World Journal of Social Science Research
ISSN 2375-9747 (Print) ISSN 2332-5534 (Online)
Vol. 5, No. 2, 2018
www.scholink.org/ojs/index.php/wjssr

Original Paper

Psychometric Features of the General Teacher Test under the *D*-Scoring Model: The Case of Teacher Certification Assessment in Saudi Arabia

Dimiter M. Dimitrov[1, 2*] & Abdullah Alsadaawi[1, 3]

[1] National Center for Assessment, Riyadh, Saudi Arabia

[2] George Mason University, Fairfax, USA

[3] King Saud University, Riyadh, Saudi Arabia

[*] Dimiter M. Dimitrov, National Center for Assessment, Riyadh, Saudi Arabia

Received: April 12, 2018 Accepted: April 28, 2018 Online Published: May 11, 2018

doi:10.22158/wjssr.v5n2p107 URL: http://dx.doi.org/10.22158/wjssr.v5n2p107

Abstract

The teachers' knowledge and skills on general standards under the Saudi National Professional Teacher Standards is assessed with the use of the General Teacher Test (GTT) administered by the National Center for Assessment (NCA) in Saudi Arabia. This paper examines the psychometric features of the GTT in the framework of a new approach to test scoring, referred to as D-scoring model, which is used with assessments at the NCA. The stability of such features across four test forms of the GTT is also examined. The study findings provide valuable information about the accuracy of the GTT scores and the validity of their interpretation and decisions regarding the licensure of teachers.

Keywords

teacher assessment, test validation, psychometric analysis, test scoring models

1. Introduction

There is a significant policy focus on the human capital of teachers in Saudi Aribia. This is motivated both by the Saudi "Vision 2030" blueprint to modernize its economy and society (http://vision2030.gov.sa/en) and the substantial body of empirical evidence showing the importance of teacher quality for student achievement (Aaronson, Bar- row, & Sander, 2007; Goldhaber & Hansen, 2013). One way that Saudi states try to ensure a high-quality teacher workforce is by requiring teacher candidates to pass licensure tests, often of both their general education skills and content knowledge, as a requirement for receiving a teaching license.

In 2010, the Saudi Ministry of Education concluded an agreement with the National Center for Assessment (NCA) to develop and conduct teacher tests. In 2011, the NCA signed a contract with the Tatweer company for educational services (https://www.t4edu.com/en) to conduct a project with the aimed at improving the teaching profession. The project included the establishment of National Professional Teacher Standards (NPTS). The NPTS consists of 12 standards divided into two categories. The first category includes general standards that represent general teaching requirement, namely: professional knowledge, promoting learning, supporting learning and professional responsibility (Dimitrov & Alsadaawi, 2014). The second category of NPTS includes subject-specific teaching standards that cover 28 teaching areas. The standards serve to guide the construction of new teacher licensure examinations, identify training needs for new teachers, and ensure the quality of teaching programs.

1.1 Purpose of the Study

The examination of the psychometric features of the NCA teacher tests in the framework of the NPTS provides information about the accuracy of the test results and the validity of their interpretation and decisions regarding the licensure of teachers. The teachers' knowledge and skills on general standards under the NPTS is assessed at the NCA with the use of the General Teacher Test (GTT), which is required for teachers in all subject areas. Since 2014, the GTT has been administered five times, with more than 400,000 candidates taking the test. Psychometric features of tests are usually examined in the framework of Item Response Theory (IRT), but a current trend in the assessment practice and research at the NCA is to implement a new approach to test scoring and equating referred to as *Delta-scoring* (*D*-scoring) method (Dimitrov, 2016, 2017). Therefore, given the key role of the GTT in the assessment of teachers on NPTS, the purpose of this study is to examine the psychometric features of the GTT in the framework of the *D*-scoring model across different test forms of the GTT. A brief description of the *D*-scoring model is provided next.

2. *D*-Scoring Model

2.1 Computation of D-Scores

Under the "delta scoring" (*D*-scoring) of a test with binary items, the *D* score of a person is based on the person's response vector weighted by the difficulties of the items for the target population of test-takers (Dimitrov, 2016, 2017). 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 $\delta_i = 1 - \pi_i$. In this study the expected δ_i values are estimated via bootstrapping (Efron, 1979).

With the δ_i values available for a test of *n* binary items, the *D* score of person *s* on the test is computed as follows

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

where X_{si} is the score (1/0) of person *s* on item *i*. Clearly, 0 $\leq Ds \leq$ 1, with Ds = 0 if the answers of all items are incorrect (X_{si} = 0; *i* = 1, ..., *n*) and Ds = 1 if all answers are correct (i.e., X_{si} = 1; *i* = 1, ..., *n*). The *D*-score of an examinee can be interpreted as the proportion of ability required for a total success on the test demonstrated by the examinee. The same interpretation holds when Equation 1 is used, say, with test items grouped by content domains, thus allowing for valid comparison of the examinees' performance on the entire test and its content domains.

2.2 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, ..., *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. Thus, one can obtain an "Item-Person Map" (IPM) by representing the frequency distributions of the *D* scores and δ_i values on the same scale.

To illustrate, Figure 1 shows the IPM obtained with one of the four Test Forms (TFs) of the GTT examined in this study. This test form, denoted TF1, consists of 75 dichotomously scored items (1 = correct response, 0 = incorrect response). The interpretation of the IPM on the *D*-scale is similar to the IPM interpretation in the framework of IRT. As shown in Figure 1, there is a relatively good overlap between the range of examinees' *D* scores and the range of expected item difficulties, δ_i (*i* = 1, ..., 72). That is, there is a good match between examinees' ability measured by the test and the difficulties of its items.

