Article

The Delta-Scoring Method of Tests With Binary Items: A Note on True Score Estimation and Equating

Educational and Psychological Measurement
2018, Vol. 78(5) 805–825

DOI: 10.1177/0013164417724187
journals.sagepub.com/home/epm
SAGE

Dimiter M. Dimitrov[1,2]

Abstract

This article presents some new developments in the methodology of an approach to scoring and equating of tests with binary items, referred to as delta scoring (D-scoring), which is under piloting with large-scale assessments at the National Center for Assessment in Saudi Arabia. This presentation builds on a previous work on delta scoring and adds procedures for scaling and equating, item response function, and estimation of true values and standard errors of D scores. Also, unlike the previous work on this topic, where D-scoring involves estimates of item and person parameters in the framework of item response theory, the approach presented here does not require item response theory calibration.

Keywords

test scoring, test equating, delta-scoring, assessment, testing

There are ongoing efforts in the theory and practice of measurement on comparing and bridging concepts and procedures from the classical test theory (CTT) and item response theory (IRT) in efforts to achieve simplicity in test scoring, equating, and interpretations under a specific context and purpose of measurement (e.g., Bechger, Maris, Verstralen, & Beguin, 2003; DeMars, 2008; Dimitrov, 2003, 2016; Fan, 1998; Hambleton & Jones, 1993; Houston, Borman, Farmer, & Bearden, 2006; Kohli,

[1]George Mason University, Fairfax, VA, USA
[2]National Center for Assessment, Riyadh, Saudi Arabia

Corresponding Author:
Dimiter M. Dimitrov, Professor Emeritus, George Mason University, 4400 University Drive, MS 6D2, Fairfax, VA 22030-4444, USA.
Email: ddimitro@gmu.edu

Koran, & Henn, 2015; Lin, 2008; MacDonald & Paunonen, 2002; Oswald, Shaw, & Farmer, 2015; Raykov & Marcoulides, 2015).

In the context of large-scale assessments with tests of binary items, Dimitrov (2016) proposed an approach to test scoring and equating, referred to as *delta scoring* (*D*-scoring) which is under successful piloting at the National Center for Assessment (NCA) in Saudi Arabia. This approach is extended here with procedures for scaling and equating of *D* scores, item response function (IRF) on the delta scale, and estimation of true values and standard errors of *D* scores, without the intermediate role of IRT calibration. A brief description of the initial delta-scoring method (Dimitrov, 2016) and its extension with new procedures is provided next.

Theoretical Framework

The Conception of "Delta Scoring"

Under the ''delta scoring'' (*D*-scoring) of tests with binary items, the *D* score of a person is based on the person's response vector weighted by the expected difficulties of the items for the target population of test takers (Dimitrov, 2016). Specifically, if π_i is the expected ''easiness'' of item *i* (the proportion of correct item responses by the targeted population), the *expected item difficulty* is $\delta_i = 1 - \pi_i$. For a specific test with *n* binary items, Dimitrov (2016) defined the *D* score of person *s* as a linear combination of the person's binary scores, X_{si} (1/0) weighted by the expected item difficulties, $\sum_{i=1}^{n} X_{si}\delta_i$. It should be noted that Dimitrov (2016) used the notation D_s for this linear combination, but hereafter this notation is used consistently for *D* scores on a scale from 0 to 1 to facilitate their interpretation. Specifically, currently adopted with automated large-scale assessments at the NCA in Saudi Arabia is the D_s scoring obtained as follows:

$$D_s = \frac{\sum_{i=1}^{n} X_{si}\delta_i}{\sum_{i=1}^{n} \delta_i}. \quad (1)$$

The D_s scores range from 0 to 1 ($0 \leq Ds \leq 1$), with $D_s = 0$ if all items are answered incorrectly ($X_{s1} = 0, \ldots, X_{sn} = 0$) and $D_s = 1$ if all answers are correct ($X_{s1} = 1, \ldots, X_{sn} = 1$). One can interpret the D_s score of a person as the proportion of the ability required for ''total success'' on the test ($D_s = 1$) demonstrated by that person. The resulting scale is referred to as ''delta scale'' (or just *D*-scale) as the D_s score is based on the person's response vector weighted by expected item difficulties, δ_i (Greek ''delta''). It should be noted that the psychometric properties of the original *D* scores (intervalness, reliability, and so forth), presented by Dimitrov (2016), are preserved with the D_s scores under the linear transformation with Equation (1)[1].

As an example, the computation of *D* scores under Equation (1) is illustrated here for a hypothetical test of five items with expected item difficulties $\delta_1 = 0.20$, $\delta_2 = 0.35$, $\delta_3 = 0.50$, $\delta_4 = 0.65$, and $\delta_5 = 0.80$. Thus, the denominator in Equation (1) is $\sum_{i=1}^{n} \delta_i = 0.20 + 0.35 + 0.50 + 0.65 + 0.80 = 2.50$, which can be seen as the total

difficulty of the test. Under this scenario, the response vectors of four persons and their respective D scores are provided in Table 1.

As can be seen, the second and third persons in Table 1 have the same total score ($X = 3$) but different D scores ($D_2 = 0.48$, $D_3 = 0.60$) because of different response vectors. The score of the first person is $D_1 = 0$ as none of the items is answered correctly. The fourth person has a perfect score ($D_4 = 1$) as all items are answered correctly (this person demonstrated 100% of the ability required for total success on the test).

Estimation of Expected Item Difficulty

Item Response Theory–Based Estimation. In the previous article on *delta scoring* (Dimitrov, 2016), the estimation of the expected difficulty of an item, δ_i, is based on a formula that represents δ_i as a function of IRT item parameters (under the 1PL, 2PL, or 3PL model). Specifically, the expected item ''easiness,'' π_i, is obtained as a function of the item parameters (a_i and b_i) under the two-parameter logistic (2PL) model in IRT (Dimitrov, 2003):

$$\pi_i = \frac{1 - erf(X_i)}{2}, \tag{2}$$

where $X_i = a_i b_i / \sqrt{2(1 + a_i^2)}$, *erf* is the known mathematics function called *error function,* a_i is the item discrimination, and b_i is the item difficulty. Then the expected item difficulty is computed as $\delta_i = 1 - \pi_i$ (for more details, see Dimitrov, 2016).

Bootstrap Estimation. To avoid the use of IRT information, the estimation of δ_i for *delta scoring* at the NCA is currently based on the method of bootstrapping (Efron, 1979). In fact, both the IRT-based and bootstrap estimation procedures are available with the computerized system for automated test scoring and equating at the NCA (SATSE; Atanasov & Dimitrov, 2015). An important advantage of the bootstrapping is that it does not require assumptions associated with the use of Equation (2) (e.g., IRT model fit, normal distribution of the examinees' ability scores, etc.). Although with very large samples of examinees (e.g., $N > 10{,}000$) the sample-based δ_i can be used, the bootstrap approach is recommended because it provides more accurate estimates of δ_i along with the distribution of its values across thousands of random samples from the target population and their standard error, $SE(\delta_i)$.

