

An Application of the Response Vector for Mastery Method of Standard Setting in Comparison with the Extended-Angoff and Cluster Analysis Methods

Ayşenur ERDEMİR* Dimiter DIMITROV** Esra OYAR*** Serpil ÇELIKTEN DEMIREL****

Abstract

This paper presents a real-data application of a recently developed method of standard setting, referred to as the response vector for mastery (RVM) method, and compares its performance with the Extended-Angoff method and cluster analysis method (CAM). The RVM enables the derivation of cutoff scores on the bounded D-scale (ranging from 0 to 1), under the D-scoring method of measurement, and the logit scale of item response theory (IRT). Using a 25-item multiple-choice test in the field of educational measurement, the study demonstrates that the RVM method offers key advantages over Extended-Angoff and CAM methods, including reduced judgmental burden for panelists, greater classification accuracy and consistency, and flexibility in computing cutoff scores at the domain (content area) level. Given its novel structure and promising empirical performance, the RVM method holds strong potential for researchers and practitioners seeking practical, defensible, and scalable approaches to standard setting in educational assessment contexts. These qualities are reflected in the findings, as the RVM method produces stable and consistent classification decisions based on response patterns, supporting defensible decision-making while minimizing expert burden and remaining scalable for large-scale applications.

Keywords: *D-scoring method of measurement, item response theory, standard setting, cutoff scores, classification accuracy, classification consistency.*

Introduction

The standard setting in educational assessment involves a challenging process of determining cutoff scores on assessment instruments in order to classify examinees into standards-based performance levels, such as “pass/fail,” “mastery/nonmastery,” or “minimal, basic, proficient, and advanced.” There are two main steps in this process, namely (a) a policymaking task of establishing standards that represent mastery of learning objectives and levels of proficiency, and (b) a technical task of deriving cutoff scores for classification of examinees into targeted standards-based levels of proficiency (e.g., Kane, 2001; Ricker, 2006). To stay in line with the terminology in the literature, methods for deriving cutoff scores are referred to here as standard-setting methods; however, the conceptual distinction between policy standards (establishing proficiency expectations) and performance standards (deriving cutoff scores) should be kept in mind. In accordance with the rules specified by the selected standard-setting method, cutoff scores are determined based on the assessments of content experts (panelists) who are typically selected based on their subject-

* Asst. Prof. Police Academy, Institute of Forensic Sciences Ankara, Türkiye, erdemiraysenur@gmail.com, ORCID ID: 0000-0001-9656-0878

** Prof. Dr., Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia, d.dimitrov@etec.gov.sa, ORCID ID: 0000-0003-1256-4842

***Res. Asst. Dr., Gazi University, Faculty of Education, Ankara, Türkiye, esratas@gazi.edu.tr, ORCID ID: 0000-0002-4337-7815

**** Asst. Prof., Dicle University, Faculty of Education, Diyarbakır, Türkiye, sserpilcelikten@gmail.com, ORCID ID: 0000-0003-3868-3807

To cite this article:

Erdemir A., Dimiter D., Oyar E., & Çelikten Demirel S. (2026) An Application of the response vector for mastery method of standard setting in comparison with the Extended-Angoff and Cluster Analysis methods. *Journal of Measurement and Evaluation in Education and Psychology*, 17(2), 63-82. <https://doi.org/10.21031/epod.1759258>

Received: 8.08.2025

Accepted: 17.03.2026

matter expertise and familiarity with the assessment context and are often provided with methodological training prior to the standard-setting process. These methods are not only technical tools but are deeply embedded in educational policy and practice, as the resulting classifications often determine students' access to further opportunities, accountability consequences for institutions, and perceptions of fairness and merit in educational systems.

The approaches to setting cutoff scores for standard setting are classified as test-centered, examinee-centered, and data-driven approaches (e.g., Jaeger, 1989; Kane, 1994; Sireci et al., 1999). These broad approaches differ mainly in their primary focus and purpose in linking assessment results to performance standards. Under test-centered approaches, standard-setting panelists examine the items to align item characteristics with targeted performance expectations. Based on this examination, they make decisions about targeted performance levels using criteria such as (a) the probability that a borderline proficient test taker will answer an item correctly (e.g., the Angoff and Bookmark methods), or (b) the comparison of item response demands with performance level descriptors (e.g., the ID matching method and RVM method). Under *examinee-centered* approaches, instead of focusing on items, the standard-setting panelists consider examinees with the purpose of identifying cutoff scores based on observed performance distributions (e.g., contrasting group method and borderline group method; Cizek, 1996; Jaeger, 1989; Kane, 1994). For instance, under the contrasting group method, the panelists select two groups of examinees: (a) upper group – the group considered to be above the targeted standard, and (b) lower group – the group below the standard. The test score that produces the lowest rate of misclassifications (false-positive or false-negative) is selected as a cutoff score. The *data-driven* approaches to setting cutoff scores are based on statistical procedures with the purpose of identifying latent performance groupings directly from test score patterns, such as cluster analysis, latent class analysis, and mixed Rasch analysis (e.g., Binici & Cuhadar, 2022; Brown, 2007; Gündeğer & Doğan, 2018; Keuning et al, 2017; Khalid, 2011; Peabody & Wind, 2019; Sireci et al, 1999).

Within these broad categories, most popular approaches to setting cutoff scores are the *Angoff's method* (1971) and various modifications of it (e.g., Plake & Cizek, 2012; Clauser et al, 2006), *bookmark method* (e.g., Cizek, 2001; Karantonis & Sireci, 2006; Lewis et al., 1999; Lewis et al., 2012; Mitzel et al., 2001), *body of work method* (e.g., Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Wyse et al., 2014;), *benchmark method* (Phillips, 2012), *item-descriptor (ID) method* (Ferrara et al, 2008), other methods (Zwick et al., 2001), and a recently developed approach referred to as *Response Vector for Mastery (RVM) method* (Dimitrov, 2022, 2023). The literature widely acknowledges that no single (“best”) approach exists for establishing cutoff scores across diverse assessment contexts and standard-setting policy frameworks (e.g., Alshawwa, 2023; Dimitrov, 2022; Zieky, 2001). Therefore, examining and comparing standard-setting methods under real-world constraints contributes not only to measurement research but also to institutional decision-making and policy formation.

There are numerous studies comparing standard-setting methods (e.g., Brennan & Lockwood, 1980; Buckendahl et al., 2002; Chang, 1999; Çetin & Gelbal, 2013; Özarkan & Doğan, 2020; Sireci et al, 1999; Strickman, 2017; Thomas et al, 2018; Yim, 2018; Yousefi, 2021). Most studies compare test-centered methods, such as the popular Angoff's and Bookmark methods (e.g., Buckendahl et al., 2002; Çetin & Gelbal, 2013; Yim, 2018; Yousefi, 2021), or modifications of the Angoff's method under different conditions (e.g., Brandon, 2004; Jalili et al, 2011; Kara & Cetin, 2020; Ricker, 2006). Fewer studies deal with comparisons involving examinee-centered and data-driven standard-setting methods (e.g., Hess et al, 2007; Sireci et al, 1999; Thomas et al, 2018). However, while many studies focus exclusively on statistical comparisons, this study also considers aspects such as method flexibility, ease of implementation, and potential alignment with domain-level content — factors relevant to practical assessment use.

