A Standard Setting Method Using the *D*-scoring Method: Procedures and Application to Assessment for Teacher Certification

Dimiter M. Dimitrov, Ph.D.
(email: ddimitro@gmu.edu)
and
Abdullah Al-Sadaawi, Ph.D.
(email: alsadaawi@gmail.com)

National Center for Assessment
Riyadh, Saudi Arabia

Abstract

This paper presents an approach to deriving cutting scores for standard setting for large-scale standardized assessments. The approach is developed in the framework of the *D*-scoring method (DSM; Dimitrov, 2018, 2019) which is adopted at the National Center for Assessment (NCA) in Saudi Arabia and gaining attention in the field of educational assessment. Under the DSM, the test score of an examinee is based on his/her response pattern on dichotomously (1/0) scored items weighted by their expected difficulty for the target population of examinees. The DSM is a classical framework, providing transparency and computational simplicity, but it also has some measurement features analogous to those in item response theory (e.g., models of item response functions). The *D*-score of an examinee indicates what percent of the ability required for total success on the test is demonstrated by that examinee. The idea behind the proposed approach is to directly align the outcome of expert judgments with the computation of cutting scores on the *D*-scale. A panel of experts on the test content and targeted standards are required to identify response patterns of binary (1/0) item scores that reflect mastery of targeted standards measured by the test. To facilitate the experts in their judgments on this task, they are provided with the content of the test items and their expected difficulty precalibrated for a norm population of examinees. Based on such response patterns, cutting scores for standards setting are computed via a simple DSM procedure. In another scenario, when expert judgments are not available, cutting scores are identified based on a norm distribution of scores on the *D*-scale. The proposed approaches are illustrated with real data from assessments for teacher certification at the NCA.

Key words: standards setting, cutting scores, *D*-scoring method

A Standard Setting Method Using the *D*-scoring Method: Procedures and Application to Assessment for Teacher Certification

A key element in commonly used procedures for standard setting on assessments is the expert's internalization of the performance standards and the conceptualization of the borderline examinee (e.g., Cizek & Bunch, 2007). In the popular bookmark method (Mitzel, Lewis, Patz, & Green, 2001), which is considered as a better alternative to the Angoff (1971) method, experts set their cutting scores by placing a marker between groups of items arranged by increasing difficulty in an ordered item booklet (OIB). The question the experts must answer for each test item is " Is it likely that the *minimally qualified* (or *border line*) examinee will answer this item correctly ?" With this question, the 'likelihood' is usually set at 67% (or a 2/3 chance of correct response). The item immediately following the bookmark is considered to be the first item of a proficiency level and the last item in the OIB a borderline examinee of that level is likely to answer correctly. Along with the strengths of the bookmark method (e.g., direct link to psychometric data), there are serious doubts among researchers about the experts' conceptualization of key concepts and understanding of the bookmark procedure. For example, Davis-Becker, Buckendahl, and Gerrow (2011) compared cutting score results of experts using OIBs with results of experts placing bookmarks in test forms where the items were randomly ordered by difficulty, and they found similar recommendations on cutting scores under both conditions.

Without taking sides in debates on the bookmark method or comparative claims, the approach outlined in the present paper is targeting clarity in the conceptualization of key concepts involved in standard-setting procedures and computational simplicity in placing cutting scores on the *D*-scale (Dimitrov, 2018, 2019). Specifically, presented next is an approach to computing cutting scores for setting standards in two scenarios, (a) mastery, based om experts' judgments, and (b) levels of proficiency (e.g., low, medium, high), based on a norming statistical distribution on the *D*-scale. Depending on the purpose of testing and subsequent decisions, any of these scenarios, or both, can be of interest to stake holders in the respective assessment. This is an initial effort to organize standard-setting procedures that are in line with the simplicity and practical efficiency of the *D*-scoring method (DSM; Dimitrov, 2018, 2019) which is under adoption at the National Center for Assessment (NCA) in Saudi Arabia.