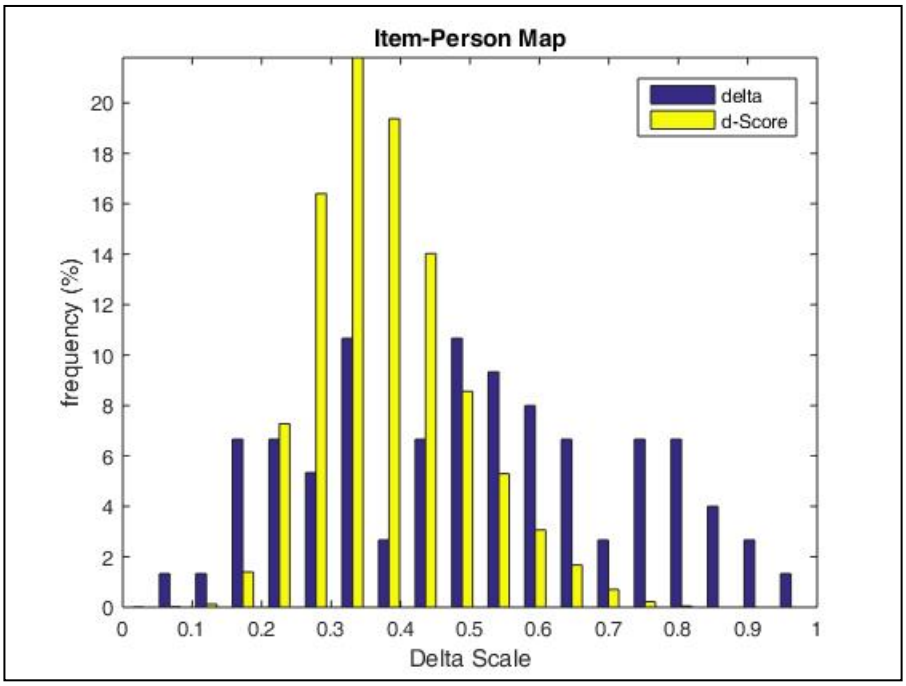

Figure 1. Item-Person Map of the Distributions of Examinees' *D*-Scores and Item Difficulties, δ, on the *D*-Scale of Test Form TF1 of the GTT.

2.3 Item Response Function on the D-Scale

Under the *D*-scoring method, 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 two-parameter logistic regression (2PLR) model:

$$\hat{X}_{si} = P(X_{si} = 1|D_s) = 1 - \frac{1}{1+\left(\frac{D_s}{b_i}\right)^{a_i}}, \quad (3)$$

where D_s is the known independent variable (predictor), obtained via Equation 1, whereas a_i and b_i are regression coefficients. In fact, $P(X_{si} = 1|D_s)$ is the true score on item *i* for a person with score D_s; (see Dimitrov, 2017).

The regression coefficients in Equation 3 are analogous to (yet different from) the IRT parameters a_i and b_i obtained under the familiar two-parameter logistic (2PL) model in IRT. Specifically, under both the IRT and *D*-scale models, 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 b_i; that is, a_i is the item discrimination at the location b_i on the logit and *D*-scale, respectively.

Let $P_i(D_s)$ is short for $P(X_{si} = 1|D_s)$ in Equation 3. The relationship between D_s scores and $P_i(D_s)$ values is referred to as Item Response Function (IRF). The graphical depiction of the IRF is referred to also as Item Characteristic Curve (ICC). For illustration, the IRFs of three items are depicted in Figure 2. These items are selected from the GTT test form denoted TF1, one the four test forms of the GTT examined in this study, with IRF parameters (*a* = slope, *b* = location) given in Table 4 as follows: item 4 (*a* = 3.3576, *b* = 0.5098), item 7 (*a* = 5.0000, *b* = 0.7992), and item 8 (*a* = 2.5760, *b* = 0.2667).

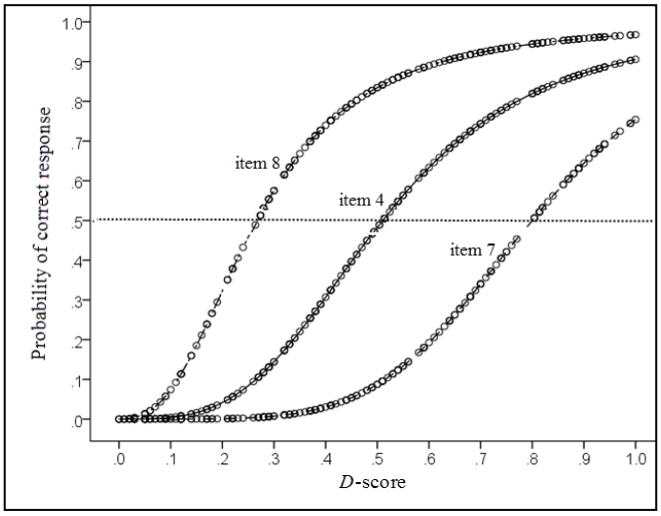

Figure 2. Item Response Functions (IRFs) on the *D*-Scale for Three Items Selected from TF1, with IRF Parameters (*a* = slope, *b* = location) Given in Table 4: Item 4 (*a* = 3.3576, b = 0.5098), Item 7 (*a* = 5.0000, *b* = 0.7992), and Item 8 (*a* = 2.5760, *b* = 0.2667)

2.4 True Values and Standard Errors of D-Scores

Note that $P_i(D_s)$, obtained via Equation 3, is the "true" (expected) value of the observed binary score X_{si} for persons with a score D_s on the *D*-scale. On the other side, the "true" 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, $P_i(D_s)$. That is,

$$E(D_s) = \frac{\sum_{i=1}^{n} P_i(D_s)\delta_i}{\sum_{i=1}^{n} \delta_i}. \quad (4)$$

The error associated with D_s, denoted $\varepsilon(D_s)$, is the difference between the score 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*- scores obtained via Equation 1, the standard error of D_s is computed as follows

$$SE(D_s) = \left(\frac{1}{\sum_{i=1}^{n} \delta_i}\right)\sqrt{\sum_{i=1}^{n} \delta_i^2 P_i(D_s)[1 - P_i(D_s)]}. \quad (5)$$

2.5 Testing for Item Fit

The testing for item fit under the 2PLR model (see Equation 3) is performed with the use of the *Mean Absolute Difference* (MAD)

$$\text{MAD} = \frac{\sum_{k=1}^{n} |O_{ki} - P_{ki}|}{n}, \quad (6)$$

where *n* is the number of *bins* (intervals) that cover the range of the *D*-scale (from 0 to 1), O_{ki} is the observed proportion of correct responses on item *i* for the examinees with *D* scores in bin *k*, and P_{ki} is the average probability of correct item response for the examinees with *D* scores in bin *k* (this probability is estimated via Equation 3, with D_s being the midpoint of bin *k*). Typically, the *D*-scale is divided into 10 bins (n = 10), with the range of each bin equal to 0.1, but other approaches to 'binning' can be used to make sure that there are enough examinees in each bin. In a simulation study on testing for item fit under the *D*-scoring model, Dimitrov and Luo (2017) found that performs well with a cutting score of 0.07. Specifically, with MAD ≥ 0.07 indicating item misfit, the Type I error rate is 0.019 (i.e., 1.9% chances that a fit item is signaled as misfit), whereas the Type II error rate is 0.061 (i.e., 6.1% chances that a misfit item is signaled as fit).