A simulation study under the development of SATSE (Atanasov & Dimitrov, 2015) showed that $SE(\delta_i)$ decreases with the increase of the sample size, but in all cases it reaches its highest value when the population $\delta_i = 0.5$ and decreases when δ_i approaches 0 or 1. This is illustrated in Table 2 with the results for two cases ($N =$ 1,000 and $N = 10{,}000$), with values of the population δ_i ranging from 0.1 to 0.9. For example, for the case of $N = 1{,}000$ the bootstrap estimate of the population $\delta_i = 0.5$is $\hat{\delta}_i = 0.49$ (rounded to two decimal digits). However, if one uses just a sample estimate of δ_i, its value can be much less accurate, ranging from 0.33 to 0.67.

Table 1. Computation of *D* Scores for Four Response Vectors on Five Binary Items.

	$\delta_1 = 0.20$	$\delta_2 = 0.35$	$\delta_3 = 0.50$	$\delta_4 = 0.65$	$\delta_5 = 0.80$			
Person	Xs_1	Xs_2	Xs_3	Xs_4	Xs_5	$\sum_{i=1}^{5} X_{si}\delta_i$	$\sum_{i=1}^{5} \delta_i$	*D* score
1	0	0	0	0	0	0	2.50	0
2	1	1	0	1	0	1.20	2.50	0.48
3	1	0	1	0	1	1.50	2.50	0.60
4	1	1	1	1	1	2.50	2.50	1

Note. The *D* scores are computed with the use of Equation (1) ($s = 1, 2, 3, 4$; $i = 1, 2, 3, 4, 5$).

Table 2. Bootstrap Estimates of Expected Item Difficulties, With Their Standard Error (*SE*) and Distribution Range.

Population δ_i	Bootstrap $\hat{\delta}_i$	$SE(\hat{\delta}_i)$	min $\hat{\delta}_i$	max $\hat{\delta}_i$
$N = 1{,}000$				
0.90	0.89	0.029	0.79	0.96
0.80	0.79	0.040	0.63	0.89
0.70	0.71	0.044	0.53	0.85
0.60	0.57	0.046	0.46	0.74
0.50	0.49	0.047	0.33	0.67
0.40	0.41	0.046	0.25	0.58
0.30	0.30	0.045	0.19	0.47
0.20	0.19	0.038	0.09	0.35
0.10	0.10	0.029	0.03	0.22
$N = 10{,}000$				
0.90	0.90	0.009	0.87	0.93
0.80	0.80	0.012	0.76	0.84
0.70	0.70	0.014	0.66	0.74
0.60	0.59	0.015	0.55	0.64
0.50	0.51	0.016	0.46	0.56
0.40	0.40	0.015	0.36	0.45
0.30	0.31	0.014	0.26	0.35
0.20	0.21	0.012	0.17	0.26
0.10	0.10	0.009	0.07	0.13

Note. The bootstrap estimate, $\hat{\delta}_i$, is the mode of the distribution of $\hat{\delta}_i$ values obtained with a bootstrap resampling over 1,000 replications.

Item–Person Map on the Delta Scale

With the use of Equation (1), the *D* scores of examinees and the expected item difficulties, δ_i ($i = 1, \ldots, n$), are represented on a same scale (from 0 to 1). Also, the *D* scores are conceptually comparable to the expected item difficulties, δ_i, as the *D* scores are direct function of δ_i values (see Equation 1). Thus, one can obtain an

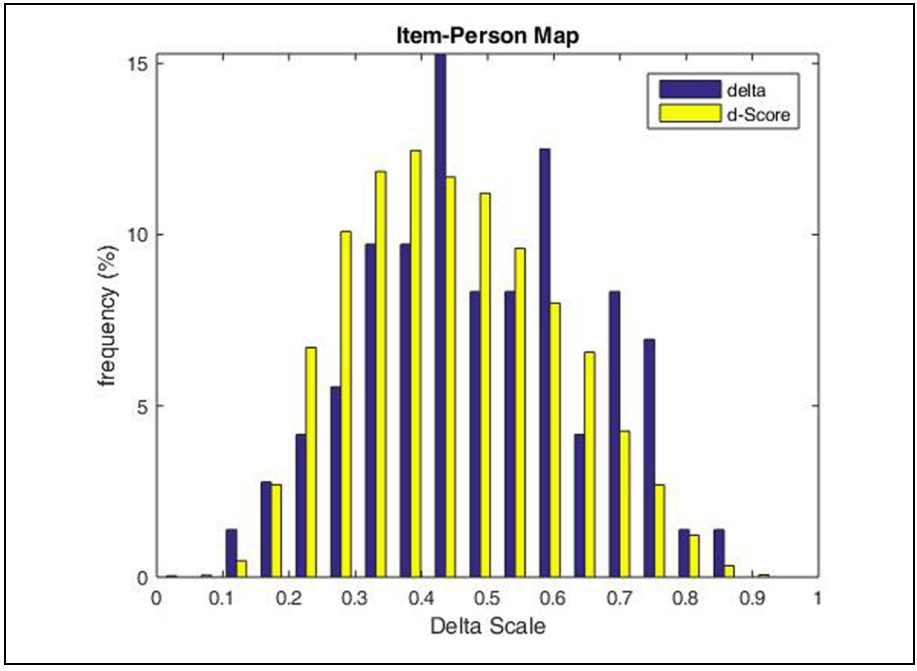

Figure 1. Item–person map for D scores and δ_i values ("deltas") obtained with data from the administration of the general aptitude test (GAT) to 3,460 high-school graduates in Saudi Arabia (GAT consists of 72 dichotomously scored item responses, 1 = *correct*, 0 = *incorrect*).

''item–person map'' (IPM) by representing the frequency distributions of the D scores and δ_i values on the same scale.

As an example, Figure 1 shows the IPM obtained with test data for 3,460 high-school graduates who took the general aptitude test (GAT) administered by the NCA in Saudi Arabia (GAT is a standardized test with 72 dichotomously scored items). Clearly, there is good overlap between the range of examinees' ability levels, as measured by their D scores, and the range of expected item difficulties, δ_i ($i = 1, \ldots, 72$). In general, the interpretation of the IPM on the D-scale is similar to the IPM interpretation in the framework of IRT.

Item Response Function on the Delta Scale

In IRT, the probability for correct response on an item, given the examinee's ability on the IRT logit scale, is estimated with the use of an appropriate logistic model. For example, under the 2PL model this probability is estimated as

$$\mathrm{P}(X_{si} = 1 | \theta_s) = \frac{e^{Da_i(\theta_s - b_i)}}{1 + e^{Da_i(\theta_s - b_i)}}, \qquad (3)$$

where X_{si} is the binary score of person s on item i (1 = *correct answer*; 0 = *otherwise*), θ_s is the ability of person s (on the logit scale), a_i is the item discrimination, and b_i is the item difficulty (D is a scaling factor, with $D = 1.7$ used to approximate the two-parameter normal ogive model).