Method

Computation of *D*-scores

Under the 'delta-scoring' (*D*-scoring) method (DSM), the *D*-score of an examinee on a set of test items (the entire test or, say, a content domain) is based on his/her response vector of (1/0) item scores and the expected difficulties of the items, δ_i. It should be noted that $\delta_i = 1 - \pi_i$, where π_i is the expected 'easiness' of the item; that is, the expected proportion of correct item responses for a targeted population of examinees). Specifically, once the δ_i values are estimated for a test of *n* binary items (e.g., via bootstrapping), the *D*-score of person *s* on the test is computed as follows

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

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; that is, $X_{si} = 1$; $i = 1, ..., n$. The *D*-score can be interpreted as the proportion (%) of the ability needed for a total success on the test demonstrated by the examinee. The same interpretation holds when Equation 1 is used with subsets of test items grouped, say, by content domains, thus allowing for valid comparisons of the examinees' performance on the entire test and its content domains.

Equating Test Forms on the *D*-scale

For validity of using given cutting scores across different test forms, the *D*-scores on such forms should be on the same scale. Although a detailed discussion on equating test forms on the *D*-scale is beyond the scope of this paper, some brief notes on this matter deserve attention. Two test forms are referred to here as '*delta-equivalent'* (δ-equivalent) is they have the same number of items and the distributions of their δ-values are identical. In the testing practice of the NCA, delta-equivalent test forms are generated via the *System for Automated Test Assembly on the D-scale* (SATA-*D*; Atanasov & Dimitrov, 2019a). Specifically, SATA-*D* assembles δ-equivalent test forms from an existing item bank, where the δ-values of all items are placed on the same scale, and the user of SATA-*D* specifies a targeted distribution of δ-values on the *D*-scale which is the same for all assembled test forms according to the purpose of the specific test. In other scenarios of testing at the NCA, when the test forms are not assembled via SATA-*D*, their equating is based on using common items among the test forms and performed by using the computer program DELTA (Atanasov & Dimitrov, 2019b).

Cutting Scores for Mastery Based on Experts' Judgments

Under the proposed method, the test items are grouped by content and/or test objectives (e.g., content domains or performance standards) and, if appropriate, such item groups are further divided into meaningful subgroups (e.g., subdomains or substandards), referred to here as 'units' of mastery. Then panel of experts are asked to identify a response vector of binary (1/0) item scores that they consider as providing evidence for 'mastery' of the respective test unit. The identification of a '***response vector for mastery***' (RVM) on each unit "automatically" provides the RVM for mastery of the respective higher-level groups of items and the entire test. In this way, the experts are asked to work in small steps on RVM units, by marking the items that they consider as providing evidence (if answered correctly) for mastering the respective unit. The experts are provided with the content of the items and their expected difficulties, δ_i. It is important to emphasize that the δ_i values are known a priori, via precalibration of test items for a norming population of examinees, and are available, say, from an existing item bank. Such an item bank is developed, for example, at the NCA where the δ_i estimates of the items are obtained via bootstrapping and rescaled across test forms on a common scale for the entire item bank.

An illustration of the computation of cutting scores on the *D*-scale for a hypothetical test of 20 binary items grouped into four domains is provided in Table 1. The δ_i values are computed for simulated data used in this example. In this case, each domain is used as RVM unit (that is, no further grouping of items by aspects of the domain is available). The D_E is used to suggest that the computation of the cutting score on the *D*-scale is based on judgements of 'experts.' As shown in Table 1, the *D*-cutting score for mastery of Domain 1 ($D_{E1} = 0.54$) is based on using Equation 1 with the RVM for this domain of six item (1 0 1 1 0 1) and their δ_i values, assuming that the RSV is identified by a panel of experts; (that is, the experts believe that the correct responses on items 1, 3, 4, and 6 provide a minimally requited evidence of mastery of the domain). Thus, an examinee with a score higher than 0.54 on the *D*-scale for the first domain (*D*

> 0.54) will be classified in the group of 'mastery' for that domain. The same procedure is used for the computation of cutting scores for mastery of the other three domains. Finally, the sequence of RVMs by domains forms the RVM for the entire test which results in a cutting score of 0.504 (or D_E = 50 on a *D*-scale from 0 to 100).