3. Method

3.1 Data

The data in this study come from the scores of teachers on four different Test Forms (TFs) of the GTT, referred to here as test forms TF1, TF2, TF3, and TF4, administered by the NCA in year 2016. The teachers come from 30 universities in Saudi Arabia. By test forms, (a) TF1 was taken by 52,160 teachers (64.7% males, 35.3% females), (b) TF2 was taken by 40,585 teachers (79.9 % males, 20.1% females), (c) TF3 was taken by 50,841 teachers (64.2 % males, 35.8% females), and (d) TF4 was taken

by 33,316 teachers (97.8 % males, 2.2% females).

3.2 Variables and Measures

Each test form (TF1, TF2, TF3, and TF4) consists of 75 dichotomously scored items (1 = correct response, 0 = incorrect response). These items are associated with four content domains of the GTT, namely (a) *professional knowledge*, 36 items, (b) *enhance learning*, 18 items, (c) *support learning*, 10 items, and (d) *professional responsibility*, 11 items. There are 15 anchor (common) items in any pair of adjacent test forms in the following sequence for linking the four test forms: TF4 → TF3 → TF2 → TF1, where TF1 is the target (base) test form. It should be noted that the anchor items in TF2 with the pair (TF2, TF1) are not the same as the anchor items in TF2 with the pair (TF2, TF3). Likewise, the anchor items in TF3 with the pair (TF3, TF2) are not the same as the anchor items in TF3 with the pair (TF3, TF4).

3.3 Statistical Analysis

In line with the purpose of this study, the analysis relates to examining key psychometric features of the GTT in the framework of the *D*-scoring model. Specifically, the consistency of such features across four test forms of the GTT (TF1, TF2, TF3, and TF4) is examined in terms of score reliability, *D*-score and δ-distributions, Item-Person Maps (IPMs), item fit (MAD values), errors associated with *D*-scores, and representativeness of the anchor items in test forms. The computations are performed with the use of customized modules incorporated in the computer program SATSE (Atanasov & Dimitrov, 2016) used with assessments at the NCA.

4. Results

4.1 Reliability Across Test Forms

As the *D*-score of a person is a linear combination of the binary scores (1/0) on the test items, X_i (i = 1, ..., n), of that person, the reliability of *D*-scores is the same as the reliability of the reliability of the raw score on the test (number correct responses) (Dimitrov, 2016). In this study, the reliability, ρ, of test scores was estimated via the Latent Variable Modeling (LVM) approach using the computer program Mplus (Muthén & Muthén, 2010) (e.g., Dimitrov, 2012, pp. 186-188; Raykov, 2007; Raykov, Dimitrov, & Asparouhov, 2010). By Test Forms (TFs), the LVM estimates of ρ, with their 95% confidence interval (CI), were found to be (a) for TF1, ρ = 0.868, with 95% CI = (0.866, 0.869), (b) for TF2, ρ = 0.865, with 95% CI = (863, 868), (c) for TF3, ρ = 0.847, with 95% CI = (0.845, 0.850), and (d) for TF4, ρ = 0.854, with 95% CI = (0.852,0.857).

4.2 δ-Values Across Test Forms

The stability of δ-values across test forms (TF1, TF2, TF3, and TF4) is examined in terms of statistics such as range, mean, standard deviation, and correlation between δ-values of the anchor items for any two adjacent test forms in the linking sequence TF4→TF3→TF2→TF1. The results are provided with Table 1. As can be seen, these statistics are very stable across the test forms for both the entire test form

and the anchor items. The mean of δ-values is close to 0.5 in all cases, thus indicating an average difficulty of the test forms and their subsets of anchor items. Also, the correlations between the δ-values of the anchor items are very high for all pairs of adjacent test forms, ranging from 0.969 to 0.988. As illustrated in Figure 3 for the pair of test forms TF1 and TF2, the δ-values of their anchor items are almost equal. The same holds for all other pairs of adjacent test forms. Thus, the comparisons across test forms remain valid without the need for rescaling the δ-values across test forms.

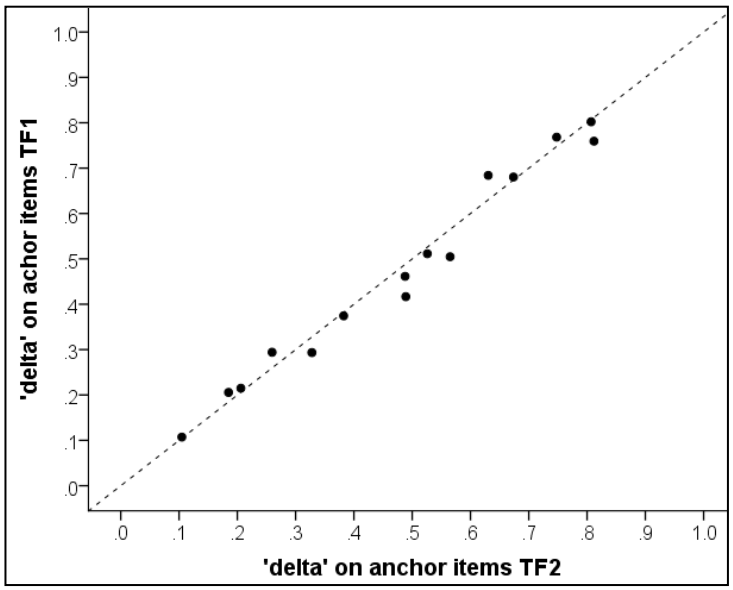

Figure 3. Relationship between δ-Values of 15 Anchor Items for Test Forms TF1 and TF2 of the GTT (correlation *R* = 0.970)

Table 1. Descriptive Statistics for Item Difficulty, δ, across Four Test Forms and Their Anchor Items

Test form	*n*	*min*	*max*	*Mean*	*SD*	*R*
TF1	75	0.068	0.948	0.514	0.223	NA
anchors with TF2	15	0.224	0.894	0.580	0.203	0.970
TF2	75	0.025	0.911	0.489	0.212	NA
anchors with TF1	15	0.224	0.911	0.606	0.224	0.970
anchors with TF3	15	0.105	0.812	0.472	0.226	0.988
TF3	75	0.106	0.930	0.533	0.199	NA
anchors with TF2	15	0.107	0.802	0.571	0.210	0.988
anchors with TF4	15	0.321	0.942	0.535	0.210	0.969
TF4	75	0.028	0.942	0.541	0.184	NA
anchors with TF3	15	0.290	0.942	0.535	0.221	0.969

Note. R = correlation between the δ-values of anchor items of two test forms (NA = not applicable).