In IRT, Equation (3) is involved in complex sequential procedures for estimating the item parameters (a_i, b_i) and the person's parameter (θ_s) using, say, the marginal maximum-likelihood parameter estimation with the EM algorithm (MML; Bock & Lieberman, 1970; Dempster, Laird, & Rubin, 1977). Under the *delta scoring* method such complexity is avoided as the estimation of D scores via Equation (1) is based on bootstrapping of expected item difficulties, δ_i. After the D scores are estimated, the probability for correct response on an item by person s, given the D_s score of that person on the *delta scale* (from 0 to 1), is estimated as a predicted item score, $\hat{X}_{si}$, with the use of the following 2PL regression[2]

$$\hat{X}_{si} = \mathrm{P}(X_{si} = 1 | D_s) = 1 - \frac{1}{1 + \left(\frac{D_s}{b_i}\right)^{a_i}}, \tag{4}$$

where D_s is the known independent variable (predictor), obtained via Equation (1), whereas a_i and b_i are regression coefficients. In fact, $\mathrm{P}(X_{si} = 1 | D_s)$ is the true score on item i for a person with score D_s (see Note 3 for the odds of the person's success on item i given the person's D_s score).[3]

The regression coefficients in Equation (4) are analogous to (yet different from) the IRT parameters a_i and b_i in Equation (3). The same notations (a_i and b_i) are used in both equations to emphasize the analogy, but in the remaining these notations will be used only with their meaning in Equation (4), so confusion will be avoided. In both Equations (3) and (4), b_i is the ''location'' of the item, that is, the location on the logit or D-scale, respectively, where the probability of correct response on the item is 0.5 (50% chances for success), whereas a_i is the slope of the response function at the location (item discrimination at b_i). It should be noted, however, a_i and b_i in Equation (3) are estimated via complex procedures such as MML within the framework of IRT, whereas a_i and b_i in Equation (4) are simply regression coefficients.

True Values and Standard Errors of D Scores

Let $P_i(D_s)$ is a short notation of $P(X_{si} = 1 | D_s)$, that is, the probability of correct item response by a person with a score D_s on the delta scale (see Equation 4). Note that $P_i(D_s)$ is the ''true'' (expected) value of the observed binary score X_{si} for subjects with a score D_s on the delta scale. On the other side, the ''true'' (expected) value of the observed D score, denoted $E(D_s)$, is obtained via Equation (1) by replacing the observed X_{si} scores with their expected values, $P_i(D_s)$, obtained with the use of Equation (4). That is,

$$E(D_s) = \frac{\sum_{i=1}^{n} P_i(D_s)\delta_i}{\sum_{i=1}^{n} \delta_i}. \quad (5)$$

The error associated with the observed score D_s, denoted $\varepsilon(D_s)$, is the difference between D_s and its expected value, $E(D_s)$, that is, $\varepsilon(D_s) = D_s - E(D_s)$. Based on a formula for $\varepsilon(D_s)$ derived by Dimitrov (2016, appendix A), which is adapted here for D_s scores obtained via Equation (1), the standard error of D_s can be computed as follows:

$$SE(D_s) = \left(\frac{1}{\sum_{i=1}^{n} \delta_i}\right) \sqrt{\sum_{i=1}^{n} \delta_i^2 P_i(D_s)[1 - P_i(D_s)]}. \quad (6)$$

Example 1: Scoring and Item Characteristic Curves on the Delta Scale

The data for this example come from simulated responses of 3,000 subjects on 20 binary items under the 1PL model in IRT (Equation 3, with $a_i = 1$) based on the random selection of (a) subjects' ability scores, θ_s ($s = 1, \ldots, 3{,}000$), from the standard normal distribution, $\theta_s \sim N(0, 1)$ and (b) item difficulty parameters from a uniform distribution in the interval from -2.0 to 2.0 (see Table 3 for the selected 1PL item parameters, denoted b_i^* to avoid confusion with the regression coefficient b_i on the D-scale metric). Using the simulated binary scores (1/0), the estimation of the expected item difficulties, δ_i ($i = 1, \ldots, 20$), was performed by using the bootstrap function in MATLAB (MathWorks, Inc., 2015) with 1,000 replications, taking the mode of the resulting distribution of δ_i values as an estimate of δ_i (the mode was found to be a slightly more accurate estimate of the population δ_i compared with the mean and median of its sampling distribution; Atanasov, 2016b). The estimates of δ_i ($i = 1, \ldots, 20$) are given in Table 3.

Figure 2 shows the distribution of D scores obtained via Equation (1) with the simulated binary scores of 3,000 subjects on 20 items, X_{si} ($s = 1, \ldots, 3{,}000$; $i = 1, \ldots, 20$), and the δ_i values in Table 3. The correlation between these D scores and the simulated 1PL-based ability values, θ_s, was very high (0.989). The IPM for the distributions of D scores and δ_i values on the *delta scale* is provided with Figure 3.

Using the 2PL regression model in Equation (4) for each item separately, with the binary scores on the item, X_{si} ($s = 1, \ldots, 3{,}000$; $i = 1, \ldots, 20$), as the dependent variable and the D scores of the subjects as the independent variable (predictor), the resulting estimates of the regression coefficients for *location* (b_i) and *discrimination* (a_i = slope at b_i) are given in Table 3. For all items, the F-ratio test for data fit of the regression model was statistically significant, with $p < .0001$. The regression analysis was performed with the statistical software MEDCALC (https://www.medcalc.org/index.php; other software, such as R, can also be used).

It is worth noting that the correlations among the parameters in Table 3 are almost perfect, with (a) 0.998 between δ_i and b_i, (b) 0.997 between δ_i and b_i^*, and (c) 0.998 between b_i and b_i^*. This should not be a surprise, given that these three parameters represent item difficulty from the perspective of different models, namely, classical (δ_i), IRT (b_i^*), and the D-scoring model with Equation (4) (b_i).

Table 3. Item Parameters With the Simulation of 20 Items in Example 1.

Item	Item difficulty (1PL simulation)	Delta-scale parameters		
	b_i^*	δ_i	a_i	b_i
1	0.4078	0.5984	2.5136	0.5145
2	0.5696	0.6273	2.7825	0.5647
3	−1.0610	0.2861	1.5294	0.2053
4	−0.2437	0.4541	2.0385	0.3734
5	0.3206	0.5669	2.4222	0.4934
6	−1.3762	0.2284	1.3477	0.1563
7	−0.9800	0.2992	1.5883	0.2277
8	−0.6881	0.3648	1.9164	0.2963
9	−0.3526	0.4305	1.8559	0.3423
10	0.2400	0.5617	2.5120	0.4833
11	0.5917	0.6063	2.6368	0.5359
12	1.8891	0.8320	3.7760	0.8090
13	−0.2690	0.4305	1.9908	0.3569
14	0.3673	0.5774	2.2533	0.5013
15	−0.9681	0.3097	1.4488	0.2208
16	−1.2601	0.2598	1.5613	0.1837
17	0.5225	0.5906	2.5973	0.5164
18	−1.3356	0.2520	1.4838	0.1813
19	0.9515	0.6877	2.8858	0.6247
20	0.9811	0.6982	2.9520	0.6314

As noted earlier, the predicted scores under the 2PL regression model in Equation (4) serve as estimates of the ''true'' values of the binary scores X_{si}. Thus, by plotting the predicted scores against the D scores, we obtain item characteristic curves (ICCs) on the delta scale. For illustration, the ICCs of four items are shown in Figure 4, namely, Item 6 (the least difficult item, $\delta_6 = 0.2284$), Item 8 ($\delta_8 = 0.3648$), Item 20 ($\delta_{20} = 0.6982$), and Item 12 (the most difficult item, $\delta_{12} = 0.8320$). Just like in IRT, by sum of the ICCs for all 20 items produces a test characteristic curve (TCC) on the delta scale, that is, the true number-correct responses (NCRs) score on the test.