Purpose of the Study

The aim of this study is to examine the effectiveness of the Response Vector for Mastery (RVM) method as a standard-setting approach by comparing its resulting classification decisions with those obtained from the Extended-Angoff method and cluster analysis.

The *response vector for mastery* (RVM) standard setting method is a test-centered method designed to operate in the classical and latent frameworks of the *D*-scoring method of measurement (DSM) and item response theory (IRT) (Dimitrov, 2022, 2023). This method has been successfully used for setting standards in the context of national large-scale assessments in Saudi Arabia, such as the National Assessment for Schools (NAFS), the National Program for Talented and Gifted Students (Mawhiba) Tests (Dimitrov et al, 2020; Sayar et al, 2024).

Methodological issues with the popular Bookmark method, in comparison with the RVM method, are discussed by Dimitrov (2022). In a practical application, the RVM method was evaluated against the Bookmark method using the scales of the *Multiple Cognitive Ability Tests* (MCAT), which are utilized for admissions to Saudi Arabia's Mawhibah programs for brilliant and gifted children (Dimitrov et al, 2020). The standard setting panelists preferred the RVM method, considering it more comprehensive, easier to understand, and inclusive of all test items, whereas they considered the Bookmark method as more difficult to apply and ignoring items that fall after the “cutting item” marked in the ordered item booklet (OIB) specified under this method (e.g., Lewis et al., 2012). Nevertheless, to the best of our knowledge, no thorough research has been done on how well the RVM approach performs in contrast to other standard-setting techniques.

As an effort in this direction, the purpose of this study is to compare the real-data performance of the RVM method with a modification of the popular extended-Angoff’s method (Angoff, 1971) and the data-driven method of cluster analysis (Sireci et al, 1999). The goal is not only to evaluate their comparative performance on accuracy and consistency metrics, but also to reflect on their potential utility and defensibility in educational assessment practice. In line with this aim, the study examines how the test cut-score is determined using different standard-setting methods, namely the Extended-Angoff method, the Cluster Analysis Model (CAM), and the Response Vector for Mastery (RVM) method. In addition, the study investigates the extent to which mastery and nonmastery decisions derived from these methods converge by comparing their classification agreement, accuracy, and consistency.

These methods were selected to represent different standard-setting paradigms: the Extended-Angoff method, a test-centered approach largely based on expert evaluations; cluster analysis, a data-driven technique; and the RVM method, an approach that includes not only expert opinion but also values derived from data in the weighting process for determining the cut-off score. This distinction allows for a meaningful comparison of methods that differ in the scope and nature of expert participation.

Theoretical Framework

Briefly described next are the standard setting methods examined here for congruency on classifications of examinees into proficiency levels: the RVM method, which is of primary interest here, an Extended-Angoff’s method, and a cluster analysis method.

Angoff Method

Under the Extended-Angoff model, the setting of cutoff scores is conducted in two stages. First, the participating panelists are asked to assess the probability that a minimally competent examinee (borderline) would answer each item correctly. The total of these probabilities across all test items, serving as an estimate of the ‘true’ score for the borderline examinee, is set as the cutoff score. Following this initial scoring, panelists engage in structured discussion to review their judgments.

In the second stage, the panelists are shown the resulting cutoff scores that they have produced and asked to reevaluate their judgments so that a final agreement (and corresponding cutoff score) is reached. The issues associated with the Angoff method arise from (a) experts’ difficulties in accurately conceptualizing a borderline examinee, and (b) their probability judgments regarding the likelihood of a correct response to each item by the borderline examinee (Busch & Jaeger, 1990; Hambleton & Plake, 1995; Ricker, 2006). For ease of reference, the Extended-Angoff model employed in this study will be referred to as the Angoff method from this point onward.

Cluster Analysis Method

The *cluster analysis method* (CAM) of setting cutoff scores is a data-driven method designed to avoid subjective judgments of panelists in the process of cutoff score derivations (Sireci et al, 1999; Hess et al, 2007; Khalid, 2011; Thomas et al, 2018). The K-means clustering procedure of CAM produces homogeneous clusters of examinees. For each examinee, the distance between the examinee and the center (centroid) of the cluster is computed, and the examinee is assigned to the cluster with the shortest distance (Anderberg, 1973). There are two main problems with using cluster analysis in order to calculate cutoff scores. First, the CAM cutoff scores are based on specific sample data, not on criterion-based standards. Second, CAM produces clusters, but it is not known whether these clusters are truly distinct performance groups of examinees. Therefore, the validity of the clusters should be examined using an external criterion, which refers to a separate variable that is not part of the clustering process and is used to verify meaningful performance differences among clusters (Sireci et al, 1999).

RVM method

Under the *response vector for mastery* (RVM) method of setting cutoff scores (Dimitrov, 2022), panelists involved in the standard setting process are not required to make probability judgments, unlike in the Angoff and Bookmark methods. Rather, test items are organized into response vector units (RVUs) based on substantial norms (e.g., content domains or standards), and panelists are instructed to identify the items they think are adequate (when the response is correct) to demonstrate mastery of the corresponding response vector unit. To perform this task, panelists match item response demands with descriptors of mastery levels for each RVU. The resulting RVMs by units and, thus, for the entire test, allow for computation of cutoff scores directly (by RVUs and for the entire test) on the *D*-scale or IRT logit scale. In both instances, the calculation of cutoff scores relies on the respective RVM and the parameters of the items it includes, in the context of either IRT or DSM. The RVM method offers a distinct advantage in that it enables panelists to operate at the content-domain level, which can be pedagogically meaningful for formative and summative assessment use. By guiding panelists to base their judgments on observable item features and mastery criteria, the RVM method may reduce some of the ambiguity inherent in probability-based approaches. Naturally, the RVM method's validity, developed by experts, relies on how clear and complete the item response requirements and the 'mastery' descriptors are—a crucial task for the effectiveness of any standard-setting method (e.g., Egan et al., 2009; Ferrara et al, 2009; Skorupski & Hambleton, 2005). Provided next are the computation formulae for the methods of (a) classical *D*-scoring, (b) latent *D*-scoring, and (c) IRT-based scoring (Dimitrov, 2022, 2023).

Classical D-scoring. Suppose that the RVM (for one test unit, RVU, or the entire test) consists of a sequence of binary scores, X_i , on n items ($X_i = 1$, if item i is selected by the panelists as necessary for mastering the respective RVU, and $X_i = 0$, otherwise). Then the cutoff scores on the scale of "weighted" classical *D*-scores, D_w , (ranging from 0 to 1; e.g., see Dimitrov, 2023) are computed as follows:

$$D_{w_cut} = \frac{\sum_{i=1}^n \delta_i X_i}{\sum_{i=1}^n \delta_i}, \quad (1)$$

where δ_i is the *expected item difficulty*; that is, $\delta_i = 1 - \pi_i$, where π_i is the "easiness" of the item (i.e., proportion of correct item responses for the target examinee groups). Bootstrapping method (Efron, 1979) can be used to estimate δ_i and is available in basic statistical programs (e.g., R and SPSS), or (b) $\hat{\delta}_i = 1 - p_i$, where p_i is the sample proportion of correct item responses if the examinee group is large enough (e.g., $N > 3,000$).

