Local Equating of Cognitively Diagnostic Modeled Observed Scores

Applied Psychological Measurement
2015, Vol. 39(1) 44–61
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614542427
apm.sagepub.com

Tao Xin¹ and Jiahui Zhang²

Abstract

A local equating method of cognitively diagnostic modeled observed score was introduced to address the problem of equating without anchor test as well as the satisfaction of equating criteria. The criterion of test construct was redefined on attribute level. Then, a true family of equating transformations given attribute mastery pattern was derived from the criterion of equity, which is the main concern of local equating. The results of simulation studies showed a substantial bias for the traditional equipercentile equating under the equivalent groups design but much smaller bias for the local method. The proposed method also showed satisfactory performances under the nonequivalent groups design even without anchor test. The local method appeared to be quite robust with respect to changes in the sample size, but its biases became larger with increasing model misfit.

Keywords

local equating, nonequivalent groups, test construct, cognitive diagnostic model, equating bias, equipercentile equating

Resulting from the use of test forms, equating has become an indispensable process in high-stakes testing, transforming the observed scores of a new test form onto the scale of an older one, so that the scores of different test forms can be used interchangeably (Holland & Dorans, 2006). Considering the existence of measurement error, Levine (1955) introduced classical test theory (CTT), and Lord (1980) applied item response theory (IRT) to observed-score equating. This study paralleled these studies in applying cognitive diagnostic models (CDMs) to equating, but with a special purpose of addressing the problem of equating without anchor test.

In addition to measurement theory and models, five criteria or requirements have been introduced to equating, though some of them being criticized as vague and unpractical (Dorans & Holland, 2000; Holland & Dorans, 2006). Hanson (1991) and van der Linden (2000, 2006a, 2006b, 2010) are among the limited efforts to explore equating methods that satisfy the criterion of equity. This study adopts the framework of van der Linden's (2000, 2006a, 2006b, 2010) local equating. In the following sections, the criteria of equating are discussed with a focus on

¹Beijing Normal University, China

²Michigan State University, East Lansing, USA

Corresponding Author:

Tao Xin, Professor, Institute of Developmental Psychology, Beijing Normal University, 19 Xijiekou Wai Street, Beijing 100875, China.

Email: xintao@bnu.edu.cn

equity and the same construct. Then, the theorem and studies of local equating are reviewed. Finally, the criteria of equating were redefined under CDMs.

The Criteria of Equating

Traditional equating in real testing program involves a large amount of decision making, which is not fully justified by theory (Brennan, 2001). Without a sound theoretical foundation, equating procedures like equipercentile equating of observed scores are simply adaptations of scaling methods. It would be unreasonable then to claim that these so-called equating procedures could achieve a goal that scaling procedures fail to achieve. The introduction of equating criteria is an attempt to bring theoretical foundation to equating.

However, criteria of equating have been criticized as vague and unpractical (Dorans & Holland, 2000). Arguments mainly concentrate on two of them: the criteria of equity and the same construct. In contrast, the definitions of the other three criteria are more straightforward and easier to understand. For example, there are indexes to check the criterion of population invariance (Dorans & Holland, 2000) and standard statistics to compute reliability (Haertel, 2006). This study focuses on the criteria of the same construct and equity which have not received adequate attention in previous literature.

The Criterion of Same Construct

It is common sense that tests measuring different constructs should never be equated. As test forms are sets of test questions built according to certain content and statistical test specifications (Kolen & Brennan, 2004), it is intuitive to judge whether the criterion of the same construct is met by checking their test blueprint (Dorans & Holland, 2000). This requirement can also be checked by inspection of the content and wording of the test questions (Dorans & Holland, 2000). Besides those general guidelines, in CTT, correlation of scores on two tests is used as an indicator of the same construct measured (Dorans & Holland, 2000). In unidimensional IRT, it is required that the same unidimensional latent trait is measured by the two tests. Therefore, a statistical test of the simultaneous fit of an IRT model to the two test forms could serve as a check on this requirement (van der Linden, 2006b). However, either the correlation in CTT or the model fit index in IRT is limited in that they could do nothing beyond supporting the unidimensional assumption and thus provide no information regarding the specific construct of the test forms to be equated.

The Criterion of Equity

The criterion of equity, proposed by Lord (1980) out of concern about fairness of equating, requires that for each examinee in the population, the conditional distribution of the equated scores on the new form given true score should be identical to that of the scores on the old form given true score (Lord, 1980). It has been well established that the goal of equating is to allow the equated scores from different test forms to be used interchangeably for any purpose (Holland & Dorans, 2006), which distinguishes equating from scaling and predicting. It is obvious that the criterion of equity is closely associated with this goal, because should the two distributions differ from each other, using the scores interchangeably would inevitably result in inequity. According to Lord, the criterion of equity implies the criterion of reliability in that between two tests with different reliability, examinees of high ability would tend to take the more reliable one while low-ability examinees would prefer the less reliable one. The criterion of equity also implies the criterion of the same construct because examinees naturally tend to

choose a test on which they think they would score higher if two tests examine different constructs (Lord, 1980). The criterion of equity is also a sufficient condition of population invariance as the definition of equity is conditional on true ability, so if the criterion of equity holds for a population P , then it will hold for any subpopulation in P (van der Linden, 2010).