Adjusted Cutting Scores for Mastery

In some scenarios the experts can be overly "conservative" (too demanding) in their expectations for 'mastery' thus producing unrealistically high cutting scores. In such cases it might be more appropriate to adjust the expert-based cutting score, D_E, toward the average difficulty of the items in the unit under consideration. Specifically, if $\bar{\delta}$ is the mean of the δ values for the item in the test unit, the adjusted cutting score, denoted here D_a, is the midpoint between the expert-based cutting score, D_E, and the difficulty of the test unit; that is,

$$D_{\text{a}} = \left(D_E + \bar{\delta}\right)/2. \quad (2)$$

For illustration, Table 2 provides the adjusted cutting scores obtained via Equation 2 using the expert-based cutting scores, D_E, and the average difficulty of the respective unit (the entire test and each domain) presented in Table 1 with the example in the previous section. An examinee with a given *D* score is assigned to (a) *Mastery*, if $D \geq D_a$, and *No-mastery*, if $D < D_a$.

Four levels of mastery. One can further refine the two levels of mastery (Mastery vs. No-mastery) by using two additional cutting scores, denoted here D_L (for 'low' level) and D_H (for 'high' level), obtained as the midpoints between the adjusted cutting score, D_a, and left-end and right-end values on the *D*-scale, 0 and 1, respectively. That is,

$$D_L = (D_a + 0)/2, \text{ and } D_H = (D_a + 1)/2. \quad (3)$$

Thus, four levels of mastery are defined in increasing order, (a) *Nomastery-1*, if $D < D_L$, (b) *Nomastery-2*, if $D_L \leq D < D_a$, (c) *Mastery-1*, if $D_a \leq D < D_H$, and (d) *Mastery-2*, if $D \geq D_H$. A colored visual presentation of the levels of mastery described in this section is given in Table 3.

Cutting Scores for Mastery Levels Based on a Norm Distribution of *D*-scores

In some cases the standard-setting goal is to place examinees into levels of performance (e.g., low, medium, high) under a norm distribution of scores on the *D*-scale. In typical large-scale assessments at the NCA, the norm distribution is a **normal distribution on the *D*-scale** (*Mean* = 0.50, *SD* = 0.1666) obtained for a representative 'norm' sample of examinees on a test form identified as a ***base form*** of the test of interest. When scaled *D*-scores are used (from 0 to 100), the norm *D*-scale distribution is a normal distribution with *Mean* = 50 and *SD* = 16.66.

Three levels of performance. When three levels of performance (low, medium, and high) are targeted, the best cutting scores under a normal distribution and the 27th and 73rd percentile; that is P_{27} and P_{73} (Kelly, 1937). For the norm *D*-scale distribution, *N*(*Mean* = 50, *SD* = 16.66), we have $P_{27} \approx 40$ and $P_{73} \approx 60$. Therefore, with cutting scores 40 and 60 on the D-scale from 0 to 100, an examinee with a given D score is assigned to a level of performance (a) *low*, if $D < 40$, (b) *medium*, if $40 \leq D < 60$, and (c) *high*, if $D \geq 60$.

Four levels of performance. The above three levels of performance can be further refined by splitting the medium level into 'below average' and 'above average' levels. Given the mean of the *D*-scale (= 50), an examinee with a given *D* score is assigned to a level of

performance (a) low, if $D < 40$, (b) below average, if $40 \leq D < 50$, (c) above average, if $50 \leq D < 60$, and (d) high, if $D \geq 60$.