Latent D-scoring. The DSM model may be conceptualized as an IRT formulation on a bounded scale, with *D*-scores constrained between 0 and 1. Similar to IRT, item and person parameters are simultaneously estimated within the selected DSM model using latent estimation methods such as Bayesian or maximum-likelihood estimation. For example, the analytic form of the *two-parameter rational function model* (RFM2) of latent *D*-scoring is

$$P_i(X_i = 1|D, b_i, s_i) = \frac{1}{1 + \left[\frac{b_i(1-D)^{s_i}}{D(1-b_i)} \right]}, \quad (2)$$

where D is the person parameter (latent D -score), b_i is the *item difficulty* (location on the D -scale, where the probability of answering the item correctly equals 0.5), and s_i refers to the parameter for *shape* (discrimination) of the item. As a side note, the two-parameter logistic (2PL) model in IRT is mathematically equal to the RFM2 when the D -score and item difficulty, b_i , are both submitted to the logit transformation, keeping s_i the same for discrimination (e.g., Dimitrov, 2023, p. 111).

Using the DSM's item parameters (such as b_i and s_i of the RFM2), the cutoff score, D_c , that equals an RVM derived from panelists' judgments, is calculated as that RVM's maximum-likelihood estimator under latent D -scoring. Thus, D_c is the MLE solution of the log-likelihood function

$$\ln L(\mathbf{RVM}|D) = \sum_{i=1}^n \{X_i \ln[P_i(D)] + (1 - X_i) \ln[1 - P_i(D)]\}, \quad (3)$$

where the expert-based RVM of binary scores (1/0) of n test items is $\mathbf{RVM} = (X_1, X_2, \dots, X_n)$. The likelihood of the RVM is $L(\mathbf{RVM}|D)$, the natural logarithm is $\ln(\cdot)$, and the probability of correct item response estimated under the DSM model (e.g., RFM2) is $P_i(D)$.

IRT scoring. As a maximum-likelihood estimator of the expert-based RVM, item parameters of the IRT model (e.g., 2PL) are used to compute the RVM-based calculation of a cutoff score, θ_c , on the logit scale. In other words, θ_c is the log-likelihood function's MLE solution

$$\ln L(\mathbf{RVM}|\theta) = \sum_{i=1}^n \{X_i \ln[P_i(\theta)] + (1 - X_i) \ln[1 - P_i(\theta)]\}, \quad (4)$$

where $P_i(\theta)$ is the probability of a correct item response estimated within the context of an IRT model. For example, the 2PL model in IRT is (Hambleton et al, 1991):

$$P_i(\theta) = \frac{1}{1 + \exp[\alpha_i(\beta_i - \theta)]}, \quad (5)$$

where θ is the person ability, α_i is the item discrimination, and β_i is the item difficulty (location on the logit scale where the probability of correct item response equals 0.5).

Methods

Instrument and Data

In this study, the standard-setting methods RVM, Angoff, and CAM were examined for congruency on classification of examinees into “mastery/ nonmastery” levels of performance on a *measurement and evaluation achievement test* (MEAT). Within the scope of this study, conducted using a MEAT, mastery refers to examinees whose response patterns indicate sufficient proficiency to meet the established performance standard, whereas nonmastery represents examinees whose response patterns fall below this threshold. These classifications are operationalized through the cut scores determined by each standard-setting method.

The test items were developed and reviewed by two subject-matter experts using a table of specifications aligned with the course content covered during the academic term, ensuring appropriate coverage of the targeted content domains. The test consists of 25 multiple-choice items grouped into three content domains: (a) *Basic concepts*: five items, (b) *Tools in measurement and evaluation in education*: 10 items, and (c) *Test and item analysis*: 10 items. Information related to the psychometric evaluation of the test scores is presented in the Results section.

The test was administered to a random sample of 340 students enrolled in a certification program (pedagogical formation/teacher-training) in Türkiye. Monte Carlo simulation studies by Weiss and Minden (2012) indicated that, with a sample size of 200, parameter estimates obtained from the 2PLM were close to the true values in terms of mean, root mean square error (RMSE), and bias when test lengths were 25 and 50. In a similar manner, the use of empirical data by Akour and Al-Omari (2013) yielded the finding that a sample size of 200 produced acceptable parameter estimates under the 3PLM when the test length exceeded 15 items. Based on these outcomes and taking into account the sample size and test length of the present study, a sample size of 200 was deemed adequate for the implementation of the 2PLM. In addition, the presence of items with moderate difficulty and a normally distributed thetas further supports the adequacy of the sample size for the analyses conducted in this study (DeMars, 2010). This setting provided a suitable context for comparing standard-setting methods, as the assessment was domain-specific, criterion-referenced, and used for certification decisions.

Data Structure and Reliability

The structural validity of the test data was examined in the framework of a confirmatory factor analysis (CFA) model with three latent factors associated with test items of the three content domains of the test. The reliability of the test data was assessed using two indices of internal consistency, the coefficient *alpha* (α) and the *omega* (ω) coefficient based on latent variable modeling (LVM; e.g., Bollen, 1989; Raykov, 2009; Dimitrov, 2012, p. 187).

As previously mentioned, under the respective latent model (e.g., RFM2 on the D-scale or 2PL on the IRT logit scale), estimations of the item parameters are used to calculate latent cutoff scores (on the IRT logit scale or D-scale). Therefore, testing for data fit of the model was also performed to ensure dependability of the item parameter estimates.

Standard-Setting Procedure

The standard-setting work with panelists (content experts) was conducted separately for the derivation of cutoff scores under the Angoff method and RVM method. In a study in the framework of generalizability theory, Hurtz and Hertz (1999) demonstrated that a group of 10 to 15 standard setting panelists is necessary for the derivation of cutoff scores under the Angoff method. In this study, the Angoff method procedure involved 12 qualified panelists with doctoral degrees in educational and psychological measurement and evaluation. In a study conducted by Dimitrov (2022), four panelists were considered sufficient for dependable derivations of cutoff scores under the RVM method of standard setting, so five experts were used in the RVM-based procedure in this study. Three of these five experts were part of the 12 experts used in the Angoff procedure to serve as common members of the two groups of experts. The Extended-Angoff and RVM methods were carried out in two stages as described in the previous section. The level of agreement among the panelists under these two methods is presented in the Results section.

Following the RVM methodology described in the theoretical framework section, the RVM was applied in steps as follows:

Step 1. The test items were grouped into three response vector units (RVUs) based on the three content subdomains of the test, (a) RVU1: Basic concepts, (b) RVU2: Tools used in measurement and evaluation in education, and (c) RVU3: Test and item analysis.

Step 2. Five panelists (content experts) were given a list of the test items grouped into three RVUs, with the list containing the consecutive numbering of the items and their description. They were also provided with descriptors of standards for demonstrating mastery' of each RVU.