The criterion of equity was seen by some researchers as so strict that hinders the development of equating methods (Dorans & Holland, 2000; Kolen & Brennan, 2004). Their reasoning is partly based on the fact that hardly any existing method could satisfy this criterion (Brennan, 2001). Nevertheless, it should be noted that in those methods a single equating transformation is used for the chosen population. Seemingly impossible to be satisfied, the criterion of equity is often ignored in most of the observed-score equating methods; or a compromised approach is adopted by introducing a less restricted version of Lord's equity criterion, for example, first-order equity, which requires examinees with a given true score have the same mean converted score on the new form as they have on the old form (Kolen & Brennan, 2004).

The Framework of Local Equating

Local equating has been introduced with concerns about the failure of the existing observed-score equating framework to satisfy the criteria of equating, especially the criteria of equity and population invariance (van der Linden, 2000, 2006a, 2006b, 2010). An error-free or true equating transformation is defined based on equity criterion, because this criterion summarizes the goal of observed-score equating (van der Linden & Wiberg, 2010). It is assumed that both test forms examine a common latent ability θ . A family of true equating transformations instead of a single transformation was derived for a certain population:

$$\varphi_{\theta}(x) = F_{Y|\theta}^{-1}(F_{X|\theta}(x)), \theta \in R, \quad (1)$$

where $F_{Y|\theta}$ is the conditional cumulative distribution of the observed score given latent ability of θ on Form Y; $F_{X|\theta}$ is the conditional cumulative distribution of the observed score given latent ability of θ on Form X. The form of Equation 1 is similar to the expression of traditional equipercentile equating except for the conditional part; however, they differ completely in their deducing processes.

The fact that a family of transformations can meet the criterion of equity but a single transformation cannot forces researchers to reconsider the implicit assumption of using a single transformation to equate. It can be proved that Equation 1 also meets the criterion of symmetry and population invariance (van der Linden, 2010). The true equating transformation family holds for any subpopulation of P ; thus, there would not be urgent necessity of randomly sampling of test takers (van der Linden, 2010). Based on the definition of true equating transformations, equating error can be defined for any equating method, written as a function of differences between score distributions or differences between equating scores (van der Linden, 2000).

In applying local methods, the biggest challenge is to find a proxy of the unknown θ for each test taker, which indicates the test taker's membership in the equating family. In other words, there has to be a way to classify the test takers into each group pertaining to their proficiency levels examined by the test. van der Linden (2010) claimed that "even a rough estimate or a simple classification of the abilities may be better than combining them into an assumed population before conducting the equating" (p. 215), and this claim was supported by the simulation results of van der Linden and Wiberg (2010) in which the observed score of anchor test was used as the proxy of θ .

Simulation studies (van der Linden, 2006b) have revealed that the local equating method using point estimates of θ , similar to the method using the full posterior distribution of θ , has

very small equating biases; in contrast, the traditional equipercentile equating method has serious biases. In van der Linden and Wiberg (2010), two traditional methods (i.e., chain equating [CE] and poststratification equating [PSE]) show substantive biases while the local method's biases are much smaller. In addition, the local methods are barely affected by estimation errors (van der Linden, 2006b), and more robust when the difficulty and accuracy of tests change, and also sensitive to declination of anchor test's accuracy (van der Linden & Wiberg, 2010). Supported by optimistic simulation results, the introduction of local equating provided a possible solution to the long-lasting problem of failing to meet the equity requirement.

Formulation of Problem

There are occasions when test material can never be used more than once for test security reasons, where the nonequivalent anchor test (NEAT) design could not apply. Take the College Entrance Examinations in China as an example, the items of which are released soon after they are administered. Recently, there has been increasing appeal for multiple administrations per year of this test, which means equating needs to be conducted between different test forms. In traditional equating methods, without anchor tests, the only explicit information that may be available from examinees taking different test forms is noncognitive demographic information, which might be used as the anchor in equating. However, little is known about the effect of equating using demographic information, which is very likely to be questioned about its equity (Holland & Dorans, 2006). The local equating methods based on IRT are more flexible than traditional methods in that they do not depend much on random sampling. However, as the ability and item estimates are needed in equating, there is a requirement of "connectedness" (van der Linden, 2010), so that parameters from different forms can be calibrated on a common metric. Therefore, local methods based on IRT cannot solve the equating problem without anchor test either. Therefore, no existing method could address the equating problem without anchor test adequately, posing new challenges to equating researchers and practitioners.

Problems remain even when anchor test can be used. First of all, equating on the NEAT design requires the making of assumptions on missing data, so that the procedures can be interpreted as bona fide equating methods (Holland & Dorans, 2006). The two types of assumptions, the PSE type and the CE type, which are very strong and untestable assumptions, indicate population invariance of the relationship among X , Y , and A . For an equivalent groups (EG) design, both assumptions hold, and equating methods based on the two assumptions give identical results. However, it should be noted that when the two populations differ (which is exactly the case where the NEAT design is employed), these assumptions are not necessarily valid, and more importantly, there is no way to test them. Second, there are problems associated with both external and internal anchor items. The former can be easily identified by test takers in some cases (e.g., Internet-based Test of English as a Foreign Language [TOEFL] in China), so the quality of which might be influenced by examinees' motives (Holland & Dorans, 2006). The latter could avoid such motive issues, but is influenced by context effects. It has been well established that the context of an item could affect its psychometric properties substantially (e.g., Brennan, 1992; Harris & Gao, 2003). In addition, both kinds of anchor tests may subject to the problem of parameter drift. If reliable equating can be done without anchor test, then those deficiencies of anchor test designs would not be an issue.