The standard errors of the D scores, computed with the use of Equation (6), are depicted in Figure 5. They range from 0.008 to 0.109, with a mean of 0.092 and standard deviation of 0.020, which indicates high precision of measurement on the delta scale. As shown in Figure 5, the standard errors, $SE(D)$, are higher (yet relatively small) in the middle range of the delta scale and decrease toward its ends, with the decrease being more pronounced with the D scores getting closer to 0. This is a general trend for $SE(D)$ estimates on the delta scale (see Dimitrov, 2016).

Equating Test Forms on the Delta Scale

In the context of large-scale assessments at the NCA, multiple test forms are usually equated to a base form of the test using the method of IRT true score equating under

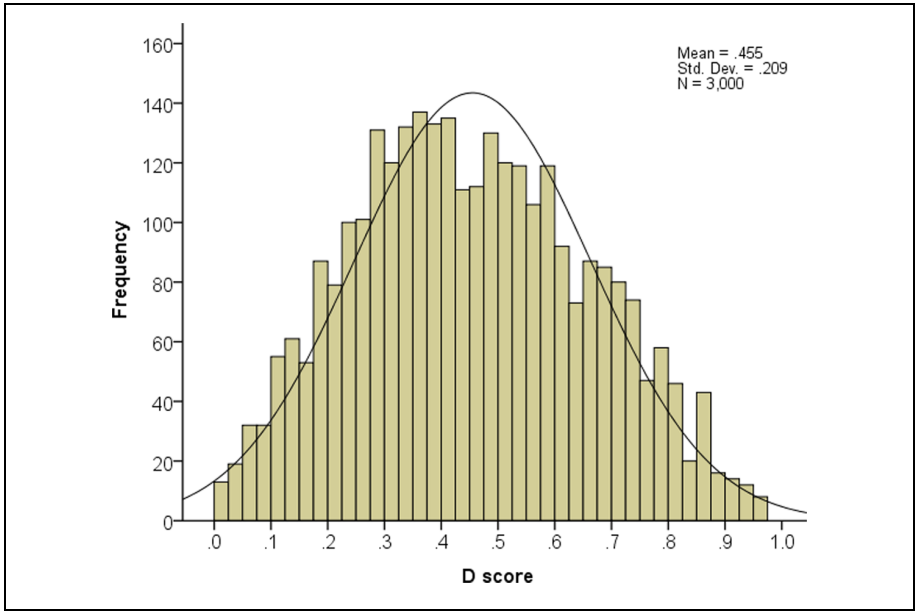

Figure 2. Frequency distribution of *D* scores obtained via Equation (1) with the binary scores of 3,000 subjects on 20 items generated with simulations under the one-parameter logistic (1PL) model, from the standard normal ability distribution, $\theta_s \sim N(0, 1)$.

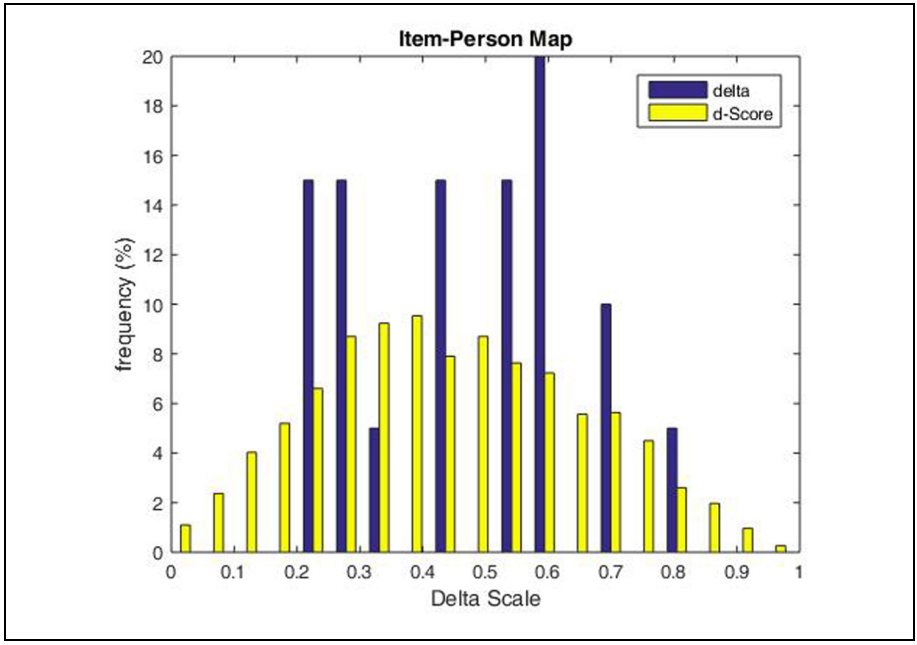

Figure 3. Item–person map for *D* scores obtained via Equation (1) with simulated binary scores of 3,000 subjects on 20 items (δ_i values in Table 3).

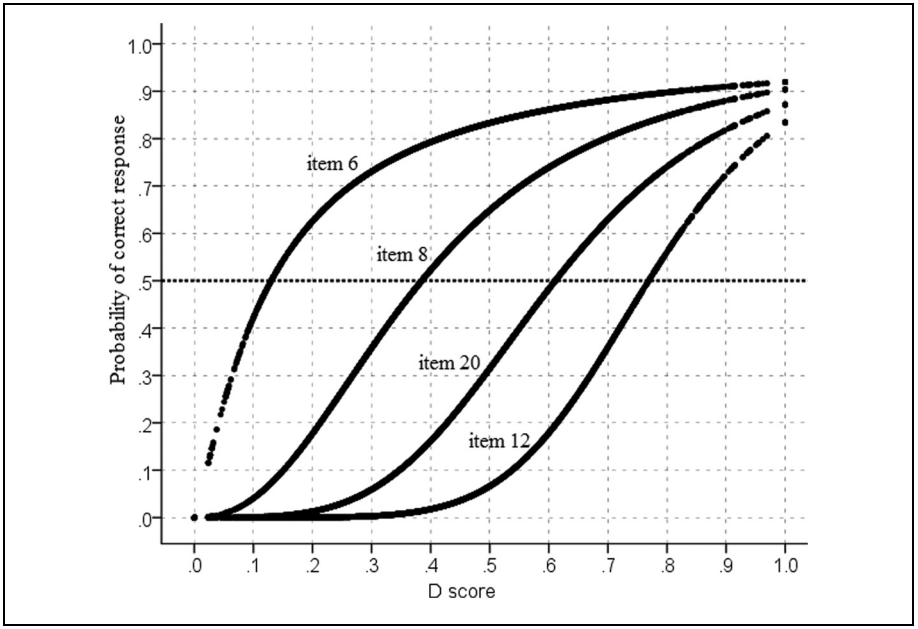

Figure 4. Item characteristic curves (ICCs) on the D-scale for four items selected from the 20 simulated items with delta parameters in Table 3: Item 6 (δ_6 = 0.2284), Item 8 (δ_8 = 0.3648), Item 12 (δ_{12} = 0.8320), and Item 20 (δ_{20} = 0.6982). The ICCs are obtained via Equation (4).

the nonequivalent groups with anchor test (NEAT) design (e.g., Angoff, 1971; Dorans, Moses, & Eignor, 2010; Kolen & Brennan, 2014; von Davier, Holland, & Thayer, 2004). This method has advantages over classical methods of test equating (e.g., see van der Linden, 2013), but its practical use involves procedures that are very complex and run into technical problems with the mapping of multiple TCCs (e.g., the Newton–Raphson procedures of tedious iterations where the choice of poor initial values leads to erroneous solutions; Kolen & Brennan, 2014, p. 194). Such problems are avoided under the approach of test equating on the delta scale described next.