Step 3. The panelists were instructed to work independently on each RVU and mark the items (circle their number in the list) that they consider sufficient (if answered correctly) for demonstrating 'mastery' on the respective RVU.

Step 4. In a templet (shown in Table 2), the marked items by a panelist were coded as "1" and unmarked items as "0" resulting a response vector for mastery (RVM) produced by that panelist (expert).

Step 5. All panelists were provided with the resulting five RVMs and instructed to work together, with focus on items with “disagreements” among them, and to come up with a final RVM (as shown in the last column in Table 2).

Step 6. Cutoff scores for ‘mastery’ were obtained using the RVM, obtained in Step 5, and estimates of item parameters obtained under the measurement model under consideration, (a) DSM-C: expected item difficulty, δ , (b) DSM-L: item location (difficulty), b , and item shape (discrimination), s , and (c) IRT: item discrimination, α , and item difficulty, β (e.g., see Table 3).

The cluster analysis method (CAM) was conducted using the K-means clustering procedure, also described in the previous section. As the purpose was to classify examinees into two (pass/fail) levels of performance, a two-cluster solution was used. After ranking the total scores of the participants in each cluster, the cutoff score was obtained by averaging the lowest score in the higher cluster and the highest score in the lower cluster.

Comparison of Classifications

The comparison of results related to the classification of examinees into categories of mastery/nonmastery, obtained under the three standard setting methods used in this study, was performed in two aspects. First, the Cohen’s kappa coefficient (Cohen, 1960) was used to assess the level of classification agreement produced by the cutoff scores computed under the three methods. The Cohen’s kappa statistic was computed, using the ‘psych’ package (Revelle, 2025) in R, to examine the classification agreements of interest. Second, indices of classification accuracy and classification consistency, obtained via the IRT-based Rudner’s method (Rudner, 2001, 2005), were compared for cutoff scores produced by the RVM method on three scales: the D -scale of DSM-C, DSM-L, and the IRT logit scale. *Classification consistency* refers to the extent to which examinees would be placed in the same performance categories across parallel versions of the same assessment, while *classification accuracy* refers to the extent to which the observed classifications align with the theoretically expected “true” classifications determined by cutoff scores on an assessment (Lee, 2010).

Under Rudner’s method, used in this study, a key element in the derivation of indices for classification accuracy and consistency is the expected probability of an examinee i to score in a performance category C , P_{iC} . This probability is given by the area corresponding to category C under a normal distribution defined by cutoff, the ability score of the examinee on the IRT logit scale, θ_i , and its standard error, σ_{θ_i} ; that is, $N(\theta_i, \sigma_{\theta_i})$.

These are organized into data vectors: vector of cut-scores, vector of thetas and vector of standard error of thetas. The first vector of cut-scores defining the boundaries of performance categories, $K = [K_1, K_2, \dots, K_{C+1}]$, which partitions the score scale into mutually exclusive performance levels. The second vector consists of examinees’ ability estimates obtained from the IRT model, $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{Ne}]^t$, and the third vector contains the corresponding standard errors associated with these ability estimates, $\hat{\sigma}_{\hat{\theta}_i} = [\hat{\sigma}_{\hat{\theta}_1}, \hat{\sigma}_{\hat{\theta}_2}, \dots, \hat{\sigma}_{\hat{\theta}_{Ne}}]$. Together, these vectors are used to model the uncertainty around examinees’ ability estimates and to determine the expected probability of classification into each performance category, which serves as the basis for computing indices of classification accuracy and classification consistency based on the formula of $\hat{P}_{iC} = \Phi(K_{Ci}, K_{Ci+1}, \hat{\theta}_i, \hat{\sigma}_{\theta_i})$. (for details, see Rudner, 2001, 2005).

For the estimation of classification accuracy and consistency on the D -scale, Rudner’s method was applied using the D -score and its standard error, $SE(D)$, of the examinee under the latent and classical DSM-estimations, DSM-L and DSM-C. These classification analyses based on the Rudner’s method were conducted in the R environment through the cacIRT R package (Lathrop, 2014). It is worth noting that the analytics of Rudner’s method imply that the index of classification consistency is always smaller in magnitude than the index of classification accuracy (Wyse & Hao, 2012).

Results

Data structure and reliability

The analyses were based on the data obtained from the measurement and evaluation achievement test (MEAT), administered to 340 examinees. Before the CFA, preliminary analyses were conducted. Within the scope of the CFA model, the assumption of normality was examined by evaluating skewness and kurtosis statistics standardized by their respective standard errors. The skewness value was very close to zero (skewness = 0.081, SE = 0.132), yielding a standardized value well within the ± 1.96 range, indicating no meaningful departure from symmetry. The kurtosis value (kurtosis = -0.746, SE = 0.264) resulted in a standardized value slightly exceeding this criterion; however, the magnitude of this deviation was modest. Since kurtosis is generally more sensitive to sampling variability than skewness, more tolerant thresholds can be considered when evaluating normality (Kline, 2023). Importantly, mild deviations in kurtosis are not considered problematic for CFA, particularly when skewness remains within acceptable limits and sample size is adequate. Overall, these findings support the assumption of approximate normality, suggesting that the data were suitable for CFA using conventional estimation methods. In addition to normality assessment, univariate outlier analysis was performed using standardized z values. The observed minimum and maximum z values ranged from -2.36 to 2.28, which are within commonly accepted thresholds, suggesting the absence of univariate outliers. Then, one-factor CFA analyses were carried out. The unidimensionality of the test data was supported by results from a confirmatory factor analysis (CFA) conducted using the computer program Mplus (version 8.11; Muthén & Muthén, 2024). The model fit indices indicated a very good data fit of a one-factor CFA model: $\chi^2(df = 275) = 353.870$, with a ratio $\chi^2/df = 1.28 (< 3)$, CFI = 0.932 (> 0.90), and RMSEA = 0.029 (< 0.08), consistent with recommended benchmarks (Hu & Bentler, 1999; see also Dimitrov, 2012, p. 135).

Testing for model-data fit of the two-parameter logistic (2PL) IRT model, used in this study, was performed to ensure dependability of the item parameter estimates (α_i, β_i) under this model. The 2PL model was preferred to the 3PL model in this study for two main reasons. First, due to problems associated with the use of a third (pseudo-guessing) parameter in the 3PL model, such as (a) lower estimation accuracy, and (b) the item difficulties are not comparable across items (Han, 2012; see also Dimitrov, 2023, p. 17). Second, the 2PL model provided a tenable model fit to the data in our study. The IRT assumption of local independence was supported by the value of the Yen's Q3 statistic computed for the study data ($Q3 < 0.3$) (Yen, 1984). The overall model-data fit of the 2PL model was supported by the following goodness-of-fit indices: CFI = 0.931 (> 0.90), SRMSR = 0.055 (< 0.06), and RMSEA = 0.034 (< 0.05). Regarding the testing for item-level fit, the ratio of the S- χ^2 fit statistic to its degrees of freedom (S- χ^2/df) was found to vary from 0.402 to 1.798 across all 25 test items, thus ensuring a good item fit according to the criterion S- $\chi^2/df < 3$ (e.g., Chernyshenko et al., 2007; Roberts, 2016). The related computations including Yen's Q3 statistic, general and item based fit indices, model and ability estimations were performed using the 'mirt' package in R. Regarding item fit on the D-scale, under DSM-C and DSM-L, the item-fit statistic *mean absolute difference* (MAD, which is $MAD < 0.07$ in this study) was found to be within the acceptable range for a good fit across all 25 test items (Dimitrov, 2023, p. 74). The computations were performed using the computer program DELTA for DSM analysis (Atanasov & Dimitrov, 2018).