As discussed earlier, the criterion of the same construct is very important but not clearly defined yet. In CTT and unidimensional IRT, the construct of tests is simplified to the true score or the latent variable, and correspondingly the check of the same construct requirement is nothing more than model-data fit or correlation. However, a unidimensional latent variable is far from enough to guarantee the equatability of the test forms. This study is an attempt to shed

some light on these problems by applying CDMs to local equating and facilitating the definition of test construct with the unique features of diagnostic measurement.

CDM-Based Local Equating

Local equating of cognitively diagnostic modeled observed scores is introduced in this section to address this problem of equating without anchor test. The new local method is similar to IRT model-based local equating method, in that the family of true equating transformations is also derived from the criterion of equity. However, the fundamental difference between the new method and the old one lies in their ability to equate nonequivalent groups (NEG) of examinees without anchor test. Besides, applying CDMs to local equating facilitates the definition and evaluation of the same construct criterion.

Notations and Models

Throughout the article, let Y be the original test form and X a new form to be equated to Y . The number-correct total scores on the two tests are denoted as Y and X , respectively, and y and x are their realizations.

CDMs refer to latent class models that employ item response functions (IRFs) to assess examinees' mastery or non-mastery on a set of skills called attributes with an explicit Q-matrix presenting attribute assignment to each item (Roussos, Templin, & Henson, 2007). The Q-matrices of the two test forms are denoted as Q_X and Q_Y , respectively, the J rows of which represent the J items and the K columns of which correspond to the K attributes (DiBello, Roussos, & Stout, 2007; Tatsuoka, 1995). In this study, elements in the Q-matrix are binary: $q_{jk} = 1$ if getting a correct response to item j without guessing requires the mastery of attribute k , and $q_{jk} = 0$ otherwise. Accordingly, the mastery or non-mastery of each attribute represents examinees' latent ability levels in the form of binary attribute vectors (e.g., DiBello, Stout, & Roussos, 1995; Tatsuoka, 1983), denoted as attribute mastery pattern (AMP) here. Let α refer to an AMP and Ω as the set of all AMPs.

Following the notations of Leighton, Gierl, and Hunka (2004) in their attribute hierarchy method (AHM), two matrices of order (k, k) (where k is the number of attributes) to describe the relationships among attributes are used: the adjacency matrix (A) and the reachability matrix (R). The A -matrix indicates the direct relationships among attributes: An element of the value of 1 in the position (j, k) means that attribute j is prerequisite to attribute k and an element of the value of 0 indicates otherwise. The R -matrix represents both the direct and indirect relationships between pairwise attributes. The relationship between R and A is straightforward, and R -matrix can be calculated from the A -matrix (Leighton et al., 2004). The A -matrix and the R -matrix of the two forms are denoted as A_X , A_Y , R_X , and R_Y , respectively. Ideal response patterns (IRPs), following the notation of Tatsuoka (1995), can be generated from the Q-matrix and the A-matrix.

Two CDMs used in this study are introduced as follows. Examinee responses were calibrated using the deterministic inputs, noisy "and" gate cognitive diagnosis (DINA) model (Junker & Sijtsma, 2001; Macready & Dayton, 1977) because of its computational ease. A latent response η_{ij} is defined as $\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{jk}}$, in which K is the number of attributes, q_{jk} is the element of the Q-matrix at (j, k) , and α_{ik} is the k th element of the vector α_i . The latent response η_{ij} has two values: $\eta_{ij} = 1$ indicates that examinee i masters all the attributes required for item j , and $\eta_{ij} = 0$ indicates that examinee i lacks at least one of the attributes required for item j . The probability of a correct response is also affected by slipping or guessing, characterized by two-item

parameters: $g_j = P(X_{ij} = 1 | \eta_{ij} = 0)$ and $s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$. The IRF of the DINA model can be written as

$$P(X_{ij} | \alpha_i) = (1 - s_j) g_j^{1 - \eta_{ij}}. \quad (2)$$

To obtain populations with different attribute distributions, data were simulated under the higher order DINA (HO-DINA) model (de la Torre & Douglas, 2004). The relationship between the attributes and higher order proficiency can be expressed using the latent logistic regression model:

$$P(\alpha_i | \theta_i) = \prod_{k=1}^K \left\{ \frac{\exp[1.7\lambda_1(\theta_i - \lambda_{0k})]}{1 + \exp[1.7\lambda_1(\theta_i - \lambda_{0k})]} \right\}, \quad (3)$$

where λ_1 and λ_{0k} are the latent discrimination and difficulty parameters, respectively. Attributes with higher λ_{0k} are deemed more difficult to master. The probability of correct response given α for the HO-DINA model is the same as Equation 2. Although the DINA and HO-DINA models are selected for this study, it should be noted that other CDMs using the Q-matrix and IRFs could also be employed in the proposed method, such as DINA model (Haertel, 1989, 1990), Noisy Input Deterministic “and” Gate (NIDA) model (Junker & Sijtsma, 2001), and Reparameterized Unified Model (DiBello et al., 1995; Hartz, 2002).