The test equating on the delta scale is based on the logic of the item pre-equating design in IRT (e.g., Kolen & Brennan, 2014). A key task is to rescale the expected item difficulties of a new test Form X, δ_X, to the delta scale of Form Y using a set of common items for the two test forms. Dimitrov (2016) described an approach to this task with the intermediate role of IRT-based item parameters. However, as the goal here is to avoid the use of IRT information in delta scoring and equating, the proposed rescaling of δ_X values is based on the logical assumption of high correlation between δ_X and δ_Y values for the common items of the test forms X and Y. This assumption is tested here with simulated data, but it was also supported in many

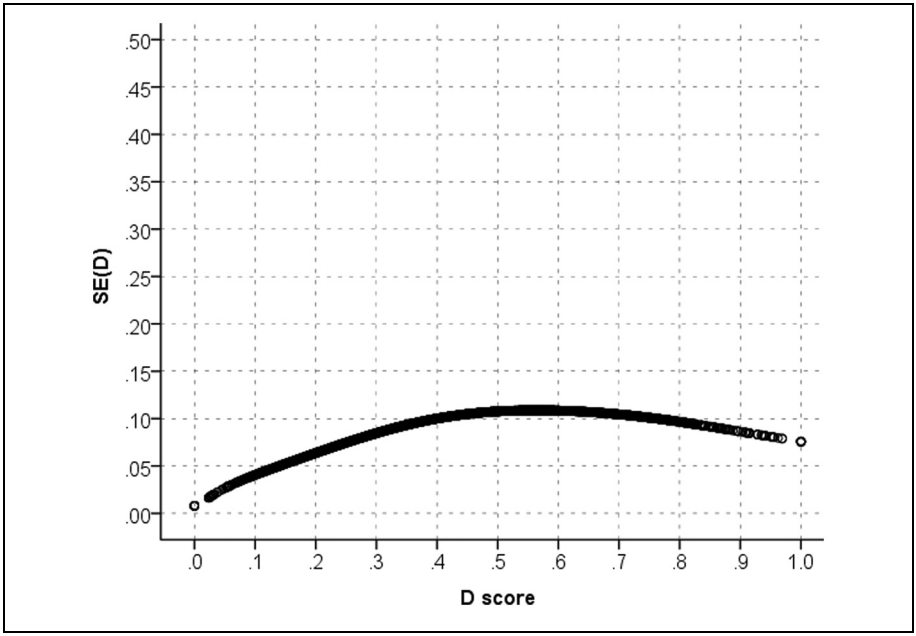

Figure 5. Standard error of measurement of *D* scores, computed via Equation (6), with the simulated data for 20 items in Example 1.

real-data analyses with large-scale assessments at the NCA (not provided here for space consideration).

Equating of Form X to Form Y on the Delta Scale

Under high correlation between the δ_X and δ_Y values for the common items of test forms X and Y, the rescaling of δ_X values to the scale of Form Y can be performed as follows. First, a simple linear regression is used with the δ_X and δ_Y values of the common items for Forms X and Y, respectively,

$$\hat{\delta}_Y = A + B\delta_X. \tag{7}$$

Second, after the regression coefficients A and B are obtained, Equation (7) is used with these coefficients and the δ_X values of all items in Form X. The resulting predicted values are, in fact, the rescaled δ_X values, denoted here δ_X^* (i.e., $\delta_X^* = \hat{\delta}_Y$). Next, the D scores on Form X, denoted here D_X, are transformed to D_X^* scores on the scale of Form Y via Equation (1) with the use of the δ_X^* values.

Delta-Equivalency of Test Forms. The comparison of transformed D_X^* scores of Form X with D_Y scores on Form Y is valid under the assumption of *delta-equivalency* of the

two test forms, which is in place if the sum of rescaled δ_X^* values of Form X is equal to the sum of δ_Y value on Form Y, that is, $\sum \delta_X^* = \sum \delta_Y$. In other words, the same level of ability is required for total success on each test form. The assumption of *delta-equivalency* can be satisfied in the practice of large-scale assessments by taking it as a restriction with optimization procedures for automated test assembly. For example, this is achieved with the system for automated test assembly (SATA; Atanasov, 2016a) designed for large-scale assessments at the NCA in Saudi Arabia, which ensures that the test forms assembled from a delta-calibrated item pool have the same number of items, with equal range of δ_i values and practical ''delta-equivalency'' (sums of δ_i values differing by not more than 0.01).

Equating of Form X to Delta-Calibrated Item Pool

The approach to test equating on the delta scale is particularly efficient in the context of large-scale assessments when new test forms are assembled from a δ-calibrated item pool. The expected item difficulties of the pool items, denoted here δ_p, are on a common scale (one can see the δ_p values as expected item difficulties for a reference population of examinees). Now suppose that a new test Form X consists of (a) operational items, assembled from the pool and (b) some nonoperational items (not used in the computation of test scores). In this scenario, all operational items of Form X are common items drawn from a δ-calibrated item pool instead of from a single old Form Y. After administering Form X, the expected item difficulties, δ_X, are estimated via bootstrapping.

To compute the D score of an examinee for the administration of Form X, Equation (1) is used with the response vector of that examinee on the operational items and their δ_X values. To place this D score onto the common scale for the reference population of the pool, Equation (1) is used with the same examinee's response vector on the operational items, but this time with their pool values, δ_p. The resulting score, denoted here D_p, is reported as the examinee's test score on Form X. This procedure can be used with multiple test forms, assembled from the pool so that they are delta-equivalent, thus, making the examinees'D_p scores comparable across such test forms.

Rescaling Nonoperational Items to the Pool Scale

Still for the above case of test Form X, another task is to rescale the δ_X values of the nonoperational (e.g., trial) items as δ_p values on the common scale of pool items, thus, extending the pool with adding these nonoperational items. To achieve this, the regression Equation (7) is used first with the δ_X values and their pool counterparts, δ_p, for the operational items to estimate the rescaling constants A and B. Then, Equation (7) is used with the rescaling constants (A and B) and the δ_X values of the nonoperational items to obtain their rescaled δ_p values for the item pool.