The internal consistency reliability of the examinees' test scores was estimated using the popular Cronbach's coefficient *alpha* ($\alpha = 0.763$) and the latent variable modeling (LVM) coefficient '*omega*' ($\omega = 0.861$, 95% CI [0.839, 0.884]). As can be noticed, the LVM-based coefficient, ω , is higher than Cronbach's α . This is not a surprise given that ω does not require the Cronbach's α assumption of *tau-equivalent measures* (i.e., all items are equally correlated with the test's latent ability) (e.g., Dimitrov, 2012, pp. 186-188).

Derivation of Cutoff scores

This section presents the cutoff scores derived for classifying study participants into mastery and nonmastery categories using the three methods described in the previous section: *Angoff's method*, *the cluster analysis method (CAM)*, and *the response vector for mastery (RVM) method*. Using the Extended-

Angoff method, which involves an initial round of probability judgments followed by panel discussion and rescaling, the 12 panelists produced cutoff scores ranging from 41.6 to 59.2 on a 0–100 scale, with a mean of 51.00 and a standard deviation of 5.42 (see Table 1). Using the mean as a cutoff score, examinees with a score above 51 are classified in the category of ‘mastery’ on test performance

Table 1

Cut-off scores determined by experts using the Angoff method

Expert	Cutoff score	Expert	Cutoff score	Expert	Cutoff score
Expert-1	58.0	Expert-5	59.2	Expert-9	44.0
Expert-2	51.6	Expert-6	52.8	Expert-10	50.8
Expert-3	41.6	Expert-7	52.4	Expert-11	46.2
Expert-4	48.4	Expert-8	50.2	Expert-12	56.8
Mean cutoff score: 51.00					

CAM cutoff score. The K-means cluster analysis was performed on the total test scores presented on a scale from 0 to 100. The test scores of examinees in the resulting two clusters ranged (a) from 16 to 56 in the first (lower level) cluster, and (b) from 60 to 100 in the second (upper level) cluster. The cutoff score of 58 was obtained as the average of the highest score in the first cluster and the lowest score in the second cluster: $(56 + 60)/2 = 58$. Following the CAM approach, the lower and upper clusters were conceptualized as non-mastery and mastery groups, respectively, and the midpoint between them was treated as the mastery threshold.

RVM cutoff score. The derivation of cutoff scores under the RVM method is based on a response vector of mastery (RVM), produced by content experts (see Table 2), and estimates of item parameters obtained under the measurement model under consideration, (a) DSM-C: expected item difficulty, δ , (b) DSM-L: item location (difficulty), b , and item shape (discrimination), s , and (c) IRT: item discrimination, α , and item difficulty, β (see Table 3). As shown in Table 2, the test items are grouped into three content domains (response vector units, RVUs), and there are five response vectors (RVMs) produced by five content experts (by RVUs and for the entire test). The final response vector, obtained upon a second-round agreement among the five experts, is used for the derivation of cutoff scores under the DSM-C, DSM-L, and IRT estimates of item parameters presented in Table 3.

Table 2*Response Vector for Mastery (RVM) Produced by Five Standard Setting Panelists*

Test subdomain (RVU*)	Test item	Expert-based RVM					Agreement in 1st round (%)	Final RVM
		E1	E2	E3	E4	E5		
RVU1 Basic Concepts	1	1	1	1	1	1	100	1
	2	1	0	0	1	0	60	0
	3	1	1	1	1	1	100	1
	4	1	1	1	1	0	80	1
	5	0	1	0	0	1	60	0
RVU2 Tools used in Measurement and Evaluation in Education	6	1	1	1	1	1	100	1
	7	1	0	1	1	0	60	1
	8	0	0	0	0	1	80	0
	9	1	1	1	1	1	100	1
	10	1	1	1	1	0	80	1
RVU3 Test and Item Analysis	11	0	1	0	1	1	60	1
	12	0	1	0	0	1	60	0
	13	1	0	1	0	1	60	1
	14	0	0	0	0	0	100	0
	15	1	1	1	0	0	60	1
RVU3 Test and Item Analysis	16	0	0	1	1	1	60	1
	17	1	0	0	0	1	60	0
	18	1	1	1	0	1	80	1
	19	0	0	1	0	0	80	0
	20	1	0	0	1	1	60	1
	21	0	0	0	0	0	100	0
	22	1	1	1	1	0	80	1
	23	0	0	0	1	0	80	0
	24	1	0	0	1	1	60	1
	25	0	0	0	0	0	100	0

Note. The mean of the agreement rates (%) among the five experts (E1, ..., E5) on individual items at the first round of their work on RVM-based standard setting is 76.80%. At the second round, they came to an agreement on the final RVM, which was used for the computation of cutoff scores.

* RVU: Response Vector Unit

Table 3

Item Parameters Obtained Under IRT and DSM (Classical and Latent) Estimations

Item	IRT			DSM	
	β	α	δ	b	s
i1	-1.58	0.75	0.25	0.25	1.04
i2	-0.64	1.04	0.36	0.39	1.25
i3	-0.08	0.72	0.47	0.51	1.01
i4	-1.36	0.61	0.31	0.29	0.87
i5	3.03	0.23	0.69	0.78	0.64
i6	-1.33	0.67	0.31	0.29	0.91
i7	-0.80	1.11	0.33	0.37	1.47
i8	-2.01	0.84	0.18	0.18	1.07
i9	0.57	1.03	0.63	0.63	1.50
i10	-1.72	1.15	0.17	0.20	1.36
i11	0.31	0.95	0.56	0.58	1.26
i12	2.48	0.33	0.71	0.78	0.78
i13	-2.55	0.45	0.26	0.14	0.61
i14	0.05	1.26	0.52	0.53	1.74
i15	-1.06	0.24	0.43	0.39	0.46
i16	-0.09	1.00	0.50	0.51	1.42
i17	-0.68	0.53	0.44	0.42	0.92
i18	-0.78	1.38	0.30	0.35	1.48
i19	0.61	0.73	0.60	0.63	1.06
i20	0.65	0.65	0.58	0.63	1.00
i21	-0.73	1.04	0.36	0.37	1.23
i22	-0.65	1.41	0.33	0.38	1.48
i23	0.30	1.15	0.56	0.59	1.41
i24	-1.20	1.51	0.23	0.27	1.56
i25	0.22	1.20	0.56	0.57	1.52

Note. β : item difficulty (location) parameter in IRT,

α : item discrimination parameter in IRT,

δ : expected item difficulty parameter in DSM,

b : item difficulty (location) parameter in DSM, and

s : item shape (discrimination) parameter in DSM.