Criteria of Equating

The same construct criterion. The criterion of the same construct is a prerequisite to check as early as the stage of test development. As Embretson (1983) suggested, cognitive diagnosis can enhance psychometric practice by casting light on the construct representation of a test. In contrast to CTT and unidimensional IRT, CDMs define the construct of tests on a more refined attribute level. To be specific, they describe the knowledge/skills required by each test item as a set of attributes with their hierarchy in a much more detailed manner. The hierarchy of attributes denoted by the A-matrix (or the R-matrix) can be regarded as a cognitive model (Leighton et al., 2004) or the “anatomical structure” of the test construct.

On attribute level, the test construct is no longer as vague as a unidimensional latent variable, but consists of several components, including (a) a set of well-defined attributes, (b) the hierarchy of attributes (in the form of the A-matrix or the R-matrix), and (c) the Q-matrix that relates attributes to items. Take a primary school math test as an example. In CTT and IRT, the construct of the test would be defined as basic math ability, and the blueprint used in the test development process may provide additional supporting evidence. Taking the approach of cognitive diagnosis, the test blueprint is in some sense quantified in the form of the Q-matrix and the A-matrix (or the R-matrix), and the general idea of the construct of the math ability is broken down into several knowledge points and mathematical skills (e.g., integer, fraction, plane geometry, simple operation, complex operation, application, etc.) as well as their hierarchy (e.g., the knowledge of integer is a prerequisite to the mastery of fraction).

Two test forms to be equated should involve the same set of attributes with identical hierarchy. Besides the same A-matrix, it would be ideal if two test forms have the same Q-matrix, which is close to the notion of “parallel test forms” but not common in practice. More often than not, the Q-matrices of test forms differ slightly from each other. Therefore, there are two kinds of situations in which the same construct criterion might be satisfied: (a) $A_X = A_Y$ (or $R_X = R_Y$) and $Q_X = Q_Y$; (b) $A_X = A_Y$ (or $R_X = R_Y$), but $Q_X \neq Q_Y$. The same A-matrix serves as a “bridge” connecting two test forms because with the same A-matrix, the two test forms share

the same set of AMPs. In cognitive diagnostic assessments, if attributes are well defined, it is assumed that the AMPs have the same interpretation between test forms (Tatsuoka, 2009). Therefore, in theory, it is assumed that a certain AMP (e.g., $\alpha = (1, 1, 0)$) indicates the same class of examinees taking either Form X or Form Y. The common A-matrix and consequently common AMPs between the two test forms provide the theoretical probability of equating without anchor test.

It should be noted that the two Q-matrices should not differ too much to satisfy the same construct criterion. For example, it is likely that a test for higher grade students involves the same attributes as another test for a lower grade population, and they also have the same A-matrix; however, the focus of measurement can be very different. A six-grade test may lay stress on complex operations (e.g., those involving a combination of simple operations), while the same attribute also constitutes a smaller proportion in a test for four graders. In this case, these two test forms with the same A-matrix but quite distinct Q-matrices should not be equated because the test constructs are clearly different. To approach the same construct criterion, Q-matrices of two tests to be equated should be at least similar to some degree. Similar Q-matrices could be defined in various ways. In this study, it is required that test forms to be equated should have the same item types (in terms of the number of attributes required by an item) and the same number of items in each type.

To check the preceding requirements, the A-matrix and the Q-matrix of each test form need to be identified. Again, there are two kinds of situations. (a) If the Q-matrices are constructed after the R-matrix and the A-matrix are identified (Leighton et al., 2004), the corresponding matrices are readily available and can be compared directly between test forms. (b) If the Q-matrices are obtained by post hoc analysis of existing test items, the A-matrix has to be calculated from the Q-matrix, which complicates things when the two Q-matrices differ (e.g., Tatsuoka, 1983, 2009). When $Q_X = Q_Y$, the same A-matrix would always be obtained from them. When $Q_X \neq Q_Y$, however, the two Q-matrices must be “sufficient” to get the same A-matrix. Sufficient Q-matrix is referred to as a Q-matrix with the R-matrix as its submatrix (Tatsuoka, 2009). It has been proved that the same R-matrix (and consequently the same A-matrix) can be obtained from two different sufficient Q-matrices (Li, 2009). Therefore, another requirement to the same construct criterion is added: The two different Q-matrices should be sufficient ones.

To conclude, the necessary and sufficient condition of the same construct criterion on the attribute level would be (a) the same Q-matrix or (b) sufficient Q-matrices that are similar to each other. Two test forms satisfying the same construct criterion have the following properties: (a) They share the same set of AMPs, and (b) each AMP corresponds to a unique IRP from each of the two test forms (Ding, Luo, Cai, Lin, & Wang, 2008). For example, two test forms X and Y both measure three attributes A1, A2, and A3 with no hierarchy. The Q-matrices are shown below, both with R-matrix (the unit matrix in this case) as a submatrix. Each test form has five items with no common item:

$$Q_X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, Q_Y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Equity criterion. With two test forms satisfying the same construct criterion as described in the previous section, AMPs are obtained as the link between two different sets of IRPs on each test form as shown in Figure 1. With slipping, guessing, and other random errors in real-world

Figure 1. A graphic illustration of the relationships among AMP, IRPs, and observed-score distributions.
 Note. AMP = attribute mastery pattern; IRP = Ideal response pattern; ORP = observed response pattern.

testing, observed response patterns (ORPs) deviate from IRPs. Then for each IRP, corresponding observed scores of each ORP form a distribution. The observed-score distribution for Form Y of a certain examinee whose AMP is α is denoted as $F_{Y|\alpha}(y)$, and his or her distribution of the equated observed scores from Form X to the scale of Form Y is denoted as $F_{\varphi(X)|\alpha}(\varphi(x))$.

