A Note on the *D*-Scoring Method Adapted for Polytomous Test Items

Educational and Psychological
Measurement

2019, Vol. 79(3) 545–557

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164418786014

journals.sagepub.com/home/epm

Dimiter M. Dimitrov^{1,2} and Yong Luo²

Abstract

An approach to scoring tests with binary items, referred to as *D*-scoring method, was previously developed as a classical analog to basic models in item response theory (IRT) for binary items. As some tests include polytomous items, this study offers an approach to *D*-scoring of such items and parallels the results with those obtained under the graded response model (GRM) for ordered polytomous items in the framework of IRT. The proposed design of using *D*-scoring with “virtual” binary items generated from polytomous items provides (a) ability scores that are consistent with their GRM counterparts and (b) item category response functions analogous to those obtained under the GRM. This approach provides a unified framework for *D*-scoring and psychometric analysis of tests with binary and/or polytomous items that can be efficient in different scenarios of educational and psychological assessment.

Keywords

D-scoring method, graded response model, polytomous items, test scoring

There are ongoing efforts in the research on classical test theory and item response theory (IRT) to achieve simplicity and efficiency in test scoring and interpretations of test scores under a specific context and purpose of measurement (e.g., DeMars, 2008; Dimitrov, 2003, 2016, 2017; Fan, 1998; Hambleton & Jones, 1993; Lin, 2008; Oswald, Shaw, & Farmer, 2015). In line with this trend, an approach to scoring and equating tests with binary items, referred to as *D*-scoring, was developed as a classical analog to basic IRT models for binary items (Dimitrov, 2016, 2017). The

¹George Mason University, Fairfax, VA, USA

²National Center for Assessment, Riyadh, Saudi Arabia

Corresponding Author:

Dimiter M. Dimitrov, Professor Emeritus, Graduate School of Education, George Mason University, 4400 University Drive, MS 6D2, Fairfax, VA 22030-4444, USA.

Email: ddimitro@gmu.edu

D-scoring method is currently implemented for automated use with large-scale assessments at the National Center for Assessment in Saudi Arabia (e.g., Atanasov & Dimitrov, 2015). As some tests include polytomous items with ordered categories (e.g., to measure levels of proficiency in language testing or teacher certification tests), the purpose of this study is to propose a *D*-scoring analog to the graded response model (GRM) in IRT (Samejima, 1969, 1996). This will provide a unified framework for efficient *D*-scoring of tests that consist of binary and/or polytomous items. Presented next is a theoretical framework of basic concepts related to the GRM and *D*-scoring method, followed by the proposed design for *D*-scoring of polytomous items with a simulation study for illustration, and discussion of the results and related issues.

Theoretical Framework

Graded Response Model

The GRM works for polytomous items with ordered categories ($x = 0, 1, \dots, m$). The analytic form of the GRM is expressed as

$$P_{ix}^*(\theta) = \frac{\exp(D\alpha_i(\theta - \tau_{ix}))}{1 + \exp(D\alpha_i(\theta - \tau_{ix}))}, \quad (1)$$

where $P_{ix}^*(\theta)$ is the probability of an examinee with ability θ on the latent trait measured by the test to score in a category x or above on item i , α_i is the item discrimination parameter, and τ_{ix} is the difficulty parameter for category x ; (τ_{ix} , referred to as *category boundary* or *threshold*, is the difficulty of scoring in category x or above, rather than scoring below category x). With $(m + 1)$ ordered categories, there are m category thresholds, $\tau_{i1}, \tau_{i2}, \dots, \tau_{im}$.

The analytic function $P_{ix}^*(\theta)$ in Equation (1) is referred to as *cumulative category response function* (CCRF). The probability of an examinee with ability θ to score in category x , denoted here as $P_{ix}(\theta)$, is obtained as

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i,x+1}^*(\theta), \quad (2)$$

where $x = 1, 2, \dots, m - 1$. The probabilities at the two extreme categories are computed as follows: (a) at $x = 0$, $P_{i0}(\theta) = 1 - P_{i1}^*(\theta)$ and (b) at $x = m$, $P_{im}(\theta) = P_{im}^*(\theta)$. The analytic function $P_{ix}(\theta)$ in Equation (2), with the additional formulas for the two extreme categories ($x = 0$ and $x = m$), is referred to as a *score category response function* (SCRF) of the GRM.

