STATISTICA, anno LXXVII, n. 4, 2017

A REVIEW OF TEST EQUATING METHODS WITH A SPECIAL FOCUS ON IRT-BASED APPROACHES

Valentina Sansivieri [1]
Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italia
Marie Wiberg
Department of Statistics, USBE, Umeå University, Sweden
Mariagiulia Matteucci
Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italia

1. INTRODUCTION

In educational and psychological measurement, a test consisting of a set of items is typically administered to a sample of subjects to make inference on the latent variables underlying the response process. Latent variables are not directly observed but are rather inferred through a statistical model from the observed, directly measured, item responses. Statistical models that aim to explain observed variables in terms of latent variables are called latent variable models. Latent variable models are used in many disciplines, including psychology, economics and the social sciences. Examples of latent variables in the field of economics include quality of life and happiness; in an educational context a typical latent variable is the examinee's ability on a specific subject (e.g., mathematics). A typical situation in these fields is that different tests are used to measure the same latent variable.

Test score equating is used to compare different test scores from different test forms (Kolen and Brennan, 2014; González and Wiberg, 2017). There are at least three reasons to have multiple forms of a test (and consequently equating). The first is security. Many testing programs administer high-stakes examinations in which performance has an important impact upon the examinee and the public: conferring a license or certificate to practice a profession, permitting admittance to a college or other training program, or granting credit for an educational experience. A second and related reason for different test forms is the current movement to open testing. Many programs find it necessary or desirable to release test items to the public (Braun, 1982). When this occurs, it is not

[1] Corresponding Author. E-mail: valentina.sansivier2@unibo.it

possible to use the released items on future test forms without providing examinees an unfair advantage. A third reason for different test forms is that test content, and therefore test items, by necessity change gradually over time.

Depending on what kind of test situation we have, we can use different data collection designs and different equating methods. Two common data collection designs are the equivalent group (EG) design, which assumes that the groups of examinees to be compared are equivalent, and the non-equivalent groups with anchor test (NEAT) design, which requires an anchor test (i.e. common items) to be administered along with the test forms to the different groups. Although the NEAT design is superior in many practical settings, there are a number of large-scale assessment tests that lack anchor tests, for example, the American College Testing (ACT, 2007). A problem with using an EG design in these situations is that it might be known that the different groups who take different test forms are non-equivalent and thus the equivalent group assumption in the EG design is not fulfilled (Lyrén and Hambleton, 2011). In these situations, an option that has been shown to work well is to use the non-equivalent groups with covariates (NEC) design (Wiberg and Bränberg, 2015). The idea with the NEC design is to use information from the covariates to adjust the differences between the groups in order to obtain comparable test scores.

The traditional equating methods include mean equating, linear equating and equipercentile equating and have been developed under all the designs. Equipercentile equating is the most general among these methods and includes the first two methods (Angoff, 1971). Kernel equating (von Davier *et al.*, 2004) is a unified approach to test equating which comprises five steps. First fitting appropriate statistical models to the raw data obtained by the data collection design (pre-smoothing). Second, estimation of the scores probabilities. Third, continuization of the discrete distributions obtained at the previous step. Fourth, equating using the equipercentile method. Fifth, calculating the standard error of equating.

Item response theory (IRT) equating (Lord, 1980) is a three-step process. In the first step, item parameters are estimated; in the second step, parameter estimates are scaled to a base IRT scale using a linear transformation; in the third step, equating is conducted by using different methods, e.g. the equipercentile equating.

The aim of this work is to propose a review of test equating methods with a focus on traditional and recent IRT-based approaches. We focus on IRT equating essentially for two reasons: the possibility of using these methods in many applications and the very recent developments in this field, which filled up some gaps. In particular, we focus on the following recent works: Andersson (2016), Andersson and Wiberg (2016), Battauz (2013), Battauz (2017), He *et al.* (2015), Lee and Lee (2016), Sansivieri and Wiberg (2017), and Tao and Cao (2016).

The structure of the remaining of this paper is as follows. In Section 2 we describe the traditional methods of equating under the different possible equating designs and we introduce the kernel method of test equating briefly. Successively, in Section 3, a detailed description of IRT equating is given which includes a comparison between IRT true-score equating and IRT observed-score equating. The section ends with a presentation

of the recent trends in IRT equating. Finally, in Section 4 we describe strengths and weaknesses of the different illustrated approaches, identifying unresolved questions and, consequently, possible future research topics.

2. Traditional equating methods and Kernel equating

In this section the methods used traditionally in test equating are illustrated. These methods include equipercentile equating, linear equating and mean equating using the different data collection designs. Additionally, the kernel method of test equating is described.

2.1. Equipercentile equating

The overall idea of the equipercentile equating is that the distribution of scores has to be the same on the two equated forms. By adopting the definitions and the notation in Kolen and Brennan (2014), Form X and Form Y represent the new form and the old form, respectively, X and Y denote the random scores for Form X and Form Y, where x and y are the corresponding realizations.

When X and Y are continuous random variables, defining F as the cumulative distribution function of X in the population and G^{-1} as the inverse of G, the equipercentile equating function under the EG design is (Braun, 1982)

$$e_Y(x) = G^{-1}[F(x)]. \tag{1}$$

In the NEAT design there is a set of common items between the equated forms, which Kolen and Brennan (2014) define as V. By following the same notation adopted for Form X and Form Y, V represents the random score for V and v is the corresponding realization. Let $f(x, v)$ refer to the joint distribution of total score and common-item score, let $f(x)$ and $h(v)$ refer to the marginal distribution of scores on Form X and on the common items V, respectively, and, finally, let $f(x \mid v)$ refer to the conditional distribution of scores on Form X given a particular score v obtained on the common items V. Then, it is trivial to show that

$$f(x, v) = f(x \mid v)h(v). \tag{2}$$

If we use the frequency estimation method, we have to define the distribution for the synthetic population. This distribution is simply obtained combining the distributions of the two populations, both of them are properly weighted (Angoff, 1971)

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x) \tag{3}$$

and

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y), \tag{4}$$

where the three subscripts s, 1 and 2 represent the synthetic population, the population which received Form X, and the population administered Form Y, respectively. We indicate with f and g the distributions for Form X and Form Y, respectively, and we define w_1 and w_2 ($w_1 + w_2 = 1$) as the weights given to each population in the definition of the synthetic population.

We can calculate the marginal distributions of scores as follows

$$f_2(x) = \sum_v f_2(x, v) = \sum_v f_1(x \mid v) h_2(v) \tag{5}$$

and

$$g_1(y) = \sum_v g_1(y, v) = \sum_v g_2(y \mid v) h_1(v). \tag{6}$$

Our aim is obtaining the syntetic populations and we can simply reach this goal by substituting the Equations (5) and (6) into Equations (3) and (4) as follows

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x \mid v) h_2(v), \tag{7}$$

$$g_s(y) = w_1 \sum_v g_2(y \mid v) h_1(v) + w_2 g_2(y). \tag{8}$$

Define $F_s(x)$ and $G_s(y)$ as the cumulative distributions, respectively, of $f_s(x)$ and $g_s(y)$. To obtain the equipercentile function for the synthetic population, we need to define four other quantities: P_s, P_s^{-1}, Q_s and Q_s^{-1} which are the percentile rank function and the percentile function for Form X and Form Y, respectively. Subsequently, we can define the equipercentile function for the synthetic population as follows

$$e_{Y_s(x)} = Q_s^{-1}[P_s(x)]. \tag{9}$$

An alternative to the frequency estimation method is the chained equipercentile equating, which involves the following steps (Angoff, 1971): using examinees from Population 1, determine the equipercentile equating relationship ($e_{V1}(x)$) which let us convert scores on Form X to the common items; using examinees from Population 2, determine the equipercentile equating relationship ($e_{V1}(x)$) which let us convert scores on the common items to scores on Form Y; finally, the Form Y equipercentile equivalent of a Form X score can be calculated as follows

$$e_{Y_{(\text{chain})}} = e_{Y2}[e_{V1}(x)]. \tag{10}$$

The chained equipercentile equating has at least one drawback. This method requires that we equate a long test to a short test, but, obviously, we cannot exchange the scores obtained on two tests which contain a very different number of items. However, this method can be used when we have two different groups, because having similar populations is not assumed.