The "anchors" are common item between adjacent test forms in the linking sequence TF4 →TF3→TF2→TF1, with TF1 being the "base" test form

4.3 D-Scores Across Test Forms

As shown in the previous section, the practical equality of δ-values for the anchor items in any pair of adjacent test forms allows for valid comparisons of *D*-scores across the test forms. The results in Table 2 and Figure 4 show that the distribution of *D*-scores in terms of shape, range, mean, and standard deviation is quite stable across the test forms TF1, TF2, TF3, and TF4.

Thus, one can treat the study groups of teachers who took these forms as practically equivalent in ability measured by the GTT. With this, the mean of the *D*-scores being close to 0.40 across all test forms indicates that on average the study sample of teachers (176,902 teachers on all four test forms) demonstrated about 40% of the ability required for total success on the GTT.

Table 2. Descriptive Statistics for *D*-Scores across Four Test Forms

Test form	*N*	*min*	*max*	*Mean*	*SD*
TF1	52,160	0.000	0.888	0.390	0.105
TF2	40,585	0.000	0.809	0.419	0.099
TF3	50,841	0.000	0.826	0.394	0.097
TF4	33,316	0.000	0.800	0.396	0.104

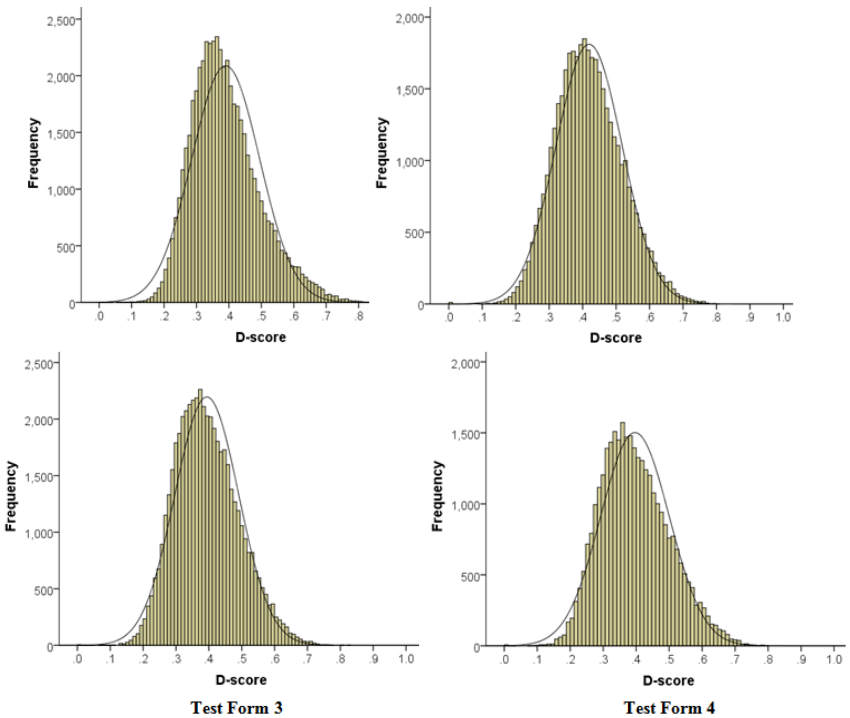

Figure 4. Distribution of *D*-Scores on Four Test Forms of the GTT

4.4 D-Score Errors Across Test Forms

The errors associated with *D*-scores are evaluated here with their standard error, $SE(D_s)$ (see Equation 5), and the correlation between the *D*-scores and their true values, $E(D_s)$ (see Equation 4). The results are presented in Table 3 and depicted in Figures 5 and 6. As given in Table 3, the mean $SE(D_s)$ is stable and very small (about 0.05) across all four test forms. From a different perspective, this finding is supported with very high (almost perfect) positive correlations between the *D*-scores and their true values (see $R_{D,TD}$ in Table 3 and Figure 6).

Table 3. Descriptive Statistics for the Standard Error of D-Scores on Four Test Forms

Test form	*N*	*min*	*max*	*Mean*	*SD*	$R_{D,TD}$
TF1	52,160	0.000	0.055	0.048	0.005	0.999
TF2	40,585	0.000	0.058	0.053	0.004	0.997
TF3	50,841	0.000	0.055	0.050	0.004	0.997
TF4	33,316	0.000	0.055	0.051	0.004	0.996

Note. $R_{D,TD}$=correlation between *D*-scores and their true values, *TD* = $E(D_s)$.