Note. When different forms of a test are used, which is usually the case in large-scale assessments, their *D*-scores have to be equated to the scale of the 'norm' distribution and then compare their equated values to the cutting scores (40, 50, 60) to place the respective examinees into levels of performance. In the assessment practice of the NCA test equating on the *D*-scale can be directly achieved by using the computer program *DELTA* (Atanasov & Dimitrov, 2019b) or the system *STATA-D* for automated assembly of 'delta-equivalent' test forms (Atanasov & Dimitrov, 2019a).

REAL-TEST EXAMPLE

The purpose of this example is to illustrate the derivation of cutting scores for levels of mastery using expert-based judgments on a teacher certification test developed at the NCA, referred to as ***General Teacher Test*** (GTT). This test consists of 75 binary (1/0) scored items associated with four content domains as follows:

Domain 1: *Professional knowledge* (36 items**).** This domain focuses on the knowledge that teachers need to plan for quality student learning opportunities; how do teachers help the students' learning in the discipline(s) that they teach ; and the curriculum and the resources that they provide to support student learning. Planning for learning includes the knowledge which is necessary to meet the standards in the other test domains.

Domain 2: *Promoting learning* (19 items). This domain describes the practices of effective teachers and the opportunities provide for student learning. It focuses on classroom engagement and the learning that teachers promote in their students, as well as the assessment practices to monitor student learning and provide helpful feedback. This domain emphasizes that teachers are responsible for promoting learning and developing the curriculum that they are expected to teach.

Domain 3: *Supporting learning* (9 items). This domain focuses on an inclusive social environment of trust and respect, and an intellectually challenging environment with high expectations for learning and achievement. It is based on the idea that effective teachers establish a classroom environment that supports student learning. As the previous domain, this domain of supporting learning also focuses on teacher practices.

Domain 4. *Professional responsibilities* (11 items). This domain relates to teachers' professional responsibilities outside the classroom, namely (a) to establish a productive relationship with parents, (b) to contribute to effective school functioning, (c) to evaluate their own practice and engagement in professional learning, (d) to report on student progress, and (e) to fulfill other responsibilities in the school.

The items in each content domain are further grouped into targeted standards for teacher candidates, but for space consideration their substantive connotation is not described here (see Tables 4 and 5). In this case, each of the five experts participating in this study were asked to provide a *response vector for mastery* (RVM) for each standards separately. The RVMs by standards generate the RVMs by domains and, eventually, for the entire test. After working independently, the experts were asked to discuss their RVM judgements as a group and to come up with an unanimously accepted decision on the RVMs by standards and, thus, by content domains and the entire test. The resulting RVMs are shown (in red color) in Table 4. The expert-based cutting scores associated with those RVMs, D_E, and their adjusted values, D_a, obtained via Equation 2, are provided in Table 5. Furthermore, using Equation 3, one can obtain additional cutting scores, D_L and D_H, for setting four mastery level (not shown her for space consideration).

Conclusion

The proposed approach to deriving cutting scores for mastery levels on large-sale assessments provides simplicity, transparency, and direct relations between expert-based judgements and the computation of cutting scores on the D-scale. The initial studies on this approach using some standardized tests at the NCA indicate its promising dependability and relevance to the content and difficulty of the test items that facilitates the judgments of experts in their work by "steps" from smaller test units to larger sets of items grouped by targeted criteria (e.g., standards and/or domains). It should be noted, however, that this approach is still in a stage of piloting, refinement, and possible modifications to reflect more adequately the specificity of the test structure and consequential validity of the classifications based on the respective cutting scores.

References

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., & Dimitrov, D. M. (2019a). *SATA-D: A System for Automated Test Assembly on the D-scale*. National Center for Assessment, Riyadh, Saudi Arabia.

Atanasov, D. V., & Dimitrov, D. M. (2019b). *DELTA: A Computer Program for D-scoring and Equating of Test Data (V. 1.0)*. National Center for Assessment, Riyadh, Saudi Arabia.

Cizek, G. J.,& Bunch, M. B.(2007). *Standard stetting: A guide to establishing and evaluating performance standards on test*. Thousand Oaks, CA: Sage.