2.2. *Linear equating*

Linear equating is less strict than equipercentile equating: in fact, while equipercentile equating requires that the scores on the two forms have the same distribution, in linear equating we assume that only the means and the standard deviations of the scores on the two forms are equal. From this it is evident that linear equating is a subcase of equipercentile equating.

In the EG design, we define the population means on Form X and Form Y $\mu(X)$ and $\mu(Y)$, respectively, while the standard deviations on the same forms are defined as $\sigma(X)$ and $\sigma(Y)$, respectively. The method gives us the linear equating transformation

$$l_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + [\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)]. \tag{11}$$

For the NEAT design, to transform the scores on X to the scale of scores on Y, the following linear equation is used

$$l_{Y_s(x)} = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y), \tag{12}$$

where s represents the synthetic population, $\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X)$, and

$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2, \tag{13}$$

where the subscripts 1 and 2 refer to populations 1 and 2, respectively and similar for Y. It is very important to underline that we cannot estimate directly the quantities $\mu_2(X)$, $\sigma_2^2(X)$, $\mu_1(Y)$, and $\sigma_1^2(Y)$ because of the NEAT design. To obtain these estimates we express the parameters in terms of directly estimable parameters and, to do this, some specific statistical assumptions are made (Kolen and Brennan, 2014, pp.103–134).

In the NEC design one improves the precision by including covariates in equating. Bränberg and Wiberg (2011) proposed a method to conduct linear equating under the NEC design. By using the same notation already introduced, one can be assume that the following linear models hold in the population

$$Y = \boldsymbol{z}^T\beta_Y + \epsilon_Y, \tag{14}$$

$$X = \boldsymbol{z}^T\beta_X + \epsilon_X, \tag{15}$$

where the vectors of covariates and of coefficients are represented by $\boldsymbol{z}$, β_Y, β_X, respectively. The covariates define the mean test scores in the subpopulations which are represented by $\boldsymbol{z}^T\beta_Y$ and $\boldsymbol{z}^T\beta_X$; and, finally, let ϵ_Y and ϵ_X refer to the difference between each examinee score and the mean, with variances σ_Y^2 and σ_X^2, respectively (while means are equal to zero). The authors define the linear equating function as follows

$$Y^* = eq_Y = b_0 + b_1X, \tag{16}$$

where $b_0 = \mu_Y - \mu_X(\sigma_Y \sigma_X^{-1})$ and $b_1 = \sigma_Y \sigma_X^{-1}$. X and Y have observed score population means equal to μ_X and μ_Y, respectively, and population standard deviations equal to σ_X and σ_Y, respectively.

2.3. *Mean equating*

This method is the least strict of all the traditional methods: it only requires that the means of scores on the two forms X and Y are equal. Obviously, it is a particular case of equipercentile and linear equating and formally defined as

$$m_Y(x) = y = x - \mu(X) + \mu(Y), \tag{17}$$

in which $m_Y(x)$ indicates that we use mean equating to transform a score x on Form X to the scale of Form Y.

2.4. *Kernel equating*

Kernel equating (von Davier *et al.*, 2004) is a unified approach to test equating and comprises the following five steps:

1. *Pre-smoothing*. In this step, statistical models are fitted to the raw data obtained by the data collection design to obtain estimates of the score distributions.
2. *Estimation of the score probabilities*. In this step the Design Function is used to transform the estimated score distributions from Step 1 into the estimated score probabilities, $\hat{r}$ and $\hat{s}$, for test X and Y on the target population, T.
3. *Continuization*. This step is necessary because the score probabilities previously estimated are discrete, so we determine continuous approximations, $\hat{F}_{h_X}(x)$ and $\hat{G}_{h_Y}(y)$, to the estimated discrete cumulative distribution functions, $\hat{F}(x)$ and $\hat{G}(y)$. Here we need to choose the kernel distribution (usually normal) and the bandwidth parameters, h_X and h_Y.
4. *Equating*. In this step we estimate the equating function as follows

$$\hat{e}_{h_X h_Y}(x) = \hat{G}^{-1}_{h_Y}(\hat{F}_{h_X}(x)). \tag{18}$$

5. *Estimation of the standard error of equating*. This step is necessary to evaluate the equating transformation.

There are many recent developments in kernel equating. For example, Lee and von Davier (2011), Rijmen *et al.* (2011), Andersson *et al.* (2013) and Wiberg *et al.* (2014). The first paper focuses on the possibility of using different kinds of kernel distributions in the

continuization step. The second paper proposes a new way to test equating differences. The third work describes the new R package **kequate** to conduct kernel equating and the final paper contains a description of local kernel equating.

3. EQUATING WITH IRT

IRT equating is the statistical instrument used to compare different test scores from different forms when IRT models are used to assemble tests. Before conducting equating using IRT we need to estimate item parameters and to scale these estimates to a base IRT scale.

3.1. Item response theory models

IRT is a paradigm for the scoring of tests measuring abilities or other variables (Lord, 1980). The leading idea of the IRT is that the probability of a correct response to an item is a function of examinee and item parameters. An example of examinee parameter is general intelligence, while item parameters are its difficulty or location representing the item position on the difficulty range, discrimination representing how steeply the rate of success of examinees varies with their ability, and a pseudoguessing parameter, characterising the (lower) asymptote at which even the least able examinees will score due to guessing. IRT entails two assumptions: an examinee parameter which is constructed as a unidimensional latent trait; local independence of items, which means that examinee responses to the items are statistically independent. The item characteristic curve (ICC) for each item relates the probability of correctly answering the item to examinee ability. Various IRT models are in use that differ in the functional form of the ICC. The three-parameter logistic model (3PLM) is (Birnbaum, 1968)

$$p_{ji} = c_i + (1-c_i)(\exp(a_i(\theta_j - b_i))/(1+\exp(a_i(\theta_j - b_i)))), \tag{19}$$

where p_{ji} is the probability of a correct response for examinee j to item i, θ_j is the ability parameter for examinee j, a_i is the discrimination parameter for item i, b_i is the difficulty or location parameter for item i and c_i is the pseudoguessing parameter. If we assume that c_i is equal to zero, we obtain the two-parameter logistic model (2PLM), which is defined as follows (Birnbaum, 1968)

$$p_{ji} = \exp(a_i(\theta_j - b_i))/(1+\exp(a_i(\theta_j - b_i))). \tag{20}$$

Finally, if we also assume that all the items have the same discrimination parameter $a_i = 1$, we obtain the one-parameter logistic model (1PLM).