Therefore, the equity criterion for the full distributions of observed scores on X and Y given AMP can be written as

$$F_{Y|\alpha}(y) = F_{\varphi(X)|\alpha}(\varphi(x)), \alpha \in \Omega. \quad (4)$$

The equity criterion can also be expressed as the requirement of zero equating error for all AMPs:

$$e(x; \alpha) \equiv F_{Y|\alpha}(y) - F_{\varphi(X)|\alpha}(\varphi(x)), \alpha \in \Omega. \quad (5)$$

The Family of True Transformations

In the following section, the family of true transformations is derived from the equity criterion. When the error for each attribute pattern in Equation 3 is required to be 0, y can be solved by taking the inverse of $F_{Y|\alpha}$:

$$y = \varphi_{\alpha}(x) = F_{Y|\alpha}^{-1} F_{\varphi(X)|\alpha}(\varphi(x)), \alpha \in \Omega. \quad (6)$$

Because the function $\varphi(\cdot)$ should be monotone, $F_{\varphi(X)|\alpha}(\varphi(x)) = F_{X|\alpha}(x)$. Then, the following equation is obtained

$$\varphi_{\alpha}(x) = F_{Y|\alpha}^{-1}(F_{X|\alpha}(x)), \alpha \in \Omega, \quad (7)$$

as the family of true equating transformation.

The derivation of the family of true equating transformation under cognitive diagnosis models is similar to that under IRT. It is a comparatively simple step from equating conditional on θ to equating conditional on α as they both represent the latent ability level of examinees.

The following theorem is offered:

Theorem. For the population of test takers P for which test scores X and Y measure the same construct, equating with the family of transformations $\varphi_{\alpha}(X)$ in Equation 7 has the following

properties: (a) equity for each $p \in P$, (b) symmetry in X and Y for each $p \in P$, and (c) invariance within P .

Proof. (a) For each $p \in P$, there is a corresponding AMP denoted as a vector α , and for each α , the transformation in Equation 7 matches the conditional distributions of $\phi(X)$ and Y given α . (b) The inverse of $F_{Y|\alpha}^{-1}(F_{X|\alpha}(x))$ is $F_{X|\alpha}^{-1}(F_{Y|\alpha}(y))$, which is Equation 7 for equating from Y to X . (c) The conditional formulation of Equation 7 implies independence from the distribution of α over P . As a consequence, the family holds for any subpopulation of P .

A CDM-Based Local Equating Method

Before equating, a CDM is applied to examinee responses to obtain estimates for item parameters and AMPs. The item parameter estimates are used to calculate the conditional distributions of observed scores given AMP using an adaptation of the recursive algorithm of Lord and Wingersky (1984). The probabilities and distributions are conditional on discrete α instead of continuous θ , then the family of transformations in Equation 7 can be obtained. Examinees taking either test form have their AMP estimates, so the equating transformation for a given examinee is chosen from the family of estimated equating transformations according to the estimated AMP denoted as $\hat{\alpha}$. Upon substitution of $\hat{\alpha}$ into Equation 7, the following estimated true transformation is obtained:

$$\phi_{\hat{\alpha}}(x) = F_{Y|\hat{\alpha}}^{-1}(F_{X|\hat{\alpha}}(x)), \hat{\alpha} \in \Omega. \quad (8)$$

The equated score for a certain examinee of observed score x can then be located.

Data and Simulation Design

The empirical study aimed to compare the proposed local method with traditional equipercentile equating and further evaluate the robustness of the local method. Because the family of true equating transformations is generally unknown for real-world tests, the traditional and local methods were evaluated using simulated data.

Equating was conducted between two simulated test forms, Form X and Form Y with no common items. Several aspects of the simulated data were held constant: The number of attributes was fixed to $K = 6$ for each test form, and the test length was set to $J = 36$. Correct response to each item requires one to three attributes. It has been well established that the accuracy of test linking would increase with test length (Kolen & Brennan, 2004). Therefore, as Form X and Form Y were chosen to be smaller than typical operational tests (van der Linden & Wiberg, 2010), it allows us to generalize the results to longer tests. The six attributes were independent, resulting in 64 AMPs. The number of items in each type (in terms of the number of attributes per item) and the Q-matrices are given in Table 1. The criterion of the same construct as defined earlier was satisfied. The slipping and guessing parameters were randomly sampled from $U(0.05, 0.25)$, respectively, for each test form.

For the distributions that involve the HO-DINA model, three ability distributions with means $\mu_0 = (-0.5, 0.0, 0.5)$ and a common standard deviation $\sigma_0 = 1.0$ were used. The elements of μ_0 stand for three groups of examinees, that is, a low-ability, a middle-ability, and a high-ability group. The higher order parameters were fixed to $\lambda_{0k} = (-1.0, -0.5, 0.0, 0.5, 1.0, 1.5)$ and $\lambda_1 = 1.0$ across all conditions. The item parameters were estimated via an expectation-maximization (EM) implementation of the marginal maximum likelihood estimation (MMLE/EM) that was implemented using the computer program R (R Development Core Team, 2011).