Dimitrov, D. M. (2018). The delta-scoring method of tests with binary items: A note on true score estimation and equating. *Educational and Psychological Measurement, 78*(5) 805–825. (first published online, 2017: DOI: 10.1177/0013164417724187).

Dimitrov, D. M. (2019). Modeling of item response functions under the *D*-scoring method. *Educational and Psychological Measurement* (published online first, DOI: 10.1177/0013164419854176).

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R.(2001). *The bookmark procedure: Psychological perspectives*. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods and perspectives (pp.249–281).Mahwah, NJ: Lawrence Erlbaum.

Table 1. *Computing Cutting Scores on the D-Scale for 'Mastery/Nonmastery': An Illustration for a Hypothetical Test of 20 Items Grouped in Four Domains*

Domain	Item	Response vector for mastery (RVM)	δ	Computation of the cutting *D*-score	*D*-cut score
1	1	**1**	0.583	$D_{E1} = \frac{\delta_1+\delta_3+\delta_4+\delta_6}{\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+\delta_6}$	0.559 (or 56)
1	2	**0**	0.641		
1	3	**1**	0.274		
1	4	**1**	0.449		
1	5	**0**	0.567		
1	6	**1**	0.224		
2	7	**0**	0.314	$D_{E2} = \frac{\delta_8+\delta_9+\delta_{11}}{\delta_7+\delta_8+\delta_9+\delta_{10}+\delta_{11}+\delta_{12}}$	0.450 (or 45)
2	8	**1**	0.364		
2	9	**1**	0.422		
2	10	**0**	0.551		
2	11	**1**	0.615		
2	12	**0**	0.844		
3	13	**1**	0.435	$D_{E3} = \frac{\delta_{13}+\delta_{15}+\delta_{17}}{\delta_{13}+\delta_{14}+\delta_{15}+\delta_{16}+\delta_{17}}$	0.622 (or 62)
3	14	**0**	0.565		
3	15	**1**	0.314		
3	16	**0**	0.253		
3	17	**1**	0.596		
4	18	**1**	0.259	$D_{c4} = \frac{\delta_{18}+\delta_{20}}{\delta_{18}+\delta_{19}+\delta_{20}}$	0.578 (or 58)
4	19	**0**	0.702		
4	20	**1**	0.702		
TOTAL Test	$D_E = \frac{\delta_1+\delta_3+\delta_4+\delta_6+\delta_8+\delta_9+\delta_{11}+\delta_{13}+\delta_{15}+\delta_{17}+\delta_{18}+\delta_{20}}{\delta_1+\delta_2+\cdots+\delta_{18}+\delta_{20}}$				.504 (or 50)

***Notes*:** 1. **δ** = expected item difficulty (for a target population of examinees).

2. For each domain, content experts have come to agreement that answering correctly the highlighted items is sufficient for "mastery" of the domain.

3. Cutting *D*-scores are computed separately for each domain and then for the entire test Equation 1 with the RVM resulting from the RVMs developed for the four domains of the test.

5. The cutting scores can be multiplied by 100 (and rounded to the nearest integer) for presentation on the *D*-scale from 0 to 100. For example, the cutting score for 'mastery' on the entire test (D_c = 0.504) can be reported as D_c= 50. That is, an examinee should demonstrate at least 50% of the ability necessary for total success on the test in order to be assigned to the category of 'mastery.'

Table 2

Expert-Based Cutting Scores (D_E) and their Adjusted Values (D_a) for the Results in Table 1

Test Unit	D_E	*Scaled* D_E (0-100)	*Mean* (δ), $\bar{\delta}$	D_a	Scaled D_a (0 – 100)
Entire Test	0.504	50	0.456	0.480	48
Domain 1	0.559	56	0.518	0.538	54
Domain 2	0.450	45	0.433	0.442	44
Domain 3	0.622	62	0.554	0.588	59
Domain 4	0.578	58	0.484	0.531	53

Note. $D_a = (D_E + \bar{\delta})/2$.