It is important to say that also multidimensional IRT models have been developed for tests that are intended to measure simultaneously along multiple dimensions (see e.g. Reckase, 2009, for a discussion).

3.2. *Transformations of IRT scales*

While in the EG design the parameter estimates are assumed to be on the same scale, in the NEAT design the parameter estimates are on different IRT scales, simply because the groups of examinees are not assumed to be equivalent.

Let scale I and scale J refer to three-parameter logistic IRT scales that linearly differ and θ_{Ij} and θ_{Jj} are values of the ability θ for examinee j on these two scales, respectively. If we indicate with A^* and B^* two constants, we can say that the following relationship between the ability-values for the two scales holds

$$\theta_{Jj} = A^* \theta_{Ij} + B^*. \tag{21}$$

The relationships between the item parameters can be formulated as follows

$$a_{Ji} = \frac{a_{Ii}}{A^*}, \tag{22}$$

$$b_{Ji} = A^* b_{Ii} + B^*, \tag{23}$$

$$c_{Ji} = c_{Ii}, \tag{24}$$

where a_{Ji} and a_{Ii}, b_{Ji} and b_{Ii}, c_{Ji} and c_{Ii} are the couples of discrimination, location and pseudoguessing parameters for the item i on scale J and on scale I, respectively. We have

$$A^* = \frac{\sigma(b_J)}{\sigma(b_I)}, \tag{25a}$$

$$= \frac{\mu(a_I)}{\mu(a_J)}, \tag{25b}$$

$$= \frac{\sigma(\theta(b_J))}{\sigma(\theta(b_I))}, \tag{25c}$$

$$B^* = \mu(b_J) - A^* \mu(b_I), \tag{26a}$$

$$= \mu(\theta_J) - A^* \mu(\theta_I). \tag{26b}$$

If we consider the scale I and the scale J, in the previous Equations we define the means $\mu(b_J)$, $\mu(b_I)$, $\mu(a_I)$, and $\mu(a_J)$ and the standard deviations $\sigma(b_J)$ and $\sigma(b_I)$ over one or more items having parameters that are expressed on the two considered scales; instead, the standard deviations $\sigma(\theta(b_J))$ and $\sigma(\theta(b_I))$ are defined over two or more examinees having parameters that are expressed on the two defined scales.

Marco (1977) uses the means and standard deviations of the b-parameter estimates obtained from the anchor test in place of the parameters in Equations (25a) and (26a): this method is known as the mean/sigma method.

A slightly different method called the mean/mean method was proposed by Loyd and Hoover (1980): to estimate the A^*-constant and the B^*-constant, they substitute the mean of the a-parameter estimates for the anchor test into Equation (25b) and the mean of the b-parameter estimates for the anchor test into Equation (26a), respectively. The obtained values of A^* and B^* then can be substituted into Equations (21)-(24) to calculate the rescaled parameter estimates (which are often referred to as being calibrated).

The mean/sigma method and the mean/mean method have a problem because they do not consider all of the item parameter estimates simultaneously. This fact has as a consequence that various combination of a—, b— and c— parameter estimates can produce identical ICC's over the range of ability at which most examinees score. The characteristic curve methods (Haebara, 1980; Stocking and Lord, 1983) may represent a solution to this problem. Note that, for ability scales I and J and for examinee j and item i, the probability of a correct response for examinees of a given ability is the same regardless of the scale, as follows

$$p_{ji}(\theta_{Jj};a_{Ji},b_{Ji},c_{Ji}) = p_{ji}\left(A^*\theta_{Ij}+B^*;\frac{a_{Ii}}{A^*},A^*b_{Ii}+B^*,c_{Ii}\right). \tag{27}$$

It is important to underline that if estimates are used in place of the parameters in Equation (27), then we cannot be sure that the equality will hold over all items and examinees for any A^* and B^*. The characteristic curve methods are based on this probable lack of equality.

In the method proposed by Haebara (1980), for example, the difference between the item characteristic curves is defined as the sum of the squared difference between the item characteristic curves for each item for examinees with a given ability. For a given θ_j we have

$$H\text{diff}(\theta_j) = \sum_{i:V}\left[p_{ji}(\theta_{Jj};\hat{a}_{Ji},\hat{b}_{Ji},\hat{c}_{Ji}) - p_{ji}\left(\theta_{Jj};\frac{\hat{a}_{Ii}}{A^*},A^*\hat{b}_{Ii}+B^*,\hat{c}_{Ii}\right)\right]^2. \tag{28}$$

The idea is to determine A^* and B^* minimizing the following criterion

$$H\text{crit} = \sum_j H\text{diff}(\theta_j). \tag{29}$$

Another approach is that of Stocking and Lord (1983). They used the sum, over items, the squared difference

$$SL\text{diff}(\theta_j) = \left[\sum_{i:V} p_{ji}(\theta_{Jj}; \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}) - \sum_{i:V} p_{ji}\left(\theta_{Jj}; \frac{\hat{a}_{Ii}}{A^*}, A^*\hat{b}_{Ii} + B^*, \hat{c}_{Ii}\right)\right]^2. \quad (30)$$

In this case A^* and B^* are calculated to minimize the following criterion

$$SL\text{crit} = \sum_j SL\text{diff}(\theta_j). \quad (31)$$

Ogasawara (2000) showed that the mean/mean method was more reliable than the mean/sigma method; Hanson and Béguin (2002) found that the characteristic curve methods produce more stable results than the mean/sigma and mean/mean methods.

Regardless of chosen transformation, there are two IRT equating approaches: IRT true-score equating and IRT observed-score equating which are described next.

3.3. *IRT true-score equating*

True-score equating was first introducted by Lord (1980), but here we will describe it as it was described in Kolen and Brennan (2014) which is based on the 3PLM. The true scores on Form X and on Form Y, which are equivalent to θ_j are defined respectively as

$$\tau_X(\theta_j) = \sum_{i:X} p_{ji}(\theta_j; a_i, b_i, c_i), \quad (32)$$

$$\tau_Y(\theta_j) = \sum_{i:Y} p_{ji}(\theta_j; a_i, b_i, c_i). \quad (33)$$

Equations (32) and (33) are referred to as test characteristic curves for Form X and Form Y. It is important to stress that true scores on Form X and Y are associated with a value of θ only over the following ranges

$$\sum_{i:X} c_i < \tau_X < n_X, \quad \sum_{i:Y} c_i < \tau_Y < n_Y, \quad (34)$$

because when we use the 3PLM, as θ approaches $-\infty$, p_{ji} approaches c_i instead of zero. The main assumption is that, for a given θ_j, true scores $\tau_X(\theta_j)$ and $\tau_Y(\theta_j)$ are equal. Subsequentially, defining τ_X^{-1} as the θ_j corresponding to true score τ_X, we can calculate the Form Y true score equivalent of a given true score on Form X as follows

$$\text{irt}_Y(\tau_X) = \tau_Y(\tau_X^{-1}), \quad \sum_{i:X} c_i < \tau_X < n_X. \quad (35)$$

From Equation (35) it is clear that true score equating is a process whose first step consists in specifying a true score τ_X on Form X, after it is necessary to find the θ_j that corresponds to that true score (τ_X^{-1}) and, finally, it is possible to determine the

true score on Form Y, τ_Y, that corresponds to that specific θ_j. To find τ_X^{-1} we use the Newton-Raphson method. Remembering that the range of possible true score on Form X is $\sum_{i:X} c_i < \tau_X < n_X$, we need to define a procedure for converting Form X scores outside this range. Kolen (1981) presented ad hoc procedures to handle this problem. He defined τ_X^* as a score outside the range of possible true scores, but within the range of possible observed scores. Equivalent scores are defined by the following equation

$$\begin{aligned} \mathrm{irt}_Y(\tau_X^*) &= \frac{\sum_{i:Y} c_i}{\sum_{i:X} c_i}\tau_X^*, \quad 0 \le \tau_X^* \le \sum_{i:X} c_i, \\ &= n_Y, \quad \tau_X^* = n_X, \end{aligned} \tag{36}$$

where n_X and n_Y are the number of items on Form X and on Form Y, respectively.