Table 1. The Q-Matrices for Form X and Form Y.

Item	Form X						Form Y						Number of attributes
	A1	A2	A3	A4	A5	A6	A1	A2	A3	A4	A5	A6	
1	1	0	0	0	0	0	1	0	0	0	0	0	1
2	0	1	0	0	0	0	0	1	0	0	0	0	1
3	0	0	1	0	0	0	0	0	1	0	0	0	1
4	0	0	0	1	0	0	0	0	0	1	0	0	1
5	0	0	0	0	1	0	0	0	0	0	1	0	1
6	0	0	0	0	0	1	0	0	0	0	0	1	1
7	1	0	0	0	0	0	1	0	0	0	0	0	1
8	0	1	0	0	0	0	0	1	0	0	0	0	1
9	0	0	1	0	0	0	0	0	1	0	0	0	1
10	0	0	0	1	0	0	0	0	0	1	0	0	1
11	0	0	0	0	1	0	0	0	0	0	1	0	1
12	0	0	0	0	0	1	0	0	0	0	0	1	1
13	1	0	0	1	0	0	0	0	1	0	1	0	2
14	0	1	0	1	0	0	0	1	0	0	1	0	2
15	0	0	1	0	1	0	0	0	0	0	1	1	2
16	0	0	0	0	1	1	0	1	0	1	0	0	2
17	1	0	0	1	0	0	0	0	0	1	0	1	2
18	1	0	0	0	1	0	0	0	1	1	0	0	2
19	1	0	1	0	0	0	0	0	1	0	0	1	2
20	1	0	0	0	1	0	0	0	0	1	1	0	2
21	0	0	0	1	0	1	0	0	1	1	0	0	2
22	0	0	1	0	1	0	1	1	0	0	0	0	2
23	1	0	1	0	0	0	0	1	0	0	1	0	2
24	0	0	1	0	0	1	1	0	0	1	0	0	2
25	1	1	0	0	1	0	1	0	0	1	0	1	3
26	0	1	1	0	1	0	1	0	0	1	1	0	3
27	1	0	1	0	1	0	1	0	1	1	0	0	3
28	1	1	0	0	0	1	0	1	1	1	0	0	3
29	1	1	0	1	0	0	0	1	1	0	0	1	3
30	0	0	1	0	1	1	0	1	0	1	0	1	3
31	1	0	0	1	0	1	1	0	1	0	1	0	3
32	0	1	0	0	1	1	1	0	0	1	0	1	3
33	0	1	0	1	0	1	0	1	1	0	1	0	3
34	1	0	0	1	1	0	1	0	1	1	0	0	3
35	1	1	0	0	0	1	1	0	1	0	0	1	3
36	1	0	1	1	0	0	1	0	1	0	0	1	3

The proposed local method was compared with traditional equipercentile equating under the baseline condition. Under the baseline condition, the EG design was used. The response vectors on X and Y for two middle-ability groups of test takers ($N = 10,000$, $\theta \sim N(0, 1)$). The sample size under the baseline condition was chosen to be large because the accurate estimates of the bias functions for the equating methods were of interest (van der Linden & Wiberg, 2010).

The robustness was evaluated varying the following factors.

Difference Between Two Groups of Examinees

The local method is expected to be reliable under NEG. This property was examined by conducting equating using two groups of examinees from different ability distributions, which is

referred to as the NEG design. Two pairs of data sets were used in equating, denoted as NEG1 and NEG2, respectively: (a) low-ability group ($\theta \sim N(-0.5, 1)$) for Form X and middle-ability group ($\theta \sim N(0, 1)$) for Form Y and (b) high-ability group ($\theta \sim N(0.5, 1)$) for Form X and low-ability group ($\theta \sim N(-0.5, 1)$) for Form Y.

Sample Size

Two sets of smaller sample size ($N = 5,000$; $N = 1,000$) were generated, respectively, under the EG and NEG designs.

Model Misfit

It is unlikely to fit real testing data with the CDM model perfectly. The response data generated by the DINA model were manipulated by adding 5% or 10% random errors to the item responses to create different degrees of model-data misfit.

Evaluation Criterion

Absence of any bias is deemed to be a more fundamental criterion of success in equating (van der Linden & Wiberg, 2010) than the standard error of equating, which is common in previous literature; therefore, equating bias is chosen as the evaluation criterion in this study. Bias for the two equating methods is defined as the difference between their actual transformations and the true transformations $\varphi_{\alpha}(X)$ in Equation 7. Bias is a function of x , and a different bias function was obtained for each α .