Example 2: Equating of Test Forms on the Delta Scale

The data for this example come from simulated responses of two groups of 2,000 subjects each on 40 binary items. The data were simulated with the use of the 2PL model in IRT, with the ability distribution being (a) $\theta \sim N(0, 1)$, for Group 1 and (b) $\theta \sim N(1, 1)$, for Group 2. That is, Group 2 comes from a population with higher ability compared with the population for Group 1. The 2PL item parameters for the simulated data were selected to produce test forms of equal average difficulty when presented on a common IRT scale. Thus, simulated were two test forms of equal difficulty taken by two groups with different abilities on a common IRT scale. Out of 40 items in each group, there were 10 common items. The simulations were replicated 20 times. It should be noted that, although the IRT framework was used for convenience with data simulations, the test equating approach illustrated here does not involve IRT parameters. The simulated data for Group 1 are considered here as a Form *Y* (base test form) and for Group 2 as a new Form *X*. The goal is to equate the *D* scores of Form *X* to the delta scale of Form *Y*. Based on the procedure described in the previous section, this goal is achieved as follows.

First, estimates of the expected item difficulties δ_Y and δ_X for the 40 items of Forms *Y* and *X*, respectively, were obtained with bootstrapping for all 20 replications. To illustrate, δ_Y and δ_X for one simulation are provided in Table 4, where the first 10 items (in boldface) are common items for the two test forms. The results for the common items with the simulated data in all 20 replications were practically the same, with (a) $M(\delta_Y) = 0.459$ and $SD(\delta_Y) = 0.004$, for Form *Y*, and (b) $M(\delta_X) = 0.241$ and $SD(\delta_X) = 0.003$, for form *X*.

Second, the correlation between δ_Y and δ_X values for the 10 common items were computed for all 20 simulations. For the common items in Table 4, this correlation is 0.946. For all 20 simulations, the correlations ranged from 0.927 to 0.969 ($M = 0.950$ and $SD = 0.011$).

Using the regression in Equation (7) with δ_Y and δ_X values for the 10 common items given in Table 4, the rescaling constants were found to be $A = 0.207$ and $B = 1.061$. Then, using the δ_X values of all 40 items in the equation $\delta_X^* = 0.207 + 1.061\delta_X$, the δ_X values were rescaled to δ_X^* values on the scale of Form *Y* (the δ_X^* values are given in Table 4). Summary statistics for δ_Y, δ_X, and δ_X^* are provided in Table 5 for the 10 common items and all 40 items. As the results in Tables 4 and 5 show (see also Figure 6), the expected item difficulties of the common items in Form *X* are lower than those in Form *Y* ($\delta_X < \delta_Y$) because Form *X* was generated with a higher ability population, $\theta \sim N(1, 1)$, compared with the population with Form *Y*, $\theta \sim N(0, 1)$. However, note that $\delta_X^* \approx \delta_Y$ for the rescaled common items. Also, the results in Table 5 for all items show that the rescaled item difficulties of Form *X* are almost identical to those of the Form *Y* in range, mean, and standard deviation. Thus, the two test forms can be treated as being practically ''delta-equivalent'' which validates the comparison of their *D* scores on a common delta scale. Thus, simulated were two test forms of equal difficulty taken by two groups with different abilities.

Table 4. Rescaling the Expected Item Difficulties of Form X to the Delta Scale of Form Y.

Item	Form Y δ_Y	Form X δ_X	Form X rescaled δ_X^*
1	**0.3976**	**0.2205**	**0.4409**[a]
2	**0.5630**	**0.3189**	**0.5454**[a]
3	**0.4843**	**0.2323**	**0.4534**[a]
4	**0.2874**	**0.1378**	**0.3532**[a]
5	**0.5354**	**0.2638**	**0.4869**[a]
6	**0.2874**	**0.0906**	**0.3031**[a]
7	**0.3976**	**0.1614**	**0.3783**[a]
8	**0.4803**	**0.2441**	**0.4660**[a]
9	**0.4961**	**0.2362**	**0.4576**[a]
10	**0.6772**	**0.4843**	**0.7208**[a]
11	0.5669	0.3110	0.5370
12	0.4685	0.2953	0.5203
13	0.4331	0.2126	0.4326
14	0.4921	0.2520	0.4743
15	0.5512	0.2913	0.5161
16	0.5354	0.2638	0.4869
17	0.3386	0.1378	0.3532
18	0.2598	0.1063	0.3198
19	0.2520	0.0827	0.2947
20	0.4724	0.2441	0.4660
21	0.8071	0.5551	0.7960
22	0.8307	0.5866	0.8294
23	0.2953	0.1339	0.3490
24	0.5354	0.2638	0.4869
25	0.3780	0.1457	0.3616
26	0.4331	0.2205	0.4409
27	0.3032	0.1102	0.3240
28	0.4724	0.2677	0.4911
29	0.7874	0.6142	0.8586
30	0.7402	0.5748	0.8169
31	0.2087	0.0787	0.2905
32	0.2323	0.0669	0.2780
33	0.5354	0.3268	0.5537
34	0.2559	0.1102	0.3240
35	0.5433	0.3071	0.5328
36	0.4606	0.2598	0.4827
37	0.7835	0.5591	0.8002
38	0.4409	0.1969	0.4159
39	0.8150	0.6142	0.8586
40	0.6378	0.3583	0.5871

Note. Using the regression $\hat{\delta}_Y = A + B\delta_X$ for the 10 common items (in boldface), the coefficients were found to be $A = 0.207$ and $B = 1.061$. The rescaled expected item difficulties of all 40 items in Form X are obtained with using the equation: $\delta_X^* = 0.207 + 1.061\delta_X$.

[a]For the computation of equated D scores of Form X (via Equation 1), the rescaled δ_X^* for the 10 common items are replaced with their actual values on the Y scale, δ_Y (the remaining 30 items are used with their δ_X^* values).

Table 5. Summary Statistics of the Expected Item Difficulties of Forms Y and X (Before and After Equated to Form Y).

Variable	Min	Max	*M*	*SD*
Common items (n = 10)				
δ_Y	0.287	0.677	0.461	0.122
δ_X	0.091	0.484	0.239	0.108
δ_X^*	0.303	0.721	0.460	0.115
All items (n = 40)				
δ_Y	0.209	0.831	0.487	0.173
δ_X	0.067	0.614	0.273	0.158
δ_X^*	0.278	0.859	0.497	0.168

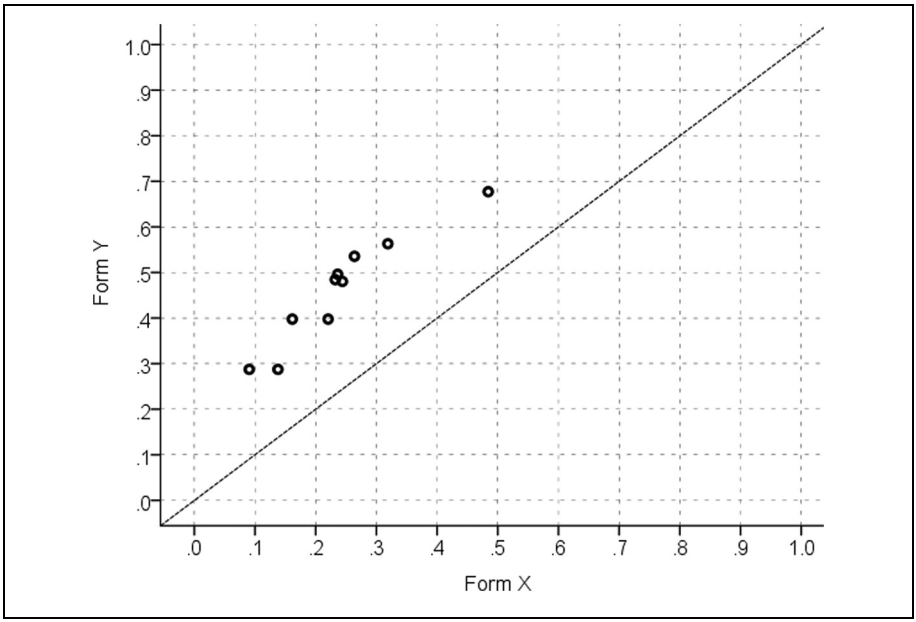

Figure 6. The δ_i values for the 10 common items on test Form *X*, prior to their rescaling, and test Form *Y* (with the simulated data in Example 2).