3.4. IRT observed-score equating

IRT observed-score equating was first introduced by Lord (1980), but we use Kolen and Brennan (2014) description of IRT observed-score equating as follows. Let θ_j refer to a specific ability level of examinees and define $f_r(x \mid \theta_j)$ as the distribution of number-correct scores over the first r items for examinees having this ability. Define $f_1(x = 0 \mid \theta_j) = (1 - p_{j1})$ and $f_1(x = 1 \mid \theta_j) = p_{j1}$ as the probabilities of earning a score of 0 and of 1 on the first item, respectively. For $r > 1$, Lord and Wingersky (1984) define the recursion formula as follows

$$\begin{aligned} f_r(x \mid \theta_j) &= f_{r-1}(x \mid \theta_j)(1 - p_{jr}), \quad x = 0, \\ &= f_{r-1}(x \mid \theta_j)(1 - p_{jr}) + f_{r-1}(x-1 \mid \theta_j)p_{jr}, \quad 0 < x < r, \\ &= f_{r-1}(x-1 \mid \theta_j)p_{jr}, \quad x = r. \end{aligned} \tag{37}$$

We underline that the recursion formula only let us calculate the observed score distribution for examinees of a given ability, so we have to accumulate the distributions obtained by using this formula to find the general observed score distribution. When the ability distribution $\psi(\theta)$ is continuous, then

$$f(x) = \int_\theta f(x \mid \theta)\psi(\theta)\, d\theta. \tag{38}$$

When the ability distribution is discrete, then

$$f(x) = \frac{1}{N}\sum_{j=1}^{N} f(x \mid \theta_j). \tag{39}$$

When we have calculated the observed score distributions for Form X and for Form Y conventional equipercentile methods are used to find score equivalents. Alternatives to the Lord-Wingersky algorithm can be found in González *et al.* (2016). Although IRT true-score equating is simple and it uses a conversion that does not depend on the distribution of ability (Kolen and Brennan, 2014, p.201), it equates true scores, which are not available in practice. IRT observed-score equating does not have this problem, because it is conducted by using observed scores.

3.5. Recent developments in IRT equating

In this section the focus is on some recent relevant developments in IRT equating. The developments regard several different areas of IRT equating.

A first area of interest is the calculation of standard errors of IRT equating. Ogasawara (2001, 2003) focused on asymptotic standard errors of IRT equating; Andersson (2016) demonstrated how to calculate the asymptotic standard errors of observed-score equating by using polytomous IRT models within the kernel equating framework. Andersson (2016) used both the EG and the NEAT (both chained equating (CE) and post-stratification equating (PSE)).

In the EG design, if we indicate with $\hat{\boldsymbol{r}}_P$ and $\hat{\boldsymbol{s}}_P$ the estimated vectors of score probabilities for the two tests X and Y, respectively, on population P, we can define the estimator of the equating function using polytomous IRT models in this design as follows

$$\hat{e}_{Y(EG)}(x;\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{s}}_P)=\hat{G}_P^{-1}(\hat{F}_P(x;\hat{\boldsymbol{r}}_P);\hat{\boldsymbol{s}}_P), \tag{40}$$

and the asymptotic variance by using the delta method (Kendall and Stuart, 1977)

$$\sigma^2_{\hat{e}_Y(x;\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{s}}_P)}=\frac{\partial e_{Y(EG)}(x;\boldsymbol{r}_P,\boldsymbol{s}_P)}{\partial \boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{s}_P)}\Sigma_{\boldsymbol{v}(\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{s}}_P)}\left[\frac{\partial e_{Y(EG)}(x;\boldsymbol{r}_P,\boldsymbol{s}_P)}{\partial \boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{s}_P)}\right]', \tag{41}$$

where $\boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{s}_P)$ is the score probability estimator and $\Sigma_{\boldsymbol{v}(\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{s}}_P)}$ its covariance matrix. The expression of the vector $\frac{\partial e_{Y(EG)}(x;\boldsymbol{r}_P,\boldsymbol{s}_P)}{\partial \boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{s}_P)}$ can be found in von Davier *et al.* (2004, p.77).

In the NEAT design we have for CE that we can indicate with $\hat{\boldsymbol{r}}_P$, $\hat{\boldsymbol{t}}_P$ the estimated vectors of score probabilities for the tests X and the anchor A on population P and with $\hat{\boldsymbol{s}}_Q$, $\hat{\boldsymbol{t}}_Q$ the estimated vectors of score probabilities for the tests Y and the anchor A on population Q, so we can define the estimator of the equating function using polytomous IRT models in this design as follows

$$\hat{e}_{Y(CE)}(x;\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{t}}_P,\hat{\boldsymbol{s}}_Q,\hat{\boldsymbol{t}}_Q)=\hat{G}_Q^{-1}(\hat{H}_Q(\hat{H}_P^{-1}(\hat{F}_P(x;\hat{\boldsymbol{r}}_P);\hat{\boldsymbol{t}}_P);\hat{\boldsymbol{t}}_Q);\hat{\boldsymbol{s}}_Q) \tag{42}$$

and the asymptotic variance is obtained by using the delta method

$$\sigma^2_{\hat{e}_{Y(CE)}(x;\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{t}}_P,\hat{\boldsymbol{s}}_Q,\hat{\boldsymbol{t}}_Q)} = \frac{\partial e_{Y(CE)}(x;\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)}{\partial \boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)} \Sigma_{\boldsymbol{v}(\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{t}}_P,\hat{\boldsymbol{s}}_Q,\hat{\boldsymbol{t}}_Q)} \left[\frac{\partial e_{Y(CE)}(x;\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)}{\partial \boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)}\right]', \quad (43)$$

where $\boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)$ is the score probability estimator and $\Sigma_{\boldsymbol{v}(\hat{\boldsymbol{r}}_P,\hat{\boldsymbol{t}}_P,\hat{\boldsymbol{s}}_Q,\hat{\boldsymbol{t}}_Q)}$ its covariance matrix. The expression of the vector $\frac{\partial e_{Y(CE)}(x;\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)}{\partial \boldsymbol{v}(\boldsymbol{r}_P,\boldsymbol{t}_P,\boldsymbol{s}_Q,\boldsymbol{t}_Q)}$ can be found in von Davier *et al.* (2004, p.82).