The bias functions for equipercentile equating and the local method are defined as

$$\text{Bias}(\varphi^*(x); \alpha) = \varepsilon_{X, Y; P}[\varphi^*(x) - \varphi_{\alpha}(x)], \alpha \in \Omega, \quad (9)$$

$$\text{Bias}(\varphi_{\hat{\alpha}}(x); \alpha) = \varepsilon_{U_1, \dots, U_n | \alpha}[\varphi_{\hat{\alpha}}(x) - \varphi_{\alpha}(x)], \alpha \in \Omega, \quad (10)$$

where $\varphi_{\alpha}(X)$ is the family of true equating transformations in Equation 7, $\varphi^*(x)$ is the equipercentile transformation, and $\varphi_{\hat{\alpha}}(x)$ is the family of equating transformations of the local method. $\varepsilon_{X, Y; P}$ denotes the expectation operation over X and Y with random sampling from P . $\varepsilon_{U_1, \dots, U_n | \alpha}$ denotes the operation of a conditional expectation over random responses to Form X given α .

The biases in Equations 9 and 10 cannot be evaluated in closed form, so they were obtained by Monte Carlo simulation. For the equipercentile method, θ s are sampled from an assumed population distribution. To calculate the expected value of the bias function in Equation 10, the response vectors were replicated 10 times for each test taker in the sample. The number of administrations for each AMP was large enough to get stable estimates of these functions. Observed scores with relative frequencies smaller than 10^{-4} for some AMPs were omitted from the study as suggested by van der Linden and Wiberg (2010). Consequently, some of the error functions in the figures in the next section do not run over the entire range of possible values of x for all AMPs.

Results

Figure 2 shows the true transformations for the two test forms. A different transformation function was obtained for each α . Transformation function curves above or below the diagonal line indicate that the relative difficulty of the two test forms differs among examinees of different

Figure 2. The family of true equating transformations of cognitively diagnostic modeled observed scores.

AMPs. For AMPs mastering only one attribute, the difference in Q-matrices does not influence their observed scores significantly, so the transformation function is very close to the diagonal line.

Equating Transformations (Baseline Condition)

The local method produced a family of equating transformations, whereas the equipercenile method has a single transformation for the given population (see the two top panels of Figure 3). There is no distinct difference between the local transformations in Figure 3 and the true transformations in Figure 2.

The Comparison of Biases (Baseline Condition)

As can be seen in Figure 3 (the two bottom panels), under the EG design, traditional equipercenile equating has much larger biases (ranging from -4.74 to 3.26) than the proposed method, especially for the examinees who master a medium number of attributes. In contrast, the local method has much smaller biases ranging from -0.73 to 0.55 for all AMPs. Specifically, 27 out of 42 (64.28%) AMPs mastering 0 to 3 attributes have seen significantly negative slopes for equipercenile equating bias functions.

Robustness of the Local Method

The robustness of the proposed local method was examined under various conditions. The transformations and bias functions of the local method with different group differences are presented in Figure 4. The traditional equipercenile equating was also conducted under the NEG

Figure 3. Equating transformations and biases of traditional equipercentile method and local method under baseline conditions.

Figure 4. Equating transformations and biases of local methods with nonequivalent groups.
Note. NEG1 = the nonequivalent groups equating design with a low-ability group and a middle-ability group; NEG2 = the nonequivalent equating design with a low-ability group and a high-ability group.

designs. As expected, the equipercentile equating suffered from even more serious bias when group difference existed, so the results are not presented here. As shown in the bottom left panel of Figure 4, the proposed method has biases as small as under the baseline condition. When the group difference becomes larger, the biases of the local method shown in the bottom right panel of Figure 4 also become larger, but the increase is not significant.

Figure 5. Equating transformations and biases of local methods under EG design and NEG design with small sample sizes.

Note. EG = equivalent groups; NEG = nonequivalent groups.

As shown in Figure 5, the local method has reliable performances with smaller sample sizes under either the EG or the NEG design. Under the EG design, the biases of the local method with $N = 5,000$ are larger compared with the biases of the local method in Figures 3 and 4 but still remain within a relatively small range of $(-1.07, 0.98)$. When group difference exists, the biases of the local method with $N = 5,000$ become slightly larger with a range of $(-0.80, 1.22)$. With the sample size of 1,000, the biases of the local method under both designs increase slightly. As can be seen in Figure 5, the local method with $N = 1,000$ has biases with a range of $(-1.07, 1.37)$ under the EG design and a range of $(-1.19, 1.41)$ under the NEG design.

When the DINA model does not fit the data very well, the biases of local method (in the bottom left panel of Figure 6) increase but a range of $(-1.08, 0.93)$ is still much smaller than that

Figure 6. Equating transformations and biases of local methods under EG design with different degrees of model misfit.
 Note. EG = equivalent groups.

of the equipercentile equating under the baseline condition. As the local method appeared to be quite robust with respect to the differences between groups, only the results under EG design are presented in Figure 6. With the degree of model misfit increasing (see the right panels of Figure 6), the bias range of the local method increases to as large as $(-1.71, 1.23)$.

Discussion

This study proposed a local equating method for cognitively diagnostic modeled observed scores. The features of CDMs facilitate local equating without anchor test. The true equating transformations of the local method were proved to satisfy the three basic criteria of equating method. Besides, this new equating method also clearly defines the same construct criterion on attribute level. Results from simulation studies suggested that the performances of this local method were quite robust even without anchor test.