Under DSM-C, the computation of cutoff scores (for each RVU and the entire test), using Equation (1), is illustrated in Table 4. Under DSM-L, the RVM (DSM-L) cutoff score for the entire test is provided in Table 5. As shown in Equation 3, this cut score is computed as a maximum-likelihood estimate (MLE) of the RVM for the test using the estimates of item parameters (b and s) in Table 3 (for a computational code in R, see Dimitrov, 2023, p. 99). A graphical depiction of the MLE solution is provided in Figure 1. Using the IRT estimates of the item parameters in Table 3, the RVM (IRT) cutoff score, obtained via MLE solution

of Equation 4, appears in Table 5 as well. A graphical depiction of the IRT-based MLE solution for the cutoff score on the entire test is shown in Figure 2.

Table 4

Computation of Cutoff scores under the RVM (DSM-C) Method of Standard Setting

Item	RVM	δ	Cutoff score formula under RVM(DSM-C)	Cutoff score
i1	1	0.25	$D_{w.c1} = \frac{\delta_1 + \delta_3 + \delta_4}{\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5}$	0.495
i2	0	0.36		
i3	1	0.47		
i4	1	0.31		
i5	0	0.69		
i6	1	0.31	$D_{w.c2} = \frac{\delta_6 + \delta_7 + \delta_9 + \delta_{10} + \delta_{11} + \delta_{13} + \delta_{15}}{\delta_6 + \delta_7 + \delta_8 + \delta_9 + \delta_{10} + \delta_{11} + \delta_{12} + \delta_{13} + \delta_{14} + \delta_{15}}$	0.656
i7	1	0.33		
i8	0	0.18		
i9	1	0.63		
i10	1	0.17		
i11	1	0.56		
i12	0	0.71		
i13	1	0.26		
i14	0	0.52		
i15	1	0.43		
i16	1	0.50	$D_{w.c3} = \frac{\delta_{16} + \delta_{18} + \delta_{20} + \delta_{22} + \delta_{24}}{\delta_{16} + \delta_{17} + \delta_{18} + \delta_{19} + \delta_{20} + \delta_{21} + \delta_{22} + \delta_{23} + \delta_{24} + \delta_{25}}$	0.435
i17	0	0.44		
i18	1	0.30		
i19	0	0.60		
i20	1	0.58		
i21	0	0.36		
i22	1	0.33		
i23	0	0.56		
i24	1	0.23		
i25	0	0.56		
For the total test			$D_w = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_6 + \delta_7 + \delta_9 + \delta_{10} + \delta_{11} + \delta_{13} + \delta_{15} + \delta_{16} + \delta_{18} + \delta_{20} + \delta_{22} + \delta_{24}}{\delta_1 + \delta_2 + \delta_3 + \dots + \delta_{24} + \delta_{25}}$	0.532

Note. Cutoff scores on the *D*-scale (from 0 to 1).

Figure 1

MLE solution of the log-likelihood function in Equation 3 on the D-scale cutoff score

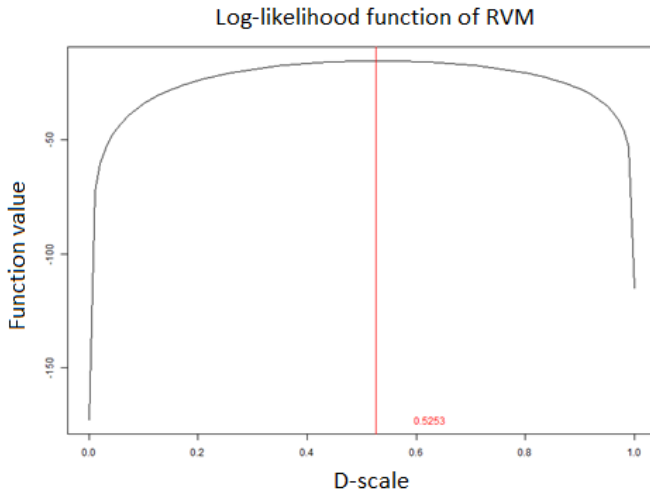
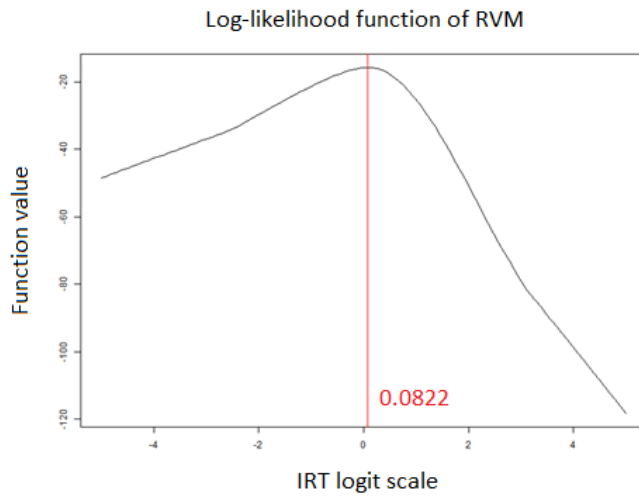


Figure 2

MLE solution of the log-likelihood function in Equation 4 on the IRT logit scale cutoff score



The cutoff scores under the Angoff, CAM, and RVM methods are provided in Table 5. The highest percentage of examinees classified into the category of ‘mastery’ (above the cutoff score, 62.35%) is produced by the Angoff cutoff score (51.00), the lowest percentage (39.41%) is produced by the CAM cutoff score (58.00), whereas the percentages produced by the RVM cutoff scores (under DSM-C, DSM-L, and IRT) are practically identical ($\approx 43\%$), with all three cutoff scores being slightly above the mean of their own scale, namely: 0.53 on the *D*-scale and 0.08 on the IRT logit scale.

Table 5*Cutoff scores Produced by the Angoff, CAM, and RVM Methods of Standard Setting*

Method	Number of items	Cutoff score	Percent examinees above the cutoff score
Angoff	25	51.00	62.35
CAM	25	58.00	39.41
RVM(DSM-C)	25	0.53	43.53
RVM(DSM-L)	25	0.53	43.24
RVM(IRT)	25	0.08	41.18

Note. CAM: Cluster Analysis Method; RVM: Response Vector for Mastery. The Angoff and CAM cutoff scores are scaled from 0 to 100, the RVM (DSM-C) and RVM (DSM-L) cutoff scores are on the *D*-scale (from 0 to 1), and the RVM (IRT) cutoff score is on the IRT logit scale.

Classification Agreement, Accuracy, and Consistency

The classifications of examinees into the categories of ‘mastery/nonmastery’ under the compared methods of standard setting were examined from the perspectives of classification agreement, accuracy, and consistency. The results are presented in Table 6. As shown, the agreement between the CAM and RVM methods (ranging from 0.90 to 0.92) is substantially higher than the agreement between the Angoff method and either the CAM or RVM methods (ranging from 0.56 to 0.64). The highest level of agreement (0.96) is associated with the comparison of RVM (DSM-L) with RVM (IRT), followed closely by the agreement between RVM (DSM-C) and RVM (DSM-L) (0.93).