In the NEAT design using PSE, if we indicate with $\hat{\boldsymbol{r}}_S$ and $\hat{\boldsymbol{s}}_S$ the estimated vectors of score probabilities for the two tests X and Y, respectively, on the synthetic population S, we can define the estimator of the equating function using polytomous IRT models in this design as follows

$$\hat{e}_{Y(PSE)}(x;\hat{\boldsymbol{r}}_S,\hat{\boldsymbol{s}}_S) = \hat{G}_S^{-1}[\hat{F}_S(x;\hat{\boldsymbol{r}}_S);\hat{\boldsymbol{s}}_S] \quad (44)$$

and the asymptotic variance is calculated by using the delta method:

$$\sigma^2_{\hat{e}_{Y(PSE)}(x;\hat{\boldsymbol{r}}_S,\hat{\boldsymbol{s}}_S)} = \frac{\partial e_{Y(PSE)}(x;\boldsymbol{r}_S,\boldsymbol{s}_S)}{\partial \boldsymbol{v}(\boldsymbol{r}_S,\boldsymbol{s}_S)} \Sigma_{\boldsymbol{v}(\hat{\boldsymbol{r}}_S,\hat{\boldsymbol{s}}_S)} \left[\frac{\partial e_{Y(PSE)}(x;\boldsymbol{r}_S,\boldsymbol{s}_S)}{\partial \boldsymbol{v}(\boldsymbol{r}_S,\boldsymbol{s}_S)}\right]', \quad (45)$$

where $\boldsymbol{v}(\boldsymbol{r}_S,\boldsymbol{s}_S)$ is the score probability estimator and $\Sigma_{\boldsymbol{v}(\hat{\boldsymbol{r}}_S,\hat{\boldsymbol{s}}_S)}$ its covariance matrix. The expression of the vector $\frac{\partial e_{Y(PSE)}(x;\boldsymbol{r}_S,\boldsymbol{s}_S)}{\partial \boldsymbol{v}(\boldsymbol{r}_S,\boldsymbol{s}_S)}$ can be found in von Davier *et al.* (2004, p.77).

A second area is the development of R packages: Battauz (2015) described the R package **equateIRT** to conduct IRT equating; Chalmers (2012) described the R package **mirt** to conduct multidimensional IRT equating.

A third area of study is multidimensional IRT equating. Brossman and Lee (2013) described multidimensional IRT equating; Lee and Lee (2016) developed a bi-factor multidimensional item response theory (BF-MIRT) observed-score equating method. Lee and Lee (2016) work with a mixed-format test containing multiple-choice (MC) and free-response (FR) items and they assume that one specific factor is measured by MC format and the other specific factor is measured by FR format. The authors indicate with θ_G the general ability, with θ_M the MC-specific factor and with θ_F the FR-specific factor and they obtain the marginal observed score distribution as follows

$$f(x) = \int\int\int_{-\infty}^{\infty} f(x|\theta_G,\theta_M,\theta_F) g(\theta_G,\theta_M,\theta_F) d\theta_G d\theta_M d\theta_F, \quad (46)$$

where $f(x|\theta_G,\theta_M,\theta_F)$ and $g(\theta_G,\theta_M,\theta_F)$ are, respectively, the conditional observed score distribution and the entire trivariate theta distribution. To find the equating relationship, the marginal observed score distributions have to be calculated for both the

old form and the new form and, finally, the traditional equipercentile equating method is used.

A fourth area of interest is the development of IRT equating methods within the complex linkage plans. Battauz (2013) focused on the problem of applying IRT equating methods within the complex linkage plans framework under the NEAT design. When we have several test forms to be equated, it is necessary to choose which forms have a direct link, considering all the factors that could have a negative impact on the quality of the equating process. For example, it could be a bad idea to put too many links to the same form, because this could imply an high exposure of the items of that form and, consequently, test security could be threatened. If two forms are linked by using two or more paths, Battauz (2013) suggests calculating the average equating coefficients. This can be done by using the generalized angle bisector method (Holland and Strawderman, 2011). The author considers the two forms 0 and l and indicates with P_{0l} the set of all the possible paths between the two forms and with A_p and B_p the linking coefficient related to path p, $p \in P_{0l}$. To transform the scale of θ_0 to the scale of θ_l we use

$$\theta_l^* = \sum_{p \in P_{0l}} \omega_p \theta_l^p, \tag{47}$$

where

$$\omega = \frac{n_p (1 + A_p^2)^{-1/2}}{\sum_{b \in P_{0l}} n_b (1 + A_b^2)^{-1/2}}, \tag{48}$$

with n_p representing proper weights. The average equating coefficients are analogously defined as

$$A_{0l}^* = \sum_{p \in P_{0l}} A_p \omega_p, \tag{49}$$

and

$$B_{0l}^* = \sum_{p \in P_{0l}} B_p \omega_p. \tag{50}$$

The asymptotic variance-covariance matrix of the average equating coefficients can be obtained by using the delta method as follows

$$a\mathrm{cov}(\hat{A_{0l}^*}, \hat{B_{0l}^*})^T = \frac{\partial (\hat{A_{0l}^*}, \hat{B_{0l}^*})^T}{\partial \alpha^T} a\mathrm{cov}(\hat{\alpha}) \frac{\partial (\hat{A_{0l}^*}, \hat{B_{0l}^*})}{\partial \alpha}. \tag{51}$$

Finally, Battauz (2013) claims that a reasonable way to calculate the weights n_p is by minimizing the average variance of θ_l^*. In Battauz (2017) the focus is on methods to put the item parameter estimates on the same scale and, in particular, she extends the methods which we have described in Section 3 for complex linkage plans (see also

Battauz (2013)). Battauz (2017) also calculates the asymptotic standard errors of the equating coefficients for each method by using the delta method. For the multiple mean-geometric mean method t is the index of the form, with $t = 1, ..., T$, and a_{jt} and b_{jt} are the item discrimination and the item difficulty of item j in the scale of form t. The item parameters expressed on the scale of the base form are indicated with a_j^* (discrimination) and b_j^* (difficulty). If we denote with $\hat{\boldsymbol{a}}$ the vector containing the elements $\hat{a}_{jt}$, with $\boldsymbol{X_1}$ a design matrix composed of $T-1$ dummy variables that indicate in which form t the item has been administered and of v dummy variables that indicate which item j is considered, and with $\hat{\boldsymbol{\beta}}_1$ the vector of the regression coefficients that is composed by $\hat{\boldsymbol{A}}$ (the vector containing the T equating coefficients) and by $\hat{\boldsymbol{a}}^*$ (the vector containing the elements a_j^*), we can estimate both the equating coefficients and the item parameters as follows

$$\hat{\boldsymbol{\beta}}_1 = \exp\left[(\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \log \hat{\boldsymbol{a}}\right]. \tag{52}$$

If we denote with $\hat{\boldsymbol{b}}$ the vector containing the elements $\hat{b}_{jt}$, with $\boldsymbol{X_2}$ a design matrix composed of $T-1$ dummy variables multiplied by -1 that indicate in which form t the item has been administered and of v dummy variables that indicate which item j is considered, with $\hat{\boldsymbol{A}}_{\boldsymbol{n}}$ the product $\boldsymbol{T}\hat{\boldsymbol{A}}$ and with $\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}$ the vector of the regression coefficients that is composed by $\hat{\boldsymbol{B}}$ (the vector containing the T equating coefficients) and by $\hat{\boldsymbol{b}}^*$ (the vector containing the elements b_j^*), we can estimate both the equating coefficients and the item parameters as follows