D-Scoring Model for Binary Items

Under the *D*-scoring of unidimensional 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 population of test takers (Dimitrov, 2016, 2017). If π_i is the

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	X_{s1}	X_{s2}	X_{s3}	X_{s4}	X_{s5}	$\sum_{i=1}^5 \delta_i X_{si}$	$\sum_{i=1}^5 \delta_i$	<i>D</i> -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 (3) ($s = 1, 2, 3, 4$; $i = 1, 2, 3, 4, 5$).

expected “easiness” of item i (the proportion of correct item responses by the targeted population), the *expected item difficulty* is $\delta_i = 1 - \pi_i$. The δ_i values are estimated via bootstrapping (Efron, 1979) using the distribution mode as an estimate of the expected δ_i value. The results from a previous simulation study, conducted with the development of a computer program for bootstrapping of δ_i values, showed that the mode of the bootstrap distribution of δ_i values is a slightly more accurate estimate of the expected δ_i value compared with the mean and median of the distribution (Atanasov, 2016).

For a test with n binary items, Dimitrov (2017) 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 δ_i as follows:

$$D_s = \frac{\sum_{i=1}^n \delta_i X_{si}}{\sum_{i=1}^n \delta_i} \quad (3)$$

The *D*-scores range from 0 to 1 ($0 \leq D_s \leq 1$), with $D_s = 0$ if all items are answered incorrectly ($X_{s1} = 0, \dots, X_{sn} = 0$) and $D_s = 1$ if all answers are correct ($X_{s1} = 1, \dots, X_{sn} = 1$). As an example, the computation of *D*-scores under Equation (3) 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 (3) is $\sum_{i=1}^5 \delta_i = 0.20 + 0.35 + 0.50 + 0.65 + 0.80 = 2.50$, 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. The *D*-score of an examinee can be interpreted as the proportion (%) of the ability required for total success on the test (i.e., $D_s = 1$) demonstrated by the examinee. For example, the second examinee in Table 1 ($D = 0.48$) has demonstrated 48%, whereas the third examinee ($D = 0.60$) has demonstrated 60% of the ability required for total success on the test, although both examinees have the same number of correct responses on the test ($X = 3$). It is worth mentioning that the *D*-scores of examinees and expected item difficulties, δ_i , are represented on the same scale (from 0 to 1), referred to here as the *D*-scale. This allows for the development of *item–person map* by mapping the

frequency distributions of the D -scores and δ_i values on the D -scale (see Dimitrov, 2017).

D-Model of Item Response Function

The probability for correct response on item i by person s , given the D_s score of that person on the D -scale, is estimated as a predicted item score, $\hat{X}_{si}$, using the two-parameter logistic regression (2PLR) for the *item response function* (IRF):

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

where D_s is the independent variable, obtained via Equation (3), whereas a_i and b_i are regression coefficients (Dimitrov, 2017). The regression coefficients a_i and b_i in Equation (4) are analogous to (yet different from) the parameters a_i and b_i under the 2PL model in IRT. Specifically, b_i is the “location” of the item on the D -scale (from 0 to 1), where the probability of correct response is 0.5 (i.e., 50% chances for success), whereas the discrimination parameter a_i is the slope of the IRF at the item location, b_i .

Let P_{si} denotes the probability under Equation (4)—that is, $P_{si} = P(X_{si} = 1 | D_s)$. By plotting the P_{si} values against the D -scores, we obtain the item characteristic curve (ICC) on the D -scale. For illustration, the ICCs of three items (3, 5, and 20) are shown in Figure 1. The data on these items come from a Monte Carlo simulation of the binary responses (1/0) of 3,000 people on 20 items, with the 2PLR parameters of the three items (under Equation 4) being as follows: item 3 ($a_3 = 1.522$, $b_3 = 0.204$), item 5 ($a_5 = 2.402$, $b_5 = 0.493$), and item 20 ($a_{20} = 2.952$, $b_{20} = 0.631$). For additional information on the D -scoring method, including the estimation of true D -scores and conditional standard error of measurement, the reader may refer to Dimitrov (2017).