While the result of traditional equipercentile equating would be a single transformation for the whole population, the results of local methods are a family of transformations conditional on latent ability level. With differential treatment of examinees with the same total score on the same test, local equating definitely has a long way to go before being adopted in practice, but the differential treatment is supported theoretically by the satisfaction equity requirement and can provide explanations for some situations in real testing world. It is a common phenomenon that sometimes a certain examinee prefers test X to test Y while another examinee finds test Y easier for himself. In the terminology of diagnostic measurement, it is likely that the construct presented by the Q-matrix of test X matches the knowledge structure of examinee p better than test Y, while the opposite is true for examinee q . For example, let us assume that tests X and Y involve three common attributes in five items: Four items in test X require the mastery of attribute No. 1, whereas only two items in test Y require attribute No. 1, and four items in test Y require attribute No. 3, whereas only two items in test X require attribute No. 3. It is also known that examinee p 's AMP is $(0\ 1\ 1)$ and examinee q 's AMP is $(1\ 1\ 0)$. Assuming all other

conditions set equal, it is easy for us to predict that test X is more difficult than test Y for examinee p , whereas test X is easier than test Y for examinee q . The relative difficulty of tests X and Y can be well interpreted from the difference in the Q -matrix, and its evidence can be found directly from conditional observed-score distributions. As it can be proved that the relative difficulty of two tests varies individually, treating all the examinees with a single equating transformation seems unreasonable. In this sense, local equating with diagnostic classification models clearly explains why people should be equated differently.

An important by-product of the new equating approach is the definition of test construct on attribute level. How test construct is defined is closely connected with the equating procedure that follows. It is advantageous to define test construct in a more refined manner. Some test programs (e.g., SAT Reasoning Test) follow very strict and thorough content and statistical specifications to assemble forms (Liu & Walker, 2007), which leads to a well-defined test construct; this evidence can be used to support the equatability of different test forms. The adoption of test forms in large-scale assessment because of test security issues gave rise to the need of equating. Equating, in turn, poses requirements on the test forms. Traditionally, the work of equating researchers begins only after the test forms are ready to be administered. The authors, however, suggested that “equating” should begin much earlier. It might be possible to further reduce the equating biases of the local method by using certain types of Q -matrices.

Although it seems to be a simple step from local equating based on IRT to local equating based on CDMs, it should be noted that the real motivation of combining CDMs with local equating is to solve the problem of equating without anchor test. The proposed method has offered an approach to reliable equating without anchor test or the need for representative samples. In the simulation studies, the performances of the local method were put through examinations under two NEG designs. The equating biases of the local method turned out to be more or less unaffected by group ability differences. Except for the slight distinction when the sample size was as small as 1,000, there is no significant difference between the biases functions under EG design and NEG design. The property of parameter invariance of the DINA model (de la Torre & Lee, 2010) plays an important role here as the item parameters are used to calculate the family of equating transformations. The robustness of the local method was also examined using different sample sizes. With a large sample size of 10,000, the equating biases are within a range of -1 and 1 . Halving the sample size, the biases did not increase drastically and were still in a range smaller than equipercentile equating under the baseline condition. The biases remained in a range between -1.5 and 1.5 even with a sample size of 1,000. This property is important because equating with small samples has always been a difficult issue.

It should be noted that the DINA model was adopted in this study only for illustration purposes. It is impossible that the DINA model could hold in any real-world testing situation, and the results of the simulation studies also demonstrated that as the degree of model misfit increases, equating biases become larger significantly. It is only reasonable to employ other CDMs in the local method if the DINA model could not fit the data. It should be a comparatively simple step to generalize the proposed method to other CDMs. The only assumption made when using the method is that a certain AMP have the same meaning in different test forms. In addition, parameter invariance of other CDMs should be checked before the local method is applied.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for this research was provided by two research grants from the National Natural Science Foundation of China (30670718 and 31371047 to Tao Xin).

References

- Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225-264.
- Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice*, 20, 6-18.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, 47, 115-127.
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26, psychometrics* (pp. 979-1030). Amsterdam, the Netherlands: Elsevier.
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Mahwah, NJ: Erlbaum.
- Ding, S., Luo, F., Cai, Y., Lin, H., & Wang, X. (2008). Complement to Tatsuoaka's Q matrix theory. In K. Shigemasa, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 417-424). Tokyo, Japan: Universal Academy Press.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301-321.
- Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika*, 55, 477-494.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: Praeger.
- Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics*, 16, 93-100.
- Harris, D. J., & Gao, X. (2003, April). *A conceptual synthesis of context effect*. In Context effects: Implications for pretesting and CBT—American Educational Research Association Annual Meeting, Chicago, IL.
- Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.
- Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, 41, 205-237.
- Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS RB-55-23). Princeton, NJ: Education Testing Service.

- Li, F. (2009). *Linking between two tests without common items—By rule space model* (Unpublished doctoral dissertation). Beijing Normal University, China.
- Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109-134). New York, NY: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating." *Applied Psychological Measurement*, 8, 452-461.
- Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2, 99-120.
- R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org/>
- Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-Based latent class models. *Journal of Educational Measurement*, 44, 293-311.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge.
- van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*, 65, 437-456.
- van der Linden, W. J. (2006a). Equating an adaptive test to a linear test. *Applied Psychological Measurement*, 30, 493-508.
- van der Linden, W. J. (2006b). Equating error in observed-score equating. *Applied Psychological Measurement*, 30, 355-378.
- van der Linden, W. J. (2010). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking* (pp. 201-223). New York, NY: Springer.
- van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement*, 34, 620-640.