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{2}} = \exp\left[(\boldsymbol{X}_2^T \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T \operatorname{diag}(\hat{\boldsymbol{A}}_{\boldsymbol{n}}) \hat{\boldsymbol{b}}\right]. \tag{53}$$

By using the same notation we can estimate for the multiple mean-mean method the equating coefficient A_t as follows

$$\hat{A}_t = \frac{\sum_{j \in J_t} \hat{a}_{jt}}{\sum_{j \in J_t} \frac{\sum_{s \in U_j} \hat{a}_{js}}{\sum_{s \in U_j} \hat{A}_s}}, \tag{54}$$

where U_j is the set of forms such that item j is in J_t. The equating coefficient $\hat{B}_t$ can be estimated following the multiple mean-geometric mean method. The multiple item response function method extend the Haebara method (Haebara, 1980) to the case of multiple forms. The equating coefficients can be found by minimizing the function:

$$f_{IR}^* = \sum_{t=1}^{T} \int_{-\infty}^{\infty} \sum_{j \in J_t} (P_{jt} - P_{jt}^*)^2 h(\theta) d(\theta), \tag{55}$$

where h is the density of a standard normal distribution, P_{jt} is the probability of a positive response on item j using $\hat{a}_{jt}$ and $\hat{b}_{jt}$, and P_{jt}^* is the probability of a positive response

on item j using $\hat{a}^*_{jt}$ and $\hat{b}^*_{jt}$. The multiple item response function method extend the Stocking-Lord method (Stocking and Lord, 1983) to the case of multiple forms. The equating coefficients can be found by minimizing the function

$$f^*_{TR} = \sum_{t=1}^{T} \int_{-\infty}^{\infty} \left(\sum_{j \in J_t} P_{jt} - P^*_{jt} \right)^2 h(\theta) d(\theta), \tag{56}$$

where h is the density of a standard normal distribution, P_{jt} is the probability of a positive response on item j using $\hat{a}_{jt}$ and $\hat{b}_{jt}$, and P^*_{jt} is the probability of a positive response on item j using $\hat{a}^*_{jt}$ and $\hat{b}^*_{jt}$.

A fifth area of study is the quality of the anchor in IRT equating under the NEAT design. He *et al.* (2015) stress the importance of the quality of the anchor in IRT true-score equating under the NEAT design. The presence of item outliers in the anchor, in fact, could affect in a negative way the equated scores, increasing their errors. The problem is that, if we simply eliminate the outliers from the anchor, we could compromise its representativeness. By starting from this consideration, the researchers' idea is not to eliminate the outliers from the anchor and, instead, to use robust scale transformation methods. They propose two methods to reach this aim, which minimize the loss function L defined as follows

$$L(d_{ij}) = \sum_i \sum_j w_{ij} d^2_{ij}, \tag{57}$$

where d_{ij} is defined in Equation (28), while w_{ij} is a weight and it is calculated in a different way by the two methods. The *area-weighted method* uses the weight 1 when $|e_j| \leq k$, while uses the weight $k/|e_j|$ when $|e_j| > k$ ($k = 1.345$ and e_j is the area enclosed between two item characteristic curves within $\theta = -4$ and $\theta = 4$ for item j). The *method of least absolute values* uses the weight $1/|d_{ij}|$.

A sixth area of interest is the extension of IRT equating methods to the case in which tests are constructed by using testlets. Tao and Cao (2016) extended IRT equating methods to the dichotomous testlet response theory (TRT) model. Because a testlet is a set of items based on a single theme, in a test constructed by using testlets, local item dependence (LID) could be present. LID means that a random or secondary factor affects the students' performance on some items. This has as a consequence that the probability of the response pattern on those items it is not the product of the probabilities of the single items (which is equivalent to say that local item independence does not hold). TRT true-score equating follows exactly the same steps of IRT true-score equating (see Section 3), with the only difference being that, before conducting equating, we need to calculate the marginalized item response function of the primary factor θ_1 as follows

$$P(X_i = 1|\theta_1) = \int_{\theta_{d(i)}} P(X_i = 1|\theta_1, \theta_{d(i)})\psi(\theta_{d(i)})d\theta_{d(i)}, \tag{58}$$

where $\theta_{d(i)}$ is the random factor which affects the testlet $d(i)$ and $\psi(\theta_{d(i)})$ is its density. Also the TRT observed-score equating works as the IRT observed-score equating, with the only difference that the observed score distribution is accumulated over θ_1 instead of θ

$$f(x) = \int_{\theta_1} f(x|\theta_1)\psi(\theta_1)d\theta_1. \tag{59}$$

A seventh area of interest is the development of IRT observed-score equating within different frameworks. First we focus on IRT observed-score kernel equating (Andersson and Wiberg, 2016). To conduct IRT equating in a kernel framework, it is necessary to calculate the vectors of score probabilities implied by the IRT models, because they are used to determine the continuous distributions used to conduct equating. Let $e_{Y(D)}()$ refer to the equating function for a specific design D, we can define the kernel equating function from X to Y for all score values $0, ..., k_X$ as the following vector-valued function

$$e_{Y(D)}(x) = (e_{Y(D)}(0), e_{Y(D)}(1), ..., e_{Y(D)}(k_X))'. \tag{60}$$

The design D is replaced with the various designs available: Andersson and Wiberg (2016) focus on the two NEAT designs given in Equations (9) and (10). The authors also show that, under appropriate conditions, $\sqrt{n}(\hat{e}_{Y(D)}(x) - e_{Y(D)}(x)) \sim N(0, \Sigma_{\hat{e}_{Y(D)}(x)})$ (Andersson and Wiberg, 2016, pp.51–54).

We continue by describing IRT observed-score equating with the NEC design (Sansivieri and Wiberg, 2017). Differential item functioning (DIF) occurs when the expected score given the same latent trait θ_j is different by virtue of observed characteristic (z_j) (Hulin *et al.*, 1983). The traditional IRT-DIF procedures of Lord (1980) and Raju (1988) do not let us test multiple covariates for DIF simultaneously or use continuous covariates. The IRT-C model was proposed by Tay *et al.* (2011) to overcome the limitations described above. By using the 2PLM, we can include DIF in the IRT-C model as follows

$$P(y_{ji} \mid \theta_j, z_j) = \frac{1}{1 + \exp(-[a_i\theta_j + b_i + c_i z_j + d_i z_j \theta_j])}, \tag{61}$$

where the probability of item responding depends on θ_j and also on z_j. The additional terms in Equation (61), $c_i z_j$ and $d_i z_j \theta_j$, represent the direct and interaction effects for modeling uniform (same item discriminations) and non-uniform DIF (different item discriminations), respectively.