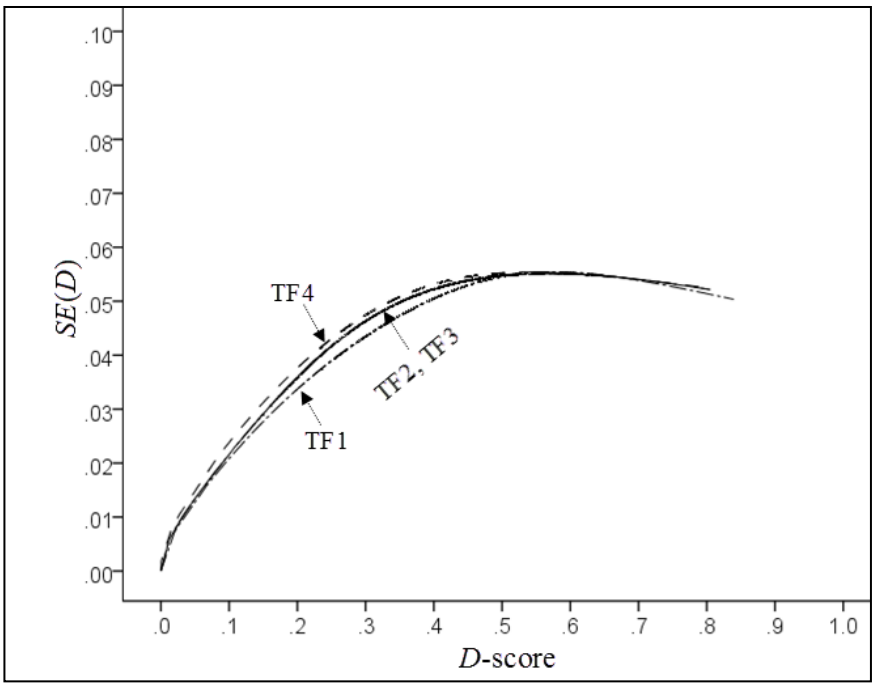

Figure 5. Standard Error of *D*-Scores, *SE*(*D*), on Four Test Forms of the GTT (TF1, TF2, TF3, TF4)

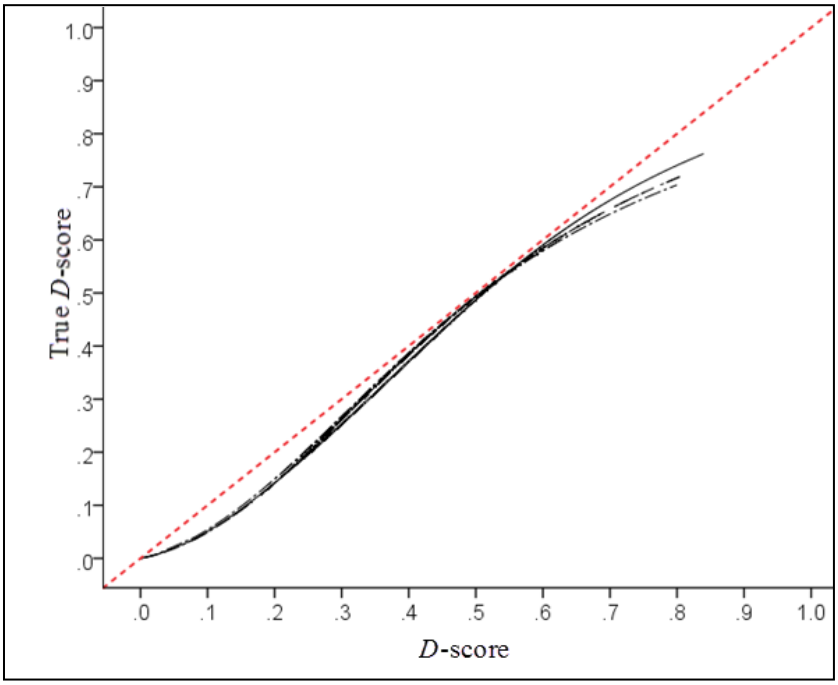

Figure 6. *D*-Sores vs True *D*-Scores on Four Test Forms of the GTT

4.5 Item-Person Maps Across Test Forms

As noted earlier, the Item-Person Map (IPM) provides information about the match between the item difficulties, δ-values, and examinees' ability levels on the *D*-scale. The IPMs depicted in Figure 7 show that there is a relatively good overlap between the range of examinees' *D* scores and the range of expected item difficulties, δ_i (i = 1, ..., 72), with the nature of this overlap being consistent across the test forms TF1, TF2, TF3, and TF4. That is, there is a consistent and similar match between examinees' ability measured by the GTT and the item difficulties across the four test forms. However, one may also notice that there are items with difficulty higher than 0.75 ($\delta > 0.75$) on the *D*-scale, but there are no examinees with ability scores in that range. At the same time, there is no enough items with difficulties in the range from, say, 0.3 to 0.4 on the *D*-scale, whereas the highest frequency of examinees is within this range. Therefore, test developers may decide to use more items with difficulty between 0.3 and 0.4 at the expense of items with difficulties higher than 0.75 on the *D*-scale.