Third, the D scores on Form Y were obtained via Equation (1), using the examinees' response vectors on Form Y and the δ_Y values in Table 4. The initial (prior to equating) D scores on Form X were also obtained via Equation (1), with the examinees' response vectors on Form X and the δ_X values in Table 4. The equated D scores on Form X (to the scale of Form Y), denoted D_X^*, were obtained in the same way, but using the δ_Y values of the 10 common items and the δ_X^* values of the remaining 30

Table 6. Summary Statistics of *D* Scores on Forms Y and X (Before and After Equated to Form Y).

D scores	Min	Max	*M*	*SD*
D_Y	0.000	1.000	0.454	0.237
D_X	0.016	1.000	0.637	0.224
D_X^*	0.030	1.000	0.674	0.212
$SE(D_Y)$	0.000	0.075	0.063	0.015
$SE(D_X)$	0.004	0.082	0.070	0.011

Note. D_X^* corresponds to D_X equated to the scale of Form *Y* (thus, D_Y and D_X^* are comparable as they are on the same scale).

items in Table 4. The δ_Y values of the common items are considered here as more accurate estimates compared with the regression-based δ_X^* estimates of these items (or even the average of δ_Y and δ_X^* values of the common items).

Descriptive statistics of the *D* scores on Forms *Y* and *X* (prior and after equating) are provided in Table 6. The comparison of D_Y and D_X^* scores is legitimate in this case because, according to the data simulation design, the two test forms are equivalent in difficulty on their common IRT scale and, thus, they remain equivalent in difficulty on their common delta scale, that is, the two test forms can be treated as practically ''delta-equivalent'' on the delta scale of the base Form *Y* ($\sum \delta_X^* \approx \sum \delta_Y$). As can be seen, the mean of D_Y scores is smaller than the mean of D_X^* scores (0.454 $<$ 0.674). This indicates higher ability of the examinees on Form *X*, compared with those on Form *Y*, which is entirely consistent with the simulation design for the IRT-based ability distributions underlying the performance on the two test forms, $N(1, 1)$ for Form *X* and $N(0, 1)$ for Form *Y*.

The standard error of *D* scores on Form *X* (prior to equating), $SE(D_X)$, is computed via Equation (6). As expected with the simulated data, the $SE(D_X)$ values are very small (see Table 6). This is supported by the high correlation between the *D* scores on Form *X* and their true values, $E(D_X)$, namely: $r = 0.995$. In fact, the squared value of this correlation represents an estimate of the reliability of the D_X scores, that is $\hat{\rho}_{XX} = 0.995^2 = 0.990$ (e.g., Allen & Yen, 1979, p. 62). It should be noted that, theoretically, the reliability of the *D* scores equals the reliability of the NCR scores (see Dimitrov, 2016).

Discussion

This article extends a previous work on an approach to test scoring for binary items, referred to as *delta-scoring* (*D*-scoring; Dimitrov, 2016). Under this approach, the *D* score of a person is based on the person's response vector weighted by the expected difficulties of the test items, δ_i (*delta*, hence the name ''delta-scoring''). It is important to emphasize that the δ_i values are expected item difficulties and, therefore, they

do not depend on the sample of examinees who took the test. In this sense, the *D*-scoring is sample independent. The new features and procedures of *D*-scoring, rescaling, and equating, which are added here to the previously published work on this topic (Dimitrov, 2016), are summarized next.

First, the *D* scores were previously presented as a linear combination of the examinee's binary scores on the test items weighted by their δ_i values (Dimitrov, 2016, Equation 3). Now, with the use of Equation (1), the *D* scores are obtained by dividing this linear combination to the sum of δ_i values of all test items, thus, putting the *D* scores on a ''delta scale'' (from 0 to 1). Also, one can interpret the *D* score of an examinee as indicating what the proportion of the ability required for total success on the test is demonstrated by that examinee. Another advantage of using Equation (1) is that the *D* scores of persons and the δ_i values of items are represented on the same scale, which allows the depiction of ''item–person map'' (IPM) as an analog of the IPM in IRT. It should be noted, however, that the properties of *D* scores, such as intervalness and reliability, presented by Dimitrov (2016), do not change under their linear transformation with Equation (1).

Second, the expected difficulty of an item, δ_i, was previously estimated as a function of IRT parameters of the item (under an appropriate, 1PL, 2PL, or 3PL, model) via Equation (2) (Dimitrov, 2016). Under the approach used here, δ_i estimates are obtained via bootstrapping thus avoiding the need of IRT calibration and testing of assumptions related to the use of Equation (2), such as IRT model fit and (close to) normal distribution of examinees' abilities on the IRT scale.

Third, a new feature to *D*-scoring added here is the use of Equation (4) to define an IRF on the delta scale, as an analog to the IRF in IRT (e.g., see Equation 3). Just like in IRT, the regression coefficients in Equation (4) represent the item location, b_i, where the probability of correct item response is 0.5, and item discrimination, a_i, which is the slope at b_i. However, the IRT person and item parameters (θ, a_i, b_i) in Equation (3) are initially unknown and then estimated with the use of complex sequential procedures, say, using MML estimations. In contrast, under the *D*-scoring model with Equation (4), the person parameter, D_s, is known (with its preliminary estimation via Equation 1) and the item parameters (a_i and b_i) are simply estimated as regression coefficients with the logistic regression in Equation (4).

Fourth, with the use of the regression in Equation (4) to obtain the probability of correct item response across *D* scores on the delta scale, the true values and the standard errors of the *D* scores are now estimated via Equations (5) and (6), respectively, without the use of IRT estimates of person and item parameters.

Fifth, the rescaling of δ_i values of one test form to the scale of another test form (or the reference scale of item pool) was previously performed with the intermediate rescaling of IRT item parameters (Dimitrov, 2016). Such IRT-based rescaling is now avoided with the bootstrap estimation of δ_i and the use of simple linear regression in Equation (7) for the rescaling of δ_i.

Sixth, the proposed equating of *D* scores from multiple test forms to a target scale is greatly simplified because, after ensuring ''delta-equivalency'' of the test forms, it

is sufficient to rescale the item parameters of any test form to the target scale, thus, avoiding IRT mapping of test characteristic curves and tedious computations and estimation errors associated with the use of Newton–Raphson iterations in such mapping (e.g., Kolen & Brennan, 2004, p. 177).

