, that entailed equating several versions and cycles of the tests.

For the first set of simulations, two tests were simulated with 2000 students each. The proficiency parameters for the first sample of students were drawn from a standard normal distribution, while the proficiency parameters for the second sample of students were drawn from a normal distribution that was either standard normal or normal with a mean 0.5 and a variance equal to 1.0. Cutoff points were varied as $\theta_0 = -0.5$ or $\theta_0 = 0.0$. The results are displayed in Table 11.4. The first column gives the length of the two tests; the tests were of equal size. 50 items is considered realistic for a high-stakes test, tests of 20 and 10 items were simulated to investigate the effects of decreasing the test length.

The second column gives the size of the anchor. The total number of items in the design displayed in the third column follows from the length of the two tests and the size of the anchor. 100 replications were made for every one of the 24 conditions. For every replication, the item parameters were redrawn from the complete pool of all item parameters of all (five) test providers. The complete pool consisted of approximately 2000 items. The last three columns give the three equating errors

Table 11.3 MAP and EAP estimates and their local reliability when the variance of the discrimination parameter is increased

Score	Freq	MAP(θ)	$Sd_{MAP}(\theta\|s)$	EAP(θ)	$Sd_{EAP}(\theta\|s)$	$Var(\theta\|s)$	$E(Var(\theta\|\mathbf{y})\|s)$	$E(Var(\theta\|\mathbf{y})\|s)$
0	2	−2.26	0.57	−2.21	0.64	0.41	0.00	0.41
1	5	−1.96	0.54	−1.99	0.61	0.37	0.00	0.37
2	23	−1.69	0.51	−1.73	0.60	0.36	0.02	0.34
3	35	−1.44	0.49	−1.48	0.57	0.33	0.04	0.29
4	47	−1.21	0.47	−1.22	0.54	0.29	0.05	0.24
5	58	−1.00	0.46	−1.00	0.53	0.28	0.06	0.23
6	67	−0.80	0.45	−0.83	0.48	0.23	0.03	0.20
7	71	−0.60	0.44	−0.61	0.47	0.22	0.04	0.19
8	63	−0.41	0.43	−0.42	0.45	0.20	0.03	0.17
9	86	−0.23	0.43	−0.21	0.45	0.21	0.04	0.17
10	99	−0.04	0.43	−0.05	0.46	0.21	0.04	0.17
11	81	0.14	0.43	0.15	0.45	0.20	0.03	0.17
12	87	0.33	0.43	0.36	0.46	0.21	0.04	0.17
13	63	0.52	0.44	0.51	0.47	0.22	0.04	0.18
14	60	0.71	0.45	0.77	0.49	0.24	0.04	0.20
15	48	0.91	0.45	0.98	0.51	0.26	0.04	0.22
16	44	1.13	0.47	1.19	0.54	0.29	0.04	0.25
17	33	1.35	0.49	1.43	0.57	0.33	0.05	0.28
18	19	1.60	0.51	1.64	0.60	0.36	0.05	0.31
19	8	1.87	0.53	1.89	0.62	0.39	0.02	0.36
20	1	2.17	0.57	2.13	0.64	0.40	0.00	0.40

Table 11.4 Simulation of equating via common items

Number of items							
Examination	Anchor	Total	θ_0	μ_B	$Se(\bar{b}_A - \bar{b}_B)$	$Se(\hat{\mu}_A - \hat{\mu}_B)$	$Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$
50	30	70	0.00	0.00	0.050	0.010	0.441
				0.50	0.053	0.010	0.441
			−0.50	0.00	0.050	0.010	0.445
				0.50	0.052	0.010	0.446
	20	80	0.00	0.00	0.054	0.014	0.441
				0.50	0.055	0.015	0.441
			−0.50	0.00	0.055	0.015	0.447
				0.50	0.056	0.016	0.447
	10	90	0.00	0.00	0.062	0.022	0.434
				0.50	0.059	0.023	0.434
			−0.50	0.00	0.059	0.022	0.441
				0.50	0.061	0.023	0.441
20	10	30	0.00	0.00	0.053	0.018	0.651
				0.50	0.054	0.020	0.651
			−0.50	0.00	0.052	0.018	0.677
				0.00	0.054	0.018	0.651
	5	35	0.00	0.00	0.082	0.028	0.666
				0.50	0.080	0.031	0.666
			−0.50	0.00	0.086	0.029	0.682
				0.50	0.079	0.031	0.683
10	5	15	0.00	0.00	0.100	0.024	0.889
				0.50	0.086	0.026	0.889
			−0.50	0.00	0.097	0.024	0.937
				0.50	0.087	0.026	0.937