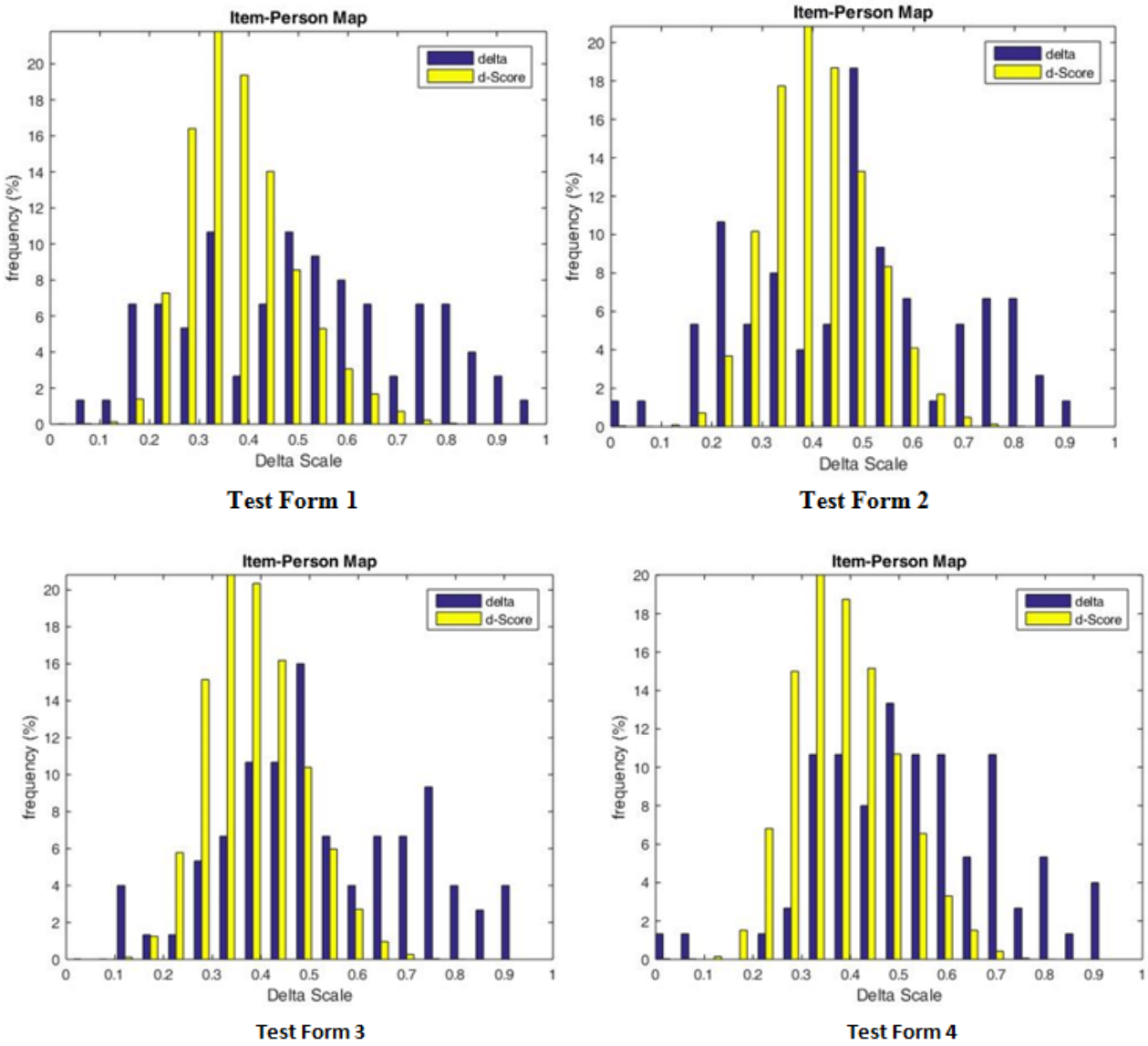

Figure 7. Item-Person Map (IPM) for Four Test Forms of the GTT

4.6 Item Fit across Test Forms

As described in the Method section, item fit under the *D*-scoring model is evaluated with the use of the MAD statistic, with MAD $\geq$ 0.07 indicating item misfit. Recall that MAD is the mean absolute difference between the observed proportions of correct responses on the item and the probabilities of correct item response under the 2PLR model in Equation 3. For illustration, the estimates of item discrimination, *a*, and item location, *b*, under this model for the items in the base test form, TF1, are provided in Table 4 (the δ-values are also provided for completeness of item parameters). The MAD values for test forms TF1, TF2, TF3, and TF4 are given in Tables 5 and 6. As can be seen, the number of misfitting items is (a) three items in TF1, (b) six items in TF2, (c) three items in TF1, and (d) one

item in TF4. It is important to note that there are no item misfits among the anchor items (numbers in bold and followed by the letter "A" in Tables 5 and 6). The graphical display of item fit is illustrated for two items in TF1 shown in Figure 8, with a misfit of item 6 (MAD = 0.0903) and a good fit of item 18 (MAD = 0.0246).

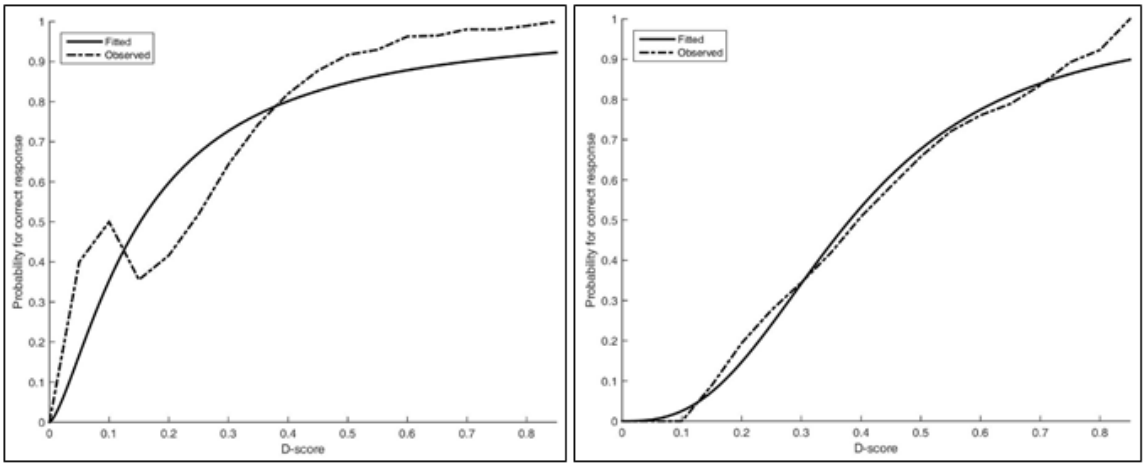

Figure 8. **Observed Proportions (Dotted Line) vs Theoretical Probability (Solid Line) of Correct Iitem Response (IRF under the 2PLR Model) in Test Form, TF1, (a) Left Panel: Item 6 (Misfit, MAD = 0.0903), and (b) Right Panel: Item 18 (Good Fit, MAD = 0.0246**)