Let $f_t(x \mid \theta_j, z_j)$ refer to the distribution of number-correct scores over the first t items for examinees of ability θ_j and covariate z_j (it was defined previously, but without covariates, in Section 3). Define $f_1(x = 0 \mid \theta_j, z_j) = (1 - p_{j1})$ and $f_1(x = 1 \mid \theta_j, z_j) = p_{j1}$ as the probabilities of earning a score of 0 and of 1 on the first item, respectively. For

$t > 1$, the recursion Formula (37) for the IRT-C Model (61) becomes (Sansivieri and Wiberg, 2017)

$$\begin{aligned} f_t(x \mid \theta_j, z_j) &= f_{t-1}(x \mid \theta_j, z_j)(1-p_{jt}), \quad x = 0, \\ &= f_{t-1}(x \mid \theta_j, z_j)(1-p_{jt}) + f_{t-1}(x-1 \mid \theta_j, z_j)p_{jt}, \quad 0 < x < t, \\ &= f_{t-1}(x-1 \mid \theta_j, z_j)p_{jt}, \quad x = t. \end{aligned} \tag{62}$$

This updated recursion formula gives the observed score distribution for examinees of a given ability and covariate. To find the observed score distribution for examinees of various abilities and with different values of covariate, the observed score distribution for examinees at each ability and value of covariate is found and then these are accumulated. When the ability distribution $\psi(\theta)$ is continuous, then

$$f(x) = \sum_{z_j} \int f(x \mid \theta, z_j)\psi(\theta)\, d\theta\, p(z_j), \tag{63}$$

where $p(z_j)$ is the distribution of z_j. To conduct observed-score equating, we have to determine observed score distributions for Form X and for Form Y and, finally, use conventional equipercentile methods to find score equivalents.

4. Some concluding remarks

Comparing the different data collection designs used in test equating, the NEC design is surely the most "incomplete"of all, probably also because it is the most recent. In fact, we have seen that only the linear equating, the kernel equating and, very recently, the IRT observed-score equating have been developed under this design, so there are still several gaps. For example the traditional equipercentile equating has not been developed yet under this design. In general, we want to underline that the most relevant aspect of the NEC design is an appropriate choice of covariates, for two main reasons. The first reason is that they should be chosen so that they can explain the differences between the groups of examinees; the second reason is that they have an impact on the probability to answer an item correctly.

Regarding the other data collection designs, the EG design is based on a very strong assumption. In many settings it is unrealistic that the groups of examinees taking the different forms are equivalent. When we have the possibility of administering common items to the examinees, the NEAT design is usually the best design. Unfortunately, sometimes we do not have access to common items and, in this case, the NEC design is a good alternative to correct for differences between the groups using covariates.

The traditional methods have been developed with the goal that, after equating, converted scores on two forms have at least some of the same score distribution characteristics in a population of examinees. Almost all methods have been developed under all

the data collection designs. For mean and linear equating, the use of sample means and standard deviations in place of the parameters typically leads to adequate equating precision, even when the sample size is small. However, equipercentile equating is often not precise enough for practical purposes because of sampling error.

The starting point of the kernel equating was the development of useful probability models for fitting the score distributions that arise in test equating. The next step was to develop equating methods that could fully exploit the log-linear models for score distributions. As kernel equating has been proposed as a unified approach to test equating one could think that it is exhaustive and totally explored by definition. Instead, there are many possibilities for new developments. For example, there are new proposals for the kernel distribution used in the continuization step.

Regarding the IRT methods, they have two main advantages. They are used in many applications and they provide an integrated psychometric framework for developing and scoring testing. The main disadvantage is that they make strong statistical assumptions, which are unlikely to hold precisely in real testing situations.

If we consider the assumptions of these models, we know that the unidimensional IRT models assume that the test forms are unidimensional and that the relationship between ability and the probability of correct response follows a specified form. These requirements are difficult to justify for many educational achievement tests and this is in contrast with a general recommendation in test equating: the equating studies should be designed to minimize the effects of violations of assumptions (Kolen and Brennan, 2014).

Considering the new methods illustrated in the last subsection of the previous section, we can do several observations.

Andersson (2016) provided very accurate approximations for the standard errors of observed-score equating with polytomous IRT models within the kernel equating framework, as we can see by the results of the simulation study. The main limitations of the work are that only twenty-five items have been used because the computational time would have been too long if too many items are used and that the Haebara and Stocking-Lord's methods have not been used. Future studies should be conducted to fill these gaps.

Lee and Lee (2016) obtained good results in their simulation study, where they showed that the bi-factor MIRT observed-score equating is more accurated than the unidimensional IRT observed-score equating when the correlation between the MC and FR factor is low, while the results of the two methods are very similar when the correlation is high. This is a logical and expected result. In fact, when the correlation between the two specific factors is low, this means that two specific factors are really present in our test and better results from the bi-factor MIRT observed-score equating are expected. On the opposite side, if there is a strong correlation between the two specific factors, this means that only one factor is present in the test and, consequently, we expect similar results from the two methods. The simulation study as an evident limitation: only ten replications have been conducted. A possible interesting extension of this work could be by using a test in which each specific factor is correlated with a different content area.

The equating in complex linkage plans (Battauz, 2013) let us gain in efficiency essentially because we use the weights. Battauz (2013) simulation study showed, in fact, that unweighted average coefficients have an higher standard error than the single equating coefficients. The method offers a strong advantage compared to the concurrent calibration: the possibility of calculating the standard errors of both the equating coefficients and the scores. Finally, by using the multiple linkage, it is also possible to control for seasonality effects. Battauz (2017) showed that all her proposed methods performed well in a simulation study and that their results are similar between them and they are similar to the results obtained in Battauz (2013). The work has however three main limitations. First a real example has not been conducted. Second, if the number of common items between test forms increases, then the computational time also increases and, finally, the mean-sigma method has not been extended. Future studies, obviously, should extend also the mean-sigma method.

About the issue of the possible presence of outliers in the common items set (He *et al.*, 2015), the two methods discussed here work well when outliers are in the anchor, but they work less well when no outliers are in the anchor. Another important consideration is that when the proportion of common items is big with respect to the total, the possible presence of outliers is negligible. Finally, as He *et al.* (2015) assume the presence of one common item outlier, future studies with multiple outliers should be conducted.

The TRT equating simulation study (Tao and Cao, 2016) shows that the observed-score equating outperforms with respect to the true-score equating, in particular when the LID is high, even if the pseudo-guessing parameter estimation badly affects the true-score equating results. For this reason,Tao and Cao (2016) stress the importance of repeating the study by using a two or one parameter TRT model as well as they underline the importance of reconducting the study by using a more complex testlet design (they use only one testlet).

It is evident that also the IRT true-score equating with the NEC design and the IRT true-score kernel equating could be developed. Regarding the IRT observed-score equating with the NEC design (Sansivieri and Wiberg, 2017), other limits are evident. Only dichotomous items have been used and only the existence of one latent dimension has been assumed. Instead, in the IRT observed-score kernel equating the inclusion of covariates could improve the accuracy of the estimates. The main advantage of the IRT observed-score equating with the NEC design is the possibility of improving results simply using covariates about the examinees, which are often available without other costs. Concerning the IRT observed-score kernel equating (Andersson and Wiberg, 2016), the main advantage is that we can obtain, by using a continuous and differentiable kernel, an equating function without points of non-differentiability which is an issue when we use linear interpolation.

Summing up, although there are many new equating methods which have emerged during the past years, there are still test settings which could benefit of using improved methods.

References

ACT (2007). *ACT Technical Manual*. Act Inc., Iowa City.

B. Andersson (2016). *Asymptotic standard errors of observed-score equating with polytomous IRT models*. Journal of Educational Measurement, 53, no. 4, pp. 459–477.