Table 6*Cohen’s Kappa for Classification Agreement between the Angoff, CAM, and RVM Methods*

Compared Methods	Cohen’s kappa
Angoff and CAM	0.56*
Angoff and RVM (DSM-C)	0.64*
Angoff and RVM (DSM-L)	0.63*
Angoff and RVM (IRT)	0.59*
CAM and RVM (DSM-C)	0.92*
CAM and RVM (DSM-L)	0.91*
CAM and RVM (IRT)	0.90*
RVM (DSM-C) and RVM (DSM-L)	0.93*
RVM (DSM-C) and RVM (IRT)	0.89*
RVM (DSM-L) and RVM (IRT)	0.96*

* Statistically significant ($p < 0.01$)

As described earlier, indices of classification accuracy and classification consistency were computed using Rudner’s method (Rudner, 2001, 2005) for cutoff scores produced by the RVM method on the IRT logit scale and *D*-scale (DSM-C and DSM-L). The computations under Rudner’s method are based on estimates of the examinees’ scores and their standard errors on the respective scales, IRT logit scale and *D*-scale for DSM-L and DSM-C. The results are provided in Table 7. Rudner’s method is not applicable to the classical scoring scale (number of items correctly answered) used with the Angoff and CAM methods.

Table 7

Classification Accuracy and Consistency Produced by the RVM Method of Standard Setting

Method	Classification accuracy	Classification consistency
RVM(DSM-L)	0.88	0.83
RVM(DSM-C)	0.87	0.82
RVM(IRT)	0.87	0.82

Overall, the RVM cutoff scores derived on the *D*-scale (DSM-L) produced higher classification consistency and accuracy compared to RVM scores on the IRT logit scale. The highest classification accuracy and consistency are observed for the cutoff scores derived by the RVM method on the latent *D*-scale, RVM (DSM-L), with estimates of 0.88 and 0.83, respectively. The RVM under the classical *D*-scoring method, RVM (DSM-C), produced a classification accuracy of 0.87 and consistency of 0.82. The RVM under IRT estimations, RVM (IRT), produced similar indices of classification accuracy (0.87) and consistency (0.82). Taken together, the results provide a comprehensive comparison of the standard-setting methods, which is further discussed in the following section.

Discussion

This study presents a real-data application for the response vector for mastery (RVM) method, a recently developed method for working in the standard setting (Dimitrov, 2022, 2023, Chapter 10), and compares its performance with the well-known Extended-Angoff's method and the cluster analysis method (CAM). The RVM is used for the computation of cutoff scores on the *D*-scale (ranging from 0 to 1), in the framework of classical *D*-scoring (DSM-C) and latent *D*-scoring (DSM-L), as well as on the IRT logit scale, denoted as RVM (DSM-C), RVM (DSM-L), and RVM (IRT), respectively. Using cutoff scores derived from the real-data applications of the compared methods, the examinees (study participants) were classified into categories of 'mastery/nonmastery' on the respective scale.

The results on classification agreement between the compared approaches to deriving cutoff scores, based on the Cohen's kappa statistic, showed that the agreement between CAM and RVM applications is much higher than the agreements between the Angoff method and CAM or RVM methods. The results on classification agreement between the compared approaches to deriving cut-scores, based on the Cohen's kappa statistic, showed that the agreement between CAM and RVM applications was substantially higher than the agreement between the Angoff method and either CAM or RVM. This pattern of results may suggest that differences in classification agreement are partly related to the extent to which standard-setting approaches rely on empirical data, expert judgment, or a combination of both. Although RVM is sometimes characterized as a data-driven method, it may be more accurately described as a hybrid approach, in which item parameters are first estimated from empirical data and expert judgments are subsequently elicited based on these data-informed item characteristics. CAM, in contrast, derives cutoff scores directly from observed score distributions and thus operates at the score level without requiring expert input. The Angoff method, however, primarily relies on expert judgments, with relatively limited direct incorporation of empirical performance data. This difference in the degree to which empirical information informs the standard-setting process may help explain the lower levels of agreement observed between the Angoff method and the data-based approaches, compared to the higher agreement between CAM and RVM. The highest level of classification agreement is between the RVM (DSM-L) and RVM (IRT), followed closely by that between RVM (DSM-C) and RVM (DSM-L). This finding may be explained by the greater potential of score-based representations to capture examinees' response patterns with reduced measurement error, a pattern that has been frequently reported in studies comparing modern test theory with classical test theory. In particular, item response theory-based scoring approaches are known to provide more precise ability estimates by explicitly modeling item characteristics and individual response patterns (de Ayala, 2009; Embretson & Reise, 2000). From a different perspective, the RVM-based applications on the scales of DSM-C, DSM-L, and

IRT were compared on classification accuracy and consistency using Rudner's method (Rudner, 2001, 2005). The results revealed that the RVM cutoff scores obtained on the D-scale (DSM-L) produced higher classification consistency and accuracy compared to RVM cutoff scores obtained on the D-scale (DSM-C) and RVM scores on the IRT logit scale. The highest classification accuracy and consistency are produced by cutoff scores obtained under the RVM (DSM-L), followed by the RVM (DSM-C) and the RVM (IRT); yet, they all maintain a high level of accuracy and consistency, demonstrating the robustness of the RVM method across scoring models (Dimitrov, 2022). Importantly, the findings underscore that method selection in standard settings has broader educational consequences beyond psychometric precision. Especially in certification contexts, the choice of method can shape who is classified as proficient, how instructional goals are aligned, and what degree of transparency is ensured in the decision-making process.

From a validity perspective, the findings of this study can be interpreted within an evidence-based framework in which classification accuracy and classification consistency provide complementary evidence for mastery decisions. The accuracy of classification is indicative of the validity of mastery interpretations, as it reflects the extent to which examinees are correctly classified relative to a specified criterion. In contrast, classification consistency is more closely associated with reliability, indicating the reproducibility of these classifications (Lee, 2010). Consequently, the efficacy of standard-setting methodologies can be assessed in terms of decision validity and reproducibility by employing classification indices. The present study demonstrates that higher classification accuracy for RVM-based cut-scores, particularly within the framework of the DSM-L, supports the validity of distinguishing between mastery and non-mastery. Moreover, the findings reveal that consistently high classification accuracy and consistency across RVM applications serve as indicators of valid, reliable, and stable decisions.

The selection of cutoff scores has direct implications for mastery decisions and their educational consequences. The placement of cutoff may have a profound impact on the proportion of examinees classified as masters or non-masters, thereby influencing the risk of misclassification. The differences in classification outcomes are of particular significance in certification and accountability contexts, where decisions regarding mastery have the potential to influence progression, access to instruction, and credentialing. Furthermore, the findings indicated that RVM-based cutoff scores showed better performance under IRT estimation. IRT has a wide application in educational assessment, including large-scale international studies such as TIMSS. The findings of this study demonstrate that the RVM method yielded high classification accuracy and consistency values under IRT, thus suggesting that these cut-scores can be effectively utilised in educational decision-making.