defined above. Note that $Sd(\theta)$ was always equal to 1.0, so the equating errors were equal to the analogous standard errors.

The results are generally as expected. Note first that there was always a substantial main effect of the test length for all three indices. For a test length of 50 items, decreasing the size of the anchor increased the equating errors for the average item difficulties $Se(\bar{b}_A - \bar{b}_B)$ and the proficiency means $Se(\hat{\mu}_A - \hat{\mu}_B)$. The effect on $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ was small. This pattern was sustained for a test length of 20 items, but in that case also $Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$ increased slightly when the anchor was decreased from 10 to 5. Finally, there were no marked effects of varying the position of the cutoff points and the differences between the two proficiency distributions.

The second set of simulations was based on the design of panel c of Fig. 11.1, the design with common students. The general setup of the study was analogous to the first one, with some exceptions. All samples of students were drawn from standard normal distributions and the cutoff point was always equal to $\theta_0 = 0.0$. There were three tests in the design: two tests to be equated and a test given to the linking group. As can be seen in the first column of Table 11.5, the tests to be equated had either 40 or 20 items. In the second column, it can be seen that the linking groups were either administered tests of 20, 10, or 4 items. These linking tests always comprised of an equal number of items from the two tests to be equated. The third column shows how the size of the sample of the linking group was varied. The two tests to be equated were always administered to 2000 students. In general, the results are much worse than those displayed in Table 11.4. In fact, only the combination of two tests of 40 items with a linking group of 1600 students administered a test of 20 items comes close to the results displayed in Table 11.4. Note that linking tests of 40 items with linking groups administered 4 items completely breaks down, especially the results for $Se(\tilde{\hat{\theta}}_{SA} - \hat{\theta}_{SB})$ with 100, 400 or 800 students in the linking groups become extremely poor.

11.6 Conclusion

Transparency of scoring is one of the major requirements for the acceptance of an assessment by stakeholders such as students, teachers and parents. This is probably the reason why number-correct scores are still prominent in education. The logic of such scoring is evident: the higher the number of correct responses, the higher the student's proficiency. The alternative of using the proficiency estimates emanating from an IRT model as test scores is more complicated to explain. In some settings, such as in the setting of computerized adaptive testing, it can be made acceptable that students that respond to more difficult items get a higher proficiency estimate than students with an analogous score on more easy items. However, explaining the dependence of proficiency estimates on item-discrimination parameters is more cumbersome.

A potential solution to the problem is using the 1PLM model, where all items are assumed to have the same discrimination index, and the proficiency estimate only depends on the number of correct responses to the items. However, the 1PLM seldom fits educational test data and using the 1PLM to utilize all the advantages of IRT leads to notable loss of precision. Therefore, the 2PLM and 3PLM have become the standard models for analyzing educational test data. In this chapter, a method to combine number-correct scoring with the 2PLM and 3PLM was suggested and methods for relating standards on the number-correct scale to standards on the latent IRT scale were outlined. Indices for both the global and local reliability of number-correct scores were introduced. It was shown that the error variance for number-correct scoring can be decomposed into two components. The first component is the variance of the proficiency estimates given the response patterns conditional on