B. Andersson, K. Bränberg, M. Wiberg (2013). *Performing the kernel method of test equating with the package kequate*. Journal of Statistical Software, 55, no. 6, pp. 1–25.

B. Andersson, M. Wiberg (2016). *Item response theory observed-score kernel equating*. Psychometrika, 82, no. 1, pp. 48–66.

W. Angoff (1971). *Scales, norms and equivalent scores*. In R. L. Thorndike (ed.), *Educational Measurement*, American Council on Educations, Washington DC, pp. 508–600.

M. Battauz (2013). *IRT test equating in complex linkage plans*. Psychometrika, 78, no. 3, pp. 464–480.

M. Battauz (2015). *equateIRT: an R package for IRT test equating*. Journal of Statistical Software, 68, no. 7, pp. 1–22.

M. Battauz (2017). *Multiple equating of separate IRT calibrations*. Psychometrika, 82, no. 3, pp. 610–636.

A. Birnbaum (1968). *Some latent trait models and their use in inferring an examinee's ability*. In F. Lord, M. Novick (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading (Mass.), pp. 397–479.

K. Bränberg, M. Wiberg (2011). *Observed score linear equating with covariates*. Journal of Educational Measurement, 48, no. 4, pp. 419–440.

H. I. Braun (1982). *Observed score test equating: a mathematical analysis of some ETS equating procedures*. In P. Holland, D. Rubin (eds.), *Test Equating*, Academic Press, New York, pp. 9–49.

B. G. Brossman, W.-C. Lee (2013). *Observed score and true score equating procedures for multidimensional item response theory*. Applied Psychological Measurement, 37, no. 6, pp. 460–481.

R. Chalmers (2012). *Mirt: A multidimensional item response theory package for the R environment*. Journal of Statistical Software, 48, no. 6, pp. 1–29.

J. González, M. Wiberg (2017). *Applying Test Equating Methods Using R*. Springer, New York.

J. GONZÁLEZ, M. WIBERG, VON DAVIER A.A. (2016). *A note on the Poisson's binomial distribution in item response theory*. Applied Psychological Measurement, 40, no. 4, pp. 302–310.

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

B. HANSON, A. BÉGUIN (2002). *Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design*. Applied Psychological Measurement, 26, no. 1, pp. 3–24.

Y. HE, Z. CUI, S. OSTERLIND (2015). *New robust scale transformation methods in the presence of outlying common items*. Applied Psychological Measurement, 39, no. 8, pp. 613–626.

P. HOLLAND, W. STRAWDERMAN (2011). *How to average equating functions if you must*. In A. VON DAVIER (ed.), *Statistical Models for Test Equating, Scaling, and Linking*, Springer, New York, pp. 89–107.

C. L. HULIN, F. DRASGOW, C. K. PARSONS (1983). *Item Response Theory: Application to Psychological Measurement*. Dorsey Pr, Homewood, IL.

M. KENDALL, A. STUART (1977). *The Advanced Theory of Statistics*. Macmillan, New York.

M. KOLEN (1981). *Comparison of traditional and item response theory methods for equating tests*. Journal of Educational Measurement, 18, pp. 1–11.

M. KOLEN, R. BRENNAN (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Springer-Verlag, New York, 3rd ed.

G. LEE, W. LEE (2016). *Bi-factor MIRT observed-score equating for mixed-format tests*. Applied Measurement in Education, 29, no. 3, pp. 224–241.

Y.-H. LEE, A. VON DAVIER (2011). *Equating through alternative kernels*. In A. VON DAVIER (ed.), *Statistical Models for Test Equating, Scaling, and Linking*, Springer, New York, pp. 159–173.

F. M. LORD (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ.

F. M. LORD, M. S. WINGERSKY (1984). *Comparison of IRT true-score and equipercentile observed-score equatings*. Applied Psychological Measurement, 8, pp. 452–461.

B. H. LOYD, H. D. HOOVER (1980). *Vertical equating using the Rasch model*. Journal of Educational Measurement, 17, pp. 179–193.

P.-E. LYRÉN, R. K. HAMBLETON (2011). *Consequences of violated equating assumptions under the equivalent groups design*. International Journal of Testing, 11, no. 4, pp. 308–323.

G. L. MARCO (1977). *Item characteristic curve solutions to three intractable testing problems*. Journal of Educational Measurement, 14, pp. 139–160.

H. OGASAWARA (2000). *Asymptotic standard errors of IRT equating coefficients using moments*. Economic Review, Otaru University of Commerce, 51, no. 1, pp. 1–23.

H. OGASAWARA (2001). *Standard errors of item response theory equating/linking by response function methods*. Applied Psychological Measurement, 25, pp. 53–67.

H. OGASAWARA (2003). *Asymptotic standard errors of IRT observed-score equating methods*. Psychometrika, 68, pp. 193–211.

N. S. RAJU (1988). *The area between two item characteristic curves*. Psychometrika, 53, pp. 495–502.

M. RECKASE (2009). *Multidimensional Item Response Theory*. Springer, New York.

F. RIJMEN, Y. QU, A. VON DAVIER (2011). *Hypothesis testing of equating differences in the kernel equating framework*. In A. VON DAVIER (ed.), *Statistical Models for Test Equating, Scaling, and Linking*, Springer, New York, pp. 317–326.

V. SANSIVIERI, M. WIBERG (2017). *Item response theory equating with the non-equivalent groups with covariates design*. In L. A. VAN DER ARK ET AL. (ed.), *Quantitative Psychology. IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol 196*, Springer, Cham, pp. 275–285.

M. STOCKING, F. LORD (1983). *Developing a common metric in item response theory*. Applied Psychological Measurement, 7, pp. 201–210.

W. TAO, Y. CAO (2016). *An extension of IRT-based equating to the dichotomous testlet response theory model*. Applied Measurement in Education, 29, no. 2, pp. 108–121.

L. TAY, D. NEWMAN, J. VERMUNT (2011). *Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement Equivalence*. Organizational Research Methods, 1, no. 14, pp. 147–176.

A. A. VON DAVIER, P. W. HOLLAND, D. T. THAYER (2004). *The Kernel Method of Test Equating*. Springer-Verlag, New York.

M. WIBERG, K. BRÄNBERG (2015). *Kernel equating under the non-equivalent groups with covariates design*. Applied Psychological Measurement, 39, no. 5, pp. 1–13.

M. WIBERG, W. VAN DER LINDEN, A. VON DAVIER (2014). *Local kernel observed-score equating*. Journal of Educational Measurement, 51, no. 1, pp. 57–74.

SUMMARY

The overall aim of this work is to review test equating methods with a particularly detailed description of item response theory (IRT) equating. Test score equating is used to compare different test scores from different test forms. Several methods have been developed to conduct equating: traditional methods, kernel method, and IRT equating. We synthetically explain the traditional equating methods which include mean equating, linear equating and equipercentile equating and which have been developed under all the possible data collection designs. We also briefly describe the idea of the kernel method: this is a unified approach to test equating for which recent interesting developments have been proposed. Then we focus on IRT equating, by describing old and new methods: in particular, we define IRT observed-score kernel equating and IRT observed-score equating using covariates, as well as other recent proposals in this field. We conclude the review by describing strengths and weaknesses of the different discussed approaches and by identifying future research topics.

Keywords: Test equating; IRT test equating; Item response theory.