D-Scoring Design for Polytomous Items

Proposed here is a design for using D -scoring under Equation (3) with data on ordered categories of polytomous test items. In IRT, such data are typically analyzed using the GRM. To illustrate the idea, assume that a test consists of n polytomous items with four ordered categories (0, 1, 2, 3) indicating, say, levels of proficiency. To dichotomize the category scores for the purpose of using Equation (3), while preserving the hierarchical nature of these categories, each polytomous item i generates three “virtual” binary items, $X_{i,1}$, $X_{i,2}$, and $X_{i,3}$, under the design illustrated in Table 2. In general, if an examinee scores at category x of a polytomous item ($x = 0, 1, \dots, m$), he or she receives a score of 1 on the virtual item $X_{i,x}$ and all preceding virtual items. If $x = 0$, the examinee receives a score of 0 on all virtual items generated from the polytomous item.

Figure 1. Item characteristic curves (ICCs) on the *D*-scale for three items, selected from 20 simulated items, with 2PLR parameters as follows: item 3 ($a_3 = 1.522$, $b_3 = 0.204$), item 5 ($a_5 = 2.402$, $b_5 = 0.493$), and item 20 ($a_{20} = 2.952$, $b_{20} = 0.631$).
Note. The ICCs are obtained via Equation (4).

Table 2. Binary Scores of Three “Virtual” Items Generated by Possible Category Scores (0, 1, 2, 3) of One Polytomous Item.

Item	Category score	Binary scores on “virtual” items		
		$X_{i,1}$	$X_{i,2}$	$X_{i,3}$
<i>i</i>	3	1	1	1
	2	1	1	0
	1	1	0	0
	0	0	0	0

Note. A test of n polytomous items, with four ordered categories each, generates a test of $3n$ binary items analyzed under the *D*-scoring method.

It is important to note that the hierarchical dependency among virtual items generated by a polytomous item under the proposed scoring design is not a problem for the computation of *D*-scores because Equation (3) does not assume statistical local independence. As a research-based support on this argument, a previous simulation study showed that the *D*-scoring is fairly robust to violations of IRT assumptions, including local independence (Luo & Dimitrov, 2018). In contrast, the statistical local independence is a key assumption in IRT estimation procedures that use the likelihood function of a response vector on binary items (e.g., the widely used maximum likelihood

Table 3. Graded Response Model (GRM) Item Parameters for Generating Simulated Data on 15 Polytomous Items With Four Ordered Categories Each.

Item	Discrimination	Cumulative category thresholds		
	α_i	τ_{i1}	τ_{i2}	τ_{i3}
1	1.466	-3.218	-1.092	0.526
2	1.708	-1.516	-0.665	0.466
3	1.710	-1.783	0.402	2.424
4	1.289	-4.444	-0.838	1.442
5	1.828	-1.584	-0.490	0.146
6	0.875	0.113	0.145	2.678
7	1.309	-2.496	0.999	3.675
8	1.787	-0.973	-0.852	0.251
9	0.805	-1.547	1.583	3.417
10	1.510	-0.914	-0.343	1.919
11	2.316	0.115	1.102	3.053
12	1.623	-3.009	2.324	4.328
13	1.655	-0.510	0.671	0.830
14	1.366	-2.900	-0.973	-0.041
15	1.533	-2.361	-1.433	0.239

Note. The threshold τ_{ix} ($x = 1, 2, 3$) indicates the location on the GRM logit scale where the probability of a person scoring at or above category x is .5. It also represents the location on the logit scale where the person has equal chances of scoring at category x or its preceding category ($x - 1$).

estimation in IRT; e.g., see Hambleton, Swaminathan, & Rogers, 1991, pp. 33-35). Thus, the proposed design of using virtual binary items generated by polytomous items with ordered response categories is appropriate under the *D*-scoring method but not under maximum likelihood estimations in IRT.

It is also important to note that under the GRM, the discrimination parameter of a polytomous item, α_i , does not vary across the response categories of the item (see Equation 1). In contrast, under the *D*-scoring model in Equation (4), each virtual item has its own slope, a_{ix} , thus providing information about the discrimination level of the polytomous category, x , corresponding to that virtual item.

Illustration With Simulated Data

Data were simulated (in R) under the GRM for a test of 15 polytomous items with four ordered categories per item (0, 1, 2, 3), with the generating item parameters given in Table 3 and ability scores of 1,000 examinees randomly selected from the distribution $\theta \sim N(0,1)$. As shown in Table 2, with each polytomous item generating three “virtual” binary items, 45 such items were obtained and analyzed in the framework of *D*-scoring. The item parameters of these 45 virtual items, generated by the 15 polytomous items, are provided in Table 4. It was expected that the resulting *D*-scores would highly correlate with the θ scores obtained via the GRM on the 15

Table 4. Item Parameters of “Virtual” Binary Items Generated by Polytomous Items With Simulated Data.