Table 3

Levels of Mastery Based on Adjusted Cutting Score on the D-scale

No Mastery (NoM)		**Mastery (M)**	
NoM-1st level	**NoM-2nd level**	**M-1st level**	**M-2nd level**
0	D_L	D_a	D_H **1**

Table 4

General Teacher Test: Response Vector for Mastery (RVM) by Standards-Domains-Test

item	Domain	Standard	RVM	δ	item	Domain	Standard	RVM	δ
1	1	1	1	.3452	37	2	6	1	.7179
2	1	1	0	.3685	38	2	6	1	.5983
3	1	1	0	.5281	39	2	6	1	.5902
4	1	1	1	.2403	40	2	6	1	.6440
5	1	1	1	.7379	41	2	6	1	.4487
6	1	1	1	.6672	42	2	6	1	.4787
7	1	1	1	.7570	43	2	6	0	.3841
8	1	1	0	.7220	44	2	6	0	.5558
9	1	2	1	.2868	45	2	**7**	1	.8499
10	1	2	1	.4743	46	2	**7**	1	.5602
11	1	2	1	.6848	47	2	**7**	1	.3880
12	1	2	0	.1249	48	2	**7**	0	.7183
13	1	2	0	.5753	49	2	**7**	1	.6042
14	1	2	0	.2350	50	2	**7**	0	.3296
15	1	2	1	.7565	51	2	**7**	1	.3489
16	1	2	1	.4668	52	2	**7**	0	.7824
17	1	2	0	.8198	53	2	**7**	0	.7457
18	1	2	1	.5682	54	2	**7**	0	.2577
19	1	2	1	.6721	55	2	8	0	.0714
20	1	2	1	.6232	56	3	8	1	.4824
21	1	4	0	.5919	57	3	8	0	.2487
22	1	4	1	.6895	58	3	8	1	.1179
23	1	4	1	.4572	59	3	8	1	.6863
24	1	4	0	.5584	60	3	8	1	.4870
25	1	4	0	.4851	61	3	9	1	.1293
26	1	4	1	.6218	62	3	9	1	.4328
27	1	4	1	.4900	63	3	9	0	.1237
28	1	4	1	.4971	64	3	9	1	.4296
29	1	4	0	.1240	65	4	10	0	.8577
30	1	4	0	.4863	66	4	10	1	.2137
31	1	4	0	.7604	67	4	10	0	.6694
32	1	5	0	.6822	68	4	10	0	.5518
33	1	5	1	.7496	69	4	11	1	.7335
34	1	5	1	.7396	70	4	11	0	.2357
35	1	5	0	.8971	71	4	11	0	.6988
36	1	5	1	.5719	72	4	12	0	.1095
					73	4	12	1	.6198
					74	4	12	0	.0362
					75	4	12	1	.8892

Table 5
Cutting scores on GTT (on a D-scale 0-100)

Test/Domain/Standard	Items	Expert-based cutting score, D_E Scaled (0-100)	Average difficulty, $\bar{\delta}$ Scaled (0-100)	Adjusted cutting score, D_a Scaled (0-100)
Entire Test	**1-75**	**60**	**52**	**56**
Domain 1: *Professional knowledge*	**1-36**	**60**	**56**	**58**
Standard 1	1-8	63	54	58
Standard 2	9-20	72	52	62
Standard 3	NA			
Standard 4	21-31	48	52	50
Standard 5	32-36	57	73	64
Domain 2: *Promoting learning*	**37-55**	**62**	**53**	**58**
Standard 6	37-44	79	55	67
Standard 7	45-54	49	56	54
Domain 3: *Supporting learning*	**56-64**	**88**	**35**	**64**
Standard 8	55-60	85	35	60
Standard 9	61-64	89	28	59
Domain 4: *Professional responsibilities*	**65-75**	**44**	**52**	**48**
Standard 10	65-68	9	57	34
Standard 11	69-71	44	56	50
Standard 12	72-75	91	41	66

Note. Standard 3 is ‘not applicable’ (NA) in this procedure for cutting scores for the GTT.