Table 4. Item Parameters for Test Form TF1 under the 2PLR Model of IRF

		IRF: 2PLR				IRF: 2PLR	
Item	δ	*a*	*b*	Item	δ	*a*	*b*
1	0.5270	1.2569	0.4952	**39**	0.5845	1.4501	0.5546
2	0.6051	3.6000	0.4574	**40**	0.5310	2.5018	0.4158
3	0.9013	4.7815	0.9900	**41**	0.5168	1.3743	0.455
4	0.6655	3.3576	0.5098	**42**	0.7902	4.5289	0.6160
5	0.5442	3.2128	0.4128	**43**	0.5817	1.4509	0.5522
6	0.4952	0.9147	0.4775	**44**	0.5251	2.4641	0.4042
7	0.9339	5.0000	0.7992	**45**	0.1259	2.1084	0.1474
8	0.2988	2.5760	0.2667	**46**	0.6813	2.5713	0.5420
9	0.6104	2.5335	0.4636	**47**	0.2178	1.7750	0.2092
10	0.4656	2.1760	0.3647	**48**	0.3425	1.1280	0.2182
11	0.6760	3.4937	0.5156	**49**	0.2879	2.6856	0.2638
12	0.0682	1.5093	0.0693	**50**	0.3949	2.0380	0.2982
13	0.3951	2.2020	0.3056	**51**	0.3072	1.6038	0.2423
14	0.9485	5.0000	0.9900	**52**	0.3524	3.0130	0.3124
15	0.3268	1.5791	0.2408	**53**	0.7890	4.2143	0.6329
16	0.5891	3.3146	0.4452	**54**	0.4869	2.7710	0.3856
17	0.4724	2.3809	0.3718	**55**	0.8145	1.8699	0.9900
18	0.4820	2.7252	0.3814	**56**	0.2156	2.3808	0.2176
19	0.5457	3.1427	0.4148	**57**	0.2240	1.6981	0.1930
20	0.5444	2.7027	0.4260	**58**	0.2047	2.5565	0.2261
21	0.7862	2.5909	0.7231	**59**	0.3394	1.4651	0.2428
22	0.5127	3.0640	0.3874	**60**	0.2239	2.5347	0.2279
23	0.6347	2.3070	0.5140	**61**	0.2628	2.4804	0.2462
24	0.6972	1.1450	0.8571	**62**	0.5125	2.2431	0.3857
25	0.8937	3.1822	0.9900	**63**	0.2791	1.8645	0.2431
26	0.4259	2.8972	0.3476	**64**	0.1787	1.5703	0.1560
27	0.5393	2.7698	0.4070	**65**	0.8472	5.0000	0.6418
28	0.4283	1.1295	0.3255	**66**	0.3412	2.0086	0.2870
29	0.4577	1.9613	0.3324	**67**	0.5459	1.1170	0.4959
30	0.8155	5.000	0.5884	**68**	0.1715	2.1884	0.1738
31	0.3194	1.7937	0.2711	**69**	0.2021	1.4413	0.1522
32	0.5255	2.5742	0.3940	**70**	0.3302	2.5910	0.2863
33	0.7489	5.0000	0.5423	**71**	0.5928	4.4453	0.4437

34	0.3204	2.7636	0.2878	**72**	0.1783	1.2076	0.1142
35	0.7506	3.7171	0.5801	**73**	0.6626	2.1858	0.5587
36	0.8829	5.0000	0.9900	**74**	0.7933	2.5534	0.7016
37	0.7429	5.0000	0.5337	**75**	0.8256	5.0000	0.6445
38	0.7108	5.0000	0.5127				

Table 5. MAD Values for Item Fit in Test Forms TF1 and TF2

	Test Form 1			**Test Form 2**			
Item	MAD	Item	MAD	Item	MAD	Item	MAD
1	.0393	39	.0378	1	.0266	**39A**	.0177
2	.0466	**40A**	.0262	2	.0458	40	.0354
3	.0541	41	.0306	3	.0230	41	.0223
4	.0455	**42A**	.0571	4	.0151	**42A**	.0370
5	.0460	43	.0369	5	.0360	43	.0201
6A	.0699	**44A**	.0276	**6A**	.0230	**44A**	.0300
7	.0146	45	.0175	7	.0173	45	.0532
8	.0324	46	.0449	8	.0308	46	.0520
9A	.0639	47	.0444	**9A**	.0678	47	.0542
10	.0256	48	.0533	10	.0225	48	.0926*
11	.0445	49	.0499	11	.0621	49	.0418
12	.0093	50	.0387	12	.0341	50	.1424*
13	.0427	51	.0350	13	.0574	51	.0467
14	.0488	**52A**	.0363	14	.0352	**52A**	.0266
15	.0321	53	.0578	15	.0499	53	.0337
16	.0463	**54A**	.0218	16	.0495	**54A**	.0404
17	.0238	55	.0330	17	.0437	55	.0507
18	.0246	56	.0156	18	.0632	56	.0301
19	.0453	**57A**	.0299	19	.0928*	**57A**	.0236
20	.0317	58	.0260	20	.0519	58	.0401
21A	.0575	59	.0402	**21A**	.0454	59	.0281
22	.0563	60	.0371	22	.1066*	60	.0464
23	.0213	**61A**	.0282	23	.0769*	**61A**	.0236
24A	.0181	62	.0432	**24A**	.0106	62	.0287
25A	.0377	63	.0371	25	.0578	63	.0501
26	.0365	64	.0261	26	.0398	64	.0456
27	.0538	65	.0305	**27A**	.0226	65	.0330
28	.0568	66	.0293	**28A**	.0277	66	.0282
29	.0483	67	.0476	29	.0291	67	.0670
30A	.0340	68	.0312	30	.0211	68	.0308
31	.0241	69	.0903*	31	.0227	69	.0410
32A	.0489	70	.0314	**32A**	.0451	70	.0236
33	.0611	71	.0365	33	.0846*	71	.0252

34	.0443	72	.0295	34	.0633	72	.0516
35	.0476	73	.0164	35	.0394	73	.0219
36	.0972*	74	.0291	36	.0213	74	.0421
37	.0415	75	.0621	37	.0360	75	.0363
38A	.0538			**38A**	.0588		

Note. A = anchor item for the two test forms; * misfitting item (MAD ≥ 0.07).