Item	δ_{ix}	a_{ix}	b_{ix}
$X_{1,1}$	0.024	1.235	0.010
$X_{1,2}$	0.236	1.580	0.102
$X_{1,3}$	0.630	2.927	0.364
$X_{2,1}$	0.157	1.626	0.073
$X_{2,2}$	0.323	2.150	0.170
$X_{2,3}$	0.669	2.942	0.382
$X_{3,1}$	0.094	1.367	0.041
$X_{3,2}$	0.583	3.083	0.344
$X_{3,3}$	0.961	6.823	0.761
$X_{4,1}$	0.008	6.000	0.001
$X_{4,2}$	0.307	1.500	0.140
$X_{4,3}$	0.819	3.342	0.611
$X_{5,1}$	0.134	1.497	0.056
$X_{5,2}$	0.339	1.996	0.182
$X_{5,3}$	0.543	2.651	0.299
$X_{6,1}$	0.543	1.907	0.299
$X_{6,2}$	0.551	1.901	0.301
$X_{6,3}$	0.882	2.564	0.847
$X_{7,1}$	0.071	0.863	0.010
$X_{7,2}$	0.724	2.980	0.469
$X_{7,3}$	0.976	9.071	0.820
$X_{8,1}$	0.244	2.031	0.126
$X_{8,2}$	0.260	2.274	0.142
$X_{8,3}$	0.567	3.137	0.335
$X_{9,1}$	0.228	0.930	0.062
$X_{9,2}$	0.748	2.475	0.536
$X_{9,3}$	0.929	4.149	0.766
$X_{10,1}$	0.252	1.968	0.139
$X_{10,2}$	0.417	2.445	0.216
$X_{10,3}$	0.898	5.752	0.641
$X_{11,1}$	0.559	3.913	0.330
$X_{11,2}$	0.843	5.925	0.534
$X_{11,3}$	0.992	57.261	0.868
$X_{12,1}$	0.031	1.094	0.009
$X_{12,2}$	0.945	6.278	0.745
$X_{12,3}$	0.992	4.167	0.999
$X_{13,1}$	0.370	2.496	0.202
$X_{13,2}$	0.669	3.652	0.412
$X_{13,3}$	0.701	3.700	0.444
$X_{14,1}$	0.039	1.147	0.013
$X_{14,2}$	0.260	1.566	0.123
$X_{14,3}$	0.512	2.056	0.220
$X_{15,1}$	0.055	1.204	0.018
$X_{15,2}$	0.157	1.248	0.058
$X_{15,3}$	0.559	2.551	0.327

Note. $X_{i,x}$ = binary item generated by polytomous item i and category x ($i = 1, \dots, 15$; $x = 1, 2, 3$) as shown in Table 1; δ_{ix} = expected item difficulty; a_{ix} and b_{ix} are the regression coefficients obtained with the use of Equation (4) (a_{ix} = slope, b_{ix} = location parameter on the D-scale, where the probability of correct item response is .5—that is, 50% chances of correct item response). To facilitate the examination of the table, highlighted are the triplets of “virtual” binary items that are associated with the odd-number polytomous items (1, 3, 5, 7, 9, 11, 13, 15).

Figure 2. Cumulative category response function (CCRF) for categories 1, 2, and 3 of Item 3 with simulated data (obtained via Equation 1 under the graded response model [GRM]).

polytomous items. It was also expected that the category response functions (CCRF and SCRF) obtained under the GRM and the *D*-scoring would be similar in type of information they provide, but not directly comparable as they are represented on different scales—namely, the IRT logit scale for the GRM and the *D*-scale (from 0 to 1) under the *D*-scoring model. Reported next are the results for one simulated data set, but the results from all replications were practically the same.

The resulting *D*-scores varied from 0.002 to 0.954 (*Mean* = 0.316 and *SD* = 0.180) on the *D*-scale (from 0 to 1). As expected, the *D*-scores highly correlated with the GRM ability scores ($r_{D\theta} = 0.970$). Given the interval nature of the *D*-scale (Dimitrov, 2016; Domingue & Dimitrov, 2015), this indicates high consistency in the estimation of the underlying ability with the use of *D*-scoring or GRM. Also, the ordering of the GRM category thresholds, τ_{ix} , for any polytomous item in Table 3 is the same as the ordering of the location parameters b_{ix} on the *D*-scale for its corresponding three virtual items in Table 4.