Limitations and Future Research

In general, although the *D*-scoring framework provides some analogs to IRT concepts, such as the IRF with Equation (4) and item–person mapping on the delta scale, the classical definition of *D*-scoring (via Equation 1) entails disadvantages compared with IRT. The main limitation of the *D*-scoring approach relates to the assumption of ''delta-equivalency'' of test forms under equating, that is, the sum of the rescaled expected item difficulties, δ_i^*, should be the same across the test forms being equated. In other words, the same level of ability is required for total success on each test form to ensure valid comparisons of *D* scores across the test forms. As noted earlier, the assumption of ''delta-equivalency'' is restrictive, but its practical satisfaction can be easily achieved in the context of large-scale assessments by using appropriate procedures for automated test assembly. For example, the system for automated test assembly with large-scale assessments at the NCA in Saudi Arabia (SATA; Atanasov, 2016a) provides content representativeness and minimal measurement errors for test forms assembled from a delta-calibrated item pool under the restrictions that the test forms have the same number of items, equal range of δ_i values, and practical delta-equivalency (sums of δ_i values differing by not more than 0.01).

In other scenarios, say, when a new test Form *X* is equated to an old test Form *Y* and delta-equivalency is not in place for the two test forms, it is necessary to adjust the transformed D_X^* scores for the difference between the sums of expected item difficulties of the two test forms. To deal with this drawback, the search for appropriate adjustment procedures is underway, but their discussion is beyond the scope of the present article. Another line of future research on *D*-scoring relates to testing for item/person fit, scaling features (e.g., the performance of *D* scores with score patterns that are ordered according to their Guttman scalability), dependability of criterion-based classifications, and so forth. For example, under the *D*-scoring with large-scale assessment data at the NCA in Saudi Arabia, promising results were obtained with procedures of EM-based imputations for rescaling of δ_i values, *D*-scoring for partial credit scoring rubrics, and *D*-scoring in multistage testing, but their validation is still under investigation.

In conclusion, the methodology and procedures of *D*-scoring, rescaling, and equating of tests with binary items can be useful to researchers in both theoretical and empirical studies, as well as to the practice of large-scale assessments in the field of educational and psychological measurement.

Acknowledgments

I would like to thank Dr. Faisal Al-Mashari Al-Saud and Dr. Abdullah Al-Qataee from the National Center for Assessment (NCA) in Riyadh, Saudi Arabia, as well as Dr. Dimitar V. Atanasov from the New Bulgarian University, for their valuable comments during the piloting *D*-scoring at the NCA.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For report purposes in the framework of large-scale assessments at the NCA, the D_s scores on the scale from 0 to 1 are transformed to D'_s scores on a scale from 0 to 100 by using a simple linear transformation with a multiplication by 100, that is, $D'_s = 100D_s$. Unlike the percentile rank score, which indicates the relative performance of a person in regard to the other test takers, the D'_s score has an absolute meaning as it shows what percent of ability required for ''total success'' on the test ($D'_s = 100$) is demonstrated by that person.
2. The general form of the regression model with Equation (4) is referred to as four-parameter logistic regression:

$$\hat{Y} = d + \frac{c - d}{1 + \left(\frac{X}{b}\right)^a}, \tag{8}$$

   where Y is regressed on X, with regression coefficients d = upper asymptote, c = lower asymptote, b = location, and a = slope at b. Equation (4) is a special case of Equation (8), where the binary score of a person s on item i, X_{si}, is regressed on the subject's score D_s, with fixed asymptotes, $d = 1$ and $c = 0$, because the predicted item score, $\hat{X}_{si}$, can range from 0 to 1. If the item responses involve pseudoguessing, then the lower asymptote, c, can be freely estimated as a regression coefficient. Thus, as an analogy to the 3PL in IRT, Equation (4) will be extended to a three-parameter logistic model.
3. Under the model with Equation (4), a simple algebra reveals the following properties:
   3.1 The odds for success on item i for a person s with score D_s, denoted here O_{si}, are

$$O_{si} = \frac{P_{si}}{1 - P_{si}} = \left(\frac{D_s}{b_i}\right)^{a_i} \tag{9}$$

3.2 If two persons, with scores D_1 and D_2, have answered the same item i, the odds ratio ($OR = O_{1i}/O_{2i}$) for their success on that item is

$$OR = \left(\frac{D_1}{D_2}\right)^{a_i}. \quad (10)$$

For example, if $D_1 = 0.8$ and $D_2 = 0.4$ (on the delta scale from 0 to 1) and the item discrimination is $a_i = 2$, then $OR = 4$. That is, the odds of the first person are four times higher than the odds of the second person to answer the item correctly. For properties of odds and odds ratios for item success under an IRT model, which is beyond the scope of this article, the reader may refer to the IRT literature (e.g., Hambleton, Swaminathan, & Rogers, 1991, pp. 81-89).

References

Allen, J. M., & Yen, W. M. (1979). *Introduction to measurement theory*. Pacific Grove, CA: Brooks/Cole.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.

Atanasov, D. V. (2016a). *A system for automated test assembly* (SATA). Riyadh, Saudi Arabia: National Center for Assessment.

Atanasov, D. V. (2016b). *A computer program in MATLAB for bootstrap estimation of expected item difficulties on a test of binary items*. Riyadh, Saudi Arabia: National Center for Assessment.

Atanasov, D. V., & Dimitrov, D. M. (2015). *A system for automated test scoring and equating* (SATSE). Riyadh, Saudi Arabia: National Center for Assessment.

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model to *n* dichotomously scored items. *Psychometrika*, *35*, 179-197.

DeMars, C. (2008, April). *Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods*. *Paper presented at the annual meeting of the National Council on Measurement in Education*, New York, NY. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.492.6980&rep=rep1&type=pdf

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

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, *27*, 440-458.

Dimitrov, D. M. (2016). An approach to scoring and equating tests with binary items: Piloting with large-scale assessments. *Educational and Psychological Measurement*, *76*, 954-975.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS Research Rep. No. RR-10-29). Princeton, NJ: ETS.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*, 357-385.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS) (NPRST-TR-06-2)*. Millington, TN: Navy Personnel Research, Studies, and Technology.

Kohli, N., Jennifer Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement*, *75*, 389-405.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, linking, and scaling: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Lin, C. J. (2008). Comparison between classical test theory and item response theory in automated assembly of parallel test forms. *Journal of Technology, Learning, and Assessment*, *6*(8). Retrieved from http://files.eric.ed.gov/fulltext/EJ838620.pdf

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 921-943.

MathWorks, Inc. (2015). *Learning MATLAB (Version 8.5.0)*. Natick, MA: Author.

Oswald, F. L., Shaw, A., & Farmer, W. L. (2015). Comparing simple scoring with IRT scoring of personality measures: The Navy Computer Adaptive Personality Scales. *Applied Psychological Measurement*, *39*, 144-154.

Raykov, T., & Marcoulides, G. A. (2015). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*. doi:10.1177/0013164415576958

van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement*, *50*, 249-285.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.