Table 6. MAD Values for Item Fit in Test Forms TF3 and TF4

	Test Form 3			**Test Form 4**			
Item	MAD	Item	MAD	Item	MAD	Item	MAD
1	.0618	39	.0652	1	.0172	39	.0451
2	.0120	40	.0483	2	.0391	40	.0188
3	.0232	41	.0217	3	.0201	41	.0201
4	.0404	**42A**	.0181	4	.0274	**42A**	.0333
5A	.0388	43	.0505	**5A**	.0304	43	.0299
6	.0459	**44A**	.0292	6	.0401	**44A**	.0298
7	.0298	45	.0256	7	.0280	45	.0669
8	.0237	46	.0355	8	.0230	46	.0405
9	.0498	47	.0209	9	.0257	47	.0417
10	.0421	**48A**	.0579	10	.0351	**48A**	.0330
11	.0736*	49	.0363	11	.0388	49	.0203
12	.0347	50	.0299	12	.0397	50	.0340
13	.0416	51	.0609	13	.0269	51	.0319
14	.0482	**52A**	.0514	14	.0377	**52A**	.0261
15	.0635	**53A**	.0567	15	.0594	**53A**	.0428
16	.0429	54	.0421	16	.0309	54	.0236
17	.0475	**55A**	.0548	17	.0771*	**55A**	.0312
18	.0440	56	.0243	18	.0437	56	.0215
19	.0396	**57A**	.0241	19	.0453	57	.0369
20	.0508	58	.0362	20	.0449	**58A**	.0372
21A	.0254	**59A**	.0181	**21A**	.0194	**59A**	.0336
22	.0350	60	.0428	22	.0249	60	.0383
23	.0269	61	.0470	23	.0112	61	.0360
24	.0495	62	.0306	24	.0175	62	.0369
25	.0343	63	.0498	25	.0405	63	.0225
26	.0463	64	.0611	26	.0246	64	.0344
27A	.0287	65	.0423	**27A**	.0176	65	.0376
28	.0467	**66A**	.0339	28	.0385	**66A**	.0300
29	.0282	67	.0582	**29A**	.0474	67	.0358
30	.0469	68	.1004*	30	.0392	68	.0451
31A	.0525	69	.0395	31	.0531	69	.0367
32	.0409	70	.0333	32	.0315	70	.0537
33	.0465	**71A**	.0298	33	.0405	**71A**	.0556

34	.0689	72	.0550	34	.0416	72	.0165
35	.0614	73	.0165	35	.0183	73	.0241
36A	.0554	74	.0301	**36A**	.0600	74	.0403
37	.0844*	75	.0812*	37	.0284	75	.0367
38	.0460			38	.0425		

Note. A = anchor item for the two test forms; * misfitting item (MAD ≥ 0.07).

5. Discussion and Conclusion

Assessments for teachers' certification are conducted by the NCA in Saudi Arabia with the use of multiple test forms of a *General Teacher Test* (GTT) and specialty tests in academic areas such as math, chemistry, physics, and so forth. The GTT, which is of interest in this study, is examined here on important psychometric features and their consistency across four test forms (TF1, TF2, TF3, and TF4). Dependable psychometric features and their generalizability across multiple test forms are the key to accuracy of the test results and their valid interpretations for the intended purposes of the assessment. It should be also noted that the psychometric analyses were conducted in the measurement framework of the *D*-scoring model which is under implementation in the assessment practice and research at the NCA (Dimitrov, 2016, 2017), Prior to the implementation of the *D*-scoring method in the assessment practice at the NCA, the analysis of test results were conducted in the framework of Item Response Theory (IRT). Therefore, the examination of the psychometric features of the tests, including the GTT, was also conducted in the framework of IRT (e.g., Dimitrov & Al-Sadaawi, 2014, 2015). For this reason, the results obtained in this study, related to psychometric features of the GTT under the *D*-scoring method, are not directly comparable to those obtained in previous studies using the IRT. It should be noted that in both scenarios (IRT and *D*-scoring) the psychometric features of the GTT were in support to its validity and reliability. However, unlike previous IRT-based studies of GTT, the present study provides valuable information about the stability of psychometric features across different test forms of the GTT, thus providing support to the generalizability aspect of the validity of GTT data.

There are several main findings that stem from the results in this study. First, the GTT scores on the *D*-scale are sufficiently accurate for the intended purposes of the test, as indicated by their high reliability and small errors of the *D*-score, which were found to be very stable across the four test forms examined in this study. Second, the difficulty of the test forms was moderate (close to average on the *D*-scale) and stable across the test forms in basic statistics (range, mean, and standard deviation) for both the entire test forms and the sets of anchor items for any pair of adjacent test forms in the adopted linking sequence TF4→TF3→TF2→TF1. Also, the δ-values of the anchor items are highly correlated and practically equal for each pair of test forms, which validates their comparison (and that of resulting *D*-scores) without the necessity for rescaling the δ-values (e.g., see Dimitrov, 2017). Third, the *D*-score distributions are very similar across the test forms in terms of shape, range, mean, standard

deviation, and correlations with their true values. Fourth, as indicated by the Item-Person Maps (IPMs) across test forms, there is a good match between item difficulties (δ-values) and examinees' ability measured by the GTT, with this trend being consistent across the test forms. Fifth, the number of misfitting items, under the *D*-scoring 2PLR model with Equation 3, is relatively small, with 3, 6, 3, and 1 misfitting items in test forms TF1, TF2, TF3, and TF4, respectively. It is important to note in this regard that none of the anchor items in the test forms is signaled for misfit. In conclusion, the findings in this study support the psychometric validity of the GTT for the intended use of this test in the framework of *D*-scoring adopted in the assessment practice at the NCA in Saudi Arabia. Also, the methodology used in this study can be useful to researchers in their work on validation of test data under the *D*-scoring model.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135. https://doi.org/10.1086/508733

Atanasov, D. V., & Dimitrov, D. M. (2016). *A System for Automated Test Scoring and Equating* (SATSE). Riyadh, Saudi Arabia: National Center for Assessment.

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. https://doi.org/10.1177/0013164416631100

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*. https://doi.org/10.1177/0013164417724187

Dimitrov, D. M. (2012). *Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields*. Alexandria, VA: American Counseling Association.

Dimitrov, D. M., & Al-Sadaawi, A. (2015). *New Teacher Test-General Science: Factor structure and psychometric Features* (National Center for Assessment, Research Report No. 2). Riyadh, Saudi Arabia: National Center for Assessment.

Dimitrov, D. M., & Al-Sadaawi, A. (2014). *New Teacher Test: Factorial structure and reliability* (National Center for Assessment, Research Report No. 2). Riyadh, Saudi Arabia: National Center for Assessment.

Dimitrov, D. M., & Luo, Y. (2018). *Testing for Item Fit under the D-Scoring Model* (Technical Report: TR-4-2018). Riyadh, Saudi Arabia: National Center for Assessment.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26. https://doi.org/10.1214/aos/1176344552

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589-612. https://doi.org/10.1111/ecca.12002

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Raykov, T. (2007). Evaluation of weighted scale reliability and criterion validity: A latent variable modeling approach. *Measurement and Evaluation in Counseling and Development, 40*(1), 42-52. https://doi.org/10.1080/07481756.2007.11909805

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling, 17*, 265-279. https://doi.org/10.1080/10705511003659417