The CCRFs under the GRM were obtained via Equation (1), whereas their counterparts under the *D*-scoring model were obtained via Equation (4) (i.e., they represent the IRFs of the virtual binary items generated by the respective polytomous item). For illustration, the CCRFs under the GRM and *D*-scoring model are depicted for one polytomous item—namely, Item 3. The GRM parameters of this item are its discrimination, $\alpha_3 = 1.710$, and three cumulative category thresholds, $\tau_{31} = -1.783$, $\tau_{32} = 0.402$, and $\tau_{33} = 2.424$ (see Table 3). Figure 2 shows the resulting GRM-based CCRFs of the item on the IRT logit scale. For example, $P_{3,1}^*$ is the probability of a

Figure 3. Cumulative category response function (CCRF) for categories 1, 2, and 3 of Item 3 with simulated data (obtained via Equation 4 on the D -scale).

person with ability θ on the logit scale to score in category 1 or higher rather than at category 0 ($P_{3,2}^*$ and $P_{3,3}^*$ are interpreted in a similar way).

The D -model parameters under Equation (4) for the virtual binary items $X_{3,1}$, $X_{3,2}$, and $X_{3,3}$ in Table 4, generated by the polytomous Item 3, are (a) discrimination: $a_{31} = 1.367$, $a_{32} = 3.083$, and $a_{33} = 6.823$, and (b) location: $b_{31} = 0.041$, $b_{32} = 0.344$, and $b_{33} = 0.761$. The IRFs of these virtual items, shown in Figure 3, represent the CCRFs of the corresponding categories (1, 2, 3) of the polytomous Item 3 on the D -scale. Note that the GRM-based CCRFs in Figure 2 look parallel because the item discrimination ($\alpha_3 = 1.710$) does not vary across categories under the GRM. In contrast, the CCRFs in Figure 3 have different slopes at the locations of the virtual items on the D -scale. For example, the highest slope is for the third virtual binary item ($a_{33} = 6.823$), and therefore, its corresponding polytomous category ($x = 3$) provides the highest discrimination among examinees under the D -scoring model.

In Figures 4 and 5, the SCRFs for the polytomous Item 3 were obtained via Equation (2), with the probabilities $P_{ix}^*(\theta)$ obtained via Equations (1) and (4) for the GRM and D -scoring model, respectively, for the response categories 1 and 2 of this item. According to the computation rule for the extreme response categories, described earlier in the section on GRM, the probabilities at the two extreme categories (0 and 3) in this case are computed as follows: $P_{i0}(\theta) = 1 - P_{i1}^*(\theta)$ and $P_{i3}(\theta) = P_{i3}^*(\theta)$, for both the GRM and D -scoring model. In both cases, the intersection of the SCRF curves for two adjacent categories is the scale point where the examinees have equal chances of scoring in either of these two categories. The

Figure 4. Score category response function (SCRf) for categories 0, 1, 2, and 3 of Item 3 with simulated data under the graded response model (GRM; obtained with the $P_{3,k}^*$ in Figure 2 as follows: $P_{3,0} = 1 - P_{3,1}^*$; $P_{3,1} = P_{3,1}^* - P_{3,2}^*$; $P_{3,2} = P_{3,2}^* - P_{3,3}^*$; and $P_{3,3} = P_{3,3}^*$).

Figure 5. Score category response function (SCRf) for categories 0, 1, 2, and 3 of Item 3 with simulated data under the D-scoring model (obtained with the $P_{3,k}^*$ in Figure 3 as follows: $P_{3,0} = 1 - P_{3,1}^*$; $P_{3,1} = P_{3,1}^* - P_{3,2}^*$; $P_{3,2} = P_{3,2}^* - P_{3,3}^*$; and $P_{3,3} = P_{3,3}^*$).

difference is that this point is on the IRT logit scale for the GRM (see Figure 4) and on the D -scale for the D -scoring model (see Figure 5).

Discussion

The D -scoring method for tests with binary items was developed as a classical analog of basic IRT models for binary items (Dimitrov, 2016, 2017). This effort was motivated by practical needs for simplicity, efficiency, and transparency in automated test scoring and equating in the framework of large-scale assessments at the National Center for Assessment in Saudi Arabia. As some tests include polytomous items with ordered categories, the purpose of this article was to propose an approach to using the D -scoring method with polytomous items as an analog to the GRM in IRT (Samejima, 1969, 1996).