Table 11.5 Simulation of equating via common students

Number of items		Number of students			
Examination	Linking group	Linking group	$Se(\overline{b}_A - \overline{b}_B)$	$Se(\hat{\mu}_A - \hat{\mu}_B)$	$Se(\hat{\theta}_{SA} - \hat{\theta}_{SB})$
40	20	100	0.092	0.098	0.505
		400	0.088	0.045	0.495
		800	0.081	0.036	0.494
		1600	0.069	0.032	0.494
40	10	100	0.468	0.179	0.574
		400	0.201	0.062	0.499
		800	0.139	0.051	0.496
		1600	0.101	0.045	0.495
40	4	100	0.416	2.043	3.310
		400	0.209	0.487	3.290
		800	0.129	0.342	2.240
		1600	0.096	0.222	0.587
20	10	100	0.128	0.118	0.702
		400	0.116	0.060	0.692
		800	0.107	0.050	0.692
		1600	0.089	0.044	0.691
20	4	100	0.618	0.167	0.724
		400	0.289	0.107	0.705
		800	0.204	0.098	0.703
		1600	0.130	0.092	0.702

number-correct scores. This component can be viewed as a measure for the bias introduced by using number-correct scores as estimates for proficiency rather than estimating the proficiency under the 2PLM or 3PLM based on a student's complete response pattern. The second component can be interpreted as the average error variance when using the number-correct score. The presented simulation studies indicate that, relative to the second component, the first component is small.

When equating two tests, say an older version and a newer version, it is not only the standard error of the proficiency estimates on the two tests which is important, but also the standard error of differences between proficiency estimates on the two tests. To obtain a realistic estimate of the standard errors of these differences, the whole covariance matrix of the estimates of all item and population parameters in the model must be taken into account. The size of these standard errors depends on the strength of the link between the two tests, that is, on the number of items and students in the design and the sizes of the overlap between, respectively, items and students. The simulation studies presented in this chapter give an indication of the standard errors of these differences for various possible designs.

The procedure for number-correct scoring was presented in the framework of unidimensional IRT models for dichotomously scored items. It can be generalized in various directions. First of all, a sum score can also be defined for a test with polytomously scored items by adding the scores on the individual items in the test. These sum scores can then be related to a unidimensional IRT model for polytomously scored items such as the generalized partial credit model (Muraki 1992), the graded response model (Samejima 1969) or the sequential model (Tutz 1990) in an manner that is analogous to the procedure presented above. Also multidimensional versions of these models (Reckase 1985) present no fundamental problems: the proficiency distributions and response probabilities introduced above just become multivariate distributions in multivariate θ parameters. For the generalized definitions of reliabilities refer to van Lier et al. (2018).

A final remark concerns the statistical framework of this chapter, which was the related Bayes modal and marginal maximum likelihood framework. In the preliminaries section of this chapter, it was already mentioned that this framework has an alternative in the framework of fully Bayesian estimation supported by Markov chain Monte Carlo computational methods (Albert 1992; Johnson and Albert 1999). Besides with dedicated samplers, the IRT models discussed here can also be estimated using general purpose samplers such as Bugs (Lunn et al. 2009) and JAGS (Plummer 2003). But details of the generalizations to other models and another computational framework remain points for further study.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17,* 251–269.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27,* 319–334.

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

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B, 39,* 1–38.

Efron, B. (1977). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273–294.

Glas, C. A. W. (2016). Maximum-likelihood estimation. In W.J. van der Linden (ed.). *Handbook of Item Response Theory: Vol. 2. Statistical tools* (pp. 197–216). Boca Raton, FL: Chapman and Hall/CRC.

Johnson, V., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NJ: Springer.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–903.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8,* 453–461.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B, 44,* 226–233.
Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28,* 3049–3067.
Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Neyman, J., & Scott, E. L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16,* 1–32.
Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria. ISSN 1609-395X.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institute.
Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.
Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph, 17,* 1–100.
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.
van Lier, H. G., Siemons, L., van der Laar, M. A. F. J., & Glas, C. A. W. (2018). Estimating optimal weights for compound scores: A multidimensional IRT approach. *Multivariate Behavioral Research.* Published Online: https://www.tandfonline.com/doi/full/10.1080/00273171.2018.1478712.