Under the proposed approach, each polytomous item with ordered categories generates “virtual” binary items. If an examinee scored in a given category of the polytomous item, the scoring design assumes that he or she mastered this category and its preceding categories, so the virtual items corresponding to these categories are assigned a score of 1 (see Table 2). This design raises the issue of local dependence for the set of virtual items generated by a polytomous item. However, as noted earlier, the computation of D -scores via Equation (3) is robust to violation of the assumption of statistical local independence (e.g., Luo & Dimitrov, 2018). In contrast, this assumption plays a key role in the IRT estimation of ability under the widely used method of maximum likelihood estimation (e.g., Hambleton et al., 1991). Therefore, the proposed scoring design of generating virtual binary items that correspond to ordered response categories of a polytomous item is suitable under the D -scoring method but not under maximum likelihood methods of ability estimation in IRT.

The results from the simulation study in this article indicate that the approach to D -scoring of polytomous items provides dependable estimation of the examinees’ ability on such items, with very high correlation (0.970) between the GRM ability scores on the IRT logit scale and the D -scores as ability estimates on the D -scale (from 0 to 1). The value of this high correlation is enhanced by the fact that the IRT logit scale and the D -scale are both (close to) interval scales. Specifically, previous studies on this matter showed that the D -scale performs slightly better than the IRT logit scale in terms of intervalness by criteria of the additive conjoint measurement, with the difference tending to decrease with the increase of the test length (Dimitrov, 2016; Domingue & Dimitrov, 2015). Thus, there is a high consistency in the estimation of the underlying ability under the GRM and the D -scoring model for polytomous items.

The category response functions (CCRF and SCRF) obtained via the D -scoring model with virtual items are similar to their GRM counterparts in type of psychometric information that they provide. For example, under both the GRM and D -scoring model, the CCRF for a response category shows the probability of scoring in that category or higher (e.g., see Figures 2 and 3). Also in both cases, the intersection of

the SCRFs of two adjacent response categories shows the scale point where the examinees have equal chances of scoring in either of these two categories (e.g., see Figures 4 and 5). However, the category response functions obtained via the GRM and *D*-scoring model are not directly comparable as they are based on different probabilistic models and different scales. For example, an important difference is that the GRM item discrimination parameter does not vary across the response categories of the item, whereas under the *D*-scoring model, each virtual item has its own slope thus providing information about the discrimination level of each response category.

In conclusion, the main contribution of this article consists of the design for generating virtual binary items from polytomous items and the use of *D*-scoring for such virtual items that provides (a) ability scores that are consistent with their GRM counterparts and (b) item category response functions analogous to those obtained under the GRM. An advantage of the CCRFs obtained under the *D*-scoring model over their GRM counterparts is that they differentiate the discrimination power of the item response categories, whereas the GRM-based discrimination parameter does not vary across the item response categories. A consequential contribution of the proposed approach to using the *D*-scoring model for polytomous test items is that it provides a unified framework for scoring and psychometric analysis of tests with binary and/or polytomous items with ordered response categories. Although this approach is illustrated in the context of its implementation at the National Center for Assessment in Saudi Arabia, it can be efficiently used in the assessment practices of other institutions for educational and psychological assessment.

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.

References

- Atanasov, D. V. (2016). *A computer program in MATLAB for bootstrap estimation of expected item difficulties for binary test 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.
- 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://commons.lib.jmu.edu/cgi/viewcontent.cgi?article=1029&context=gradpsych>

- 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.
- Dimitrov, D. M. (2017). The delta scoring method of tests with binary items: A note on true score estimation and equating. *Educational and Psychological Measurement*, 78, 805-825.
- Domingue, B. W., & Dimitrov, D. M. (2015). *A comparison of IRT theta estimates and delta scores from the perspective of additive conjoint measurement* (Research Report, RR-4-2015). Riyadh, Saudi Arabia: National Center for Assessment.
- 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. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- 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 <https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1638>
- Luo, Y., & Dimitrov, D. M. (2018). *Robustness of the D-scoring model to violation of IRT assumptions* (Research Note-4-2018). Riyadh, Saudi Arabia: National Center for Assessment.
- 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.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- Samejima, F. (1996). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.