To conclude, the results of this study indicate that the RVM standard setting method has several advantages over the Angoff and CAM methods. First, the RVM method (a) does not burden the standard setting panelists (content experts) with probabilistic judgments related to the performance of a 'minimally competent' examinee, and (b) provides a higher level of classification accuracy and consistency, compared to the Angoff method. This finding holds for all three RVM applications: RVM (DSM-C), RVM (DSM-L), and RVM (IRT). Second, although the data-driven CAM approach demonstrated high levels of agreement with the RVM-based classifications, the advantage of the RVM is that it derives criterion-based cutoff scores from the judgments of content experts. Third, unlike both the Angoff and CAM methods, the RVM allows for simple and dependable derivation of cutoff scores by response-vector units (e.g., content domains) of the test using classical D-scoring, latent D-scoring, or IRT-based estimates of item parameters. Domain-specific cutoff scores of the test can be very useful when additional diagnostic information about examinees' performance is targeted. Furthermore, the alignment of the RVM method with contemporary frameworks of test validity and evidence-based standard setting may enhance its appeal for applications in large-scale or high-stakes testing programs. From a fairness and transparency perspective, the RVM's structure—especially its focus on observable content units and mastery criteria—offers opportunities for clearer communication with stakeholders and defensible decisions in educational evaluation. As such, the method has potential value in both national certification systems and international assessment contexts seeking more equitable approaches to standard setting.

Given the novelty of the RVM, its real-data application and advantages over other methods of standard setting will be of great benefit to both researchers and practitioners in the computation of dependable cutoff scores for standard setting in the field of education.

Future research may expand on these findings by examining how various stakeholders (e.g., educators, policymakers, examinees) perceive and respond to RVM-based decisions, and whether domain-level cutoff scores offer added diagnostic or policy-relevant utility in real testing systems.

Declarations

Ethical Approval: Ethical approval was obtained from the Ethics Committee of Gazi University (Decision No. 1046762, dated 10 September 2024).

Author Contribution:

Ayşenur Erdemir: Writing – original draft, Writing – review & editing, Investigation, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Dimiter Dimitrov: Writing – original draft, Writing – review & editing, Software, Methodology, Conceptualization.

Esra Oyar: Writing – original draft, Writing – review & editing, Investigation, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Serpil Çelikten-Demirel: Writing – original draft, Writing – review & editing, Investigation, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Funding: The authors received no financial support for the research, authorship, and/or publication of this study.

Consent to Participate: Informed consent was obtained from all individual participants included in the study.

Competing Interests: The authors declare that they have no competing interests.

Previous Presentation: A preliminary version of this study was presented at the *International Symposium on Measurement, Selection and Placement* in Ankara, Türkiye, October 4–6, 2024.

References

- Alshawwa, L. (2023). Standard setting: A review of methods. *Asian Journal of Education and Social Studies*, 42(2), 1-7. <https://doi.org/10.9734/ajess/2023/v42i2909>
- Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.
- Akour, M., & Al-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences*, 5(2), 291-301.
- Atanasov, D. V., & Dimitrov, D. M. (2018). *DELTA: A computer program for test scoring, equating, and item analysis under the D-scoring method*. Riyadh, Saudi Arabia: Education & Training Evaluation Commission.
- Binici, S., & Cuhadar, I. (2022). Validating performance standards via latent class analysis. *Journal of Educational Measurement*, 59(4), 502-516. <https://doi.org/10.1111/jedm.12325>
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59-88. https://doi.org/10.1207/s15324818ame1701_4
- Brennan, R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 4 (2), 219-240. <https://doi.org/10.1177/014662168000400209>
- Brown, R. S. (2007). Using latent class analysis to set academic performance standards. *Educational Assessment*, 12(3-4), 283–301. <https://doi.org/10.1080/10627190701578321>

- Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational measurement*, 39(3), 253-263. <https://doi.org/10.1111/j.1745-3984.2002.tb01177.x>
- Busch, J.C., & Jaeger, R.M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, 27, 145-163. <https://doi.org/10.1111/j.1745-3984.1990.tb00739.x>
- Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education*, 12(2), 151-165. https://doi.org/10.1207/s15324818ame1202_3
- Çetin, S., & Gelbal, S. (2013). A comparison of Bookmark and Angoff standard setting methods. *Educational Sciences: Theory and Practice*, 13(4), 2169-2175. DOI: 10.12738/estp.2013.4.1829
- Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, 19(1), 88–106. <https://doi.org/10.1037/1040-3590.19.1.88>
- Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice*, 15(1), 13-21. <https://doi.org/10.1111/j.1745-3992.1996.tb00802.x>
- Cizek, G. J. (2001). *Setting performance standards*. Lawrence Erlbaum Associates.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on test*. Sage.
- Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R.L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 701-731). Praeger Publishers.
- Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. <https://doi.org/10.1177/00131644600200010>
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- Dimitrov, D. M. (2023). *D-scoing method of measurement: Classical and latent frameworks*. Routledge.
- Dimitrov, D. M. (2022). The response vector for mastery (RVM) method of standard setting. *Educational and Psychological Measurement*, 82(4), 719-746. <https://doi.org/10.1177/00131644211032388>
- Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scales in counseling and related fields*. Wiley.
- Dimitrov, D. M., Ghamdi, H., & Alahmadi, M. (2020, December). Setting cutting scores for mastery on Mawhibah's multiple cognitive ability tests (Technical Report 12-2020), Riyadh, Saudi Arabia: Education & Training Evaluation Commission.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1), 1-26. <https://doi.org/10.1214/aos/1176344552>
- Egan, K. L., Ferrara, S., Schneider, M. C., & Barton, K. E. (2009). Writing performance level descriptors and setting performance standards for assessments of modified achievement standards: The role of innovation and importance of following conventional practice. *Peabody Journal of Education*, 84(4), 552-577. <https://doi.org/10.1080/01619560903241028>
- Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching procedure. *Journal of Applied Testing Technology*, 9(1), 1-22.
- Ferrara, S., Swaffield, S., & Mueller, L. (2009). Conceptualizing and setting performance standards for alternate assessments. In W.D. Schafer & R.W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 93–111). Brookes.
- Gündeğer, C., & Doğan, N. (2018). A statistical approach to improve the standard setting process: Cluster analysis. *Journal of Social Sciences*, 8(2), 387-397.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.). American Council on Education/Praeger Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education* 8(1), 41–55. https://doi.org/10.1207/s15324818ame0801_4
- Han, C. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, 17(1), 1-24. <https://doi.org/10.7275/f0gz-kc87>
- Hess, B., Subhiyah, R. G., & Giordano, C. (2007). Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing examinations. *Evaluation & The Health Professions*, 30(4), 362-375. <https://doi.org/10.1177/0163278707307904>

- Hu, L. T., & Bentler, P. M. (1999). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424-453. <https://doi.org/10.1037/1082-989X.3.4.424>
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). Macmillan.
- Jalili, M., Hejri, S. M., & Norcini, J. J. (2011). Comparison of two methods of standard setting: The performance of the three-level Angoff method. *Medical Education*, 45(12), 1199-1208. <https://doi.org/10.1111/j.1365-2923.2011.04073.x>
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461. <https://doi.org/10.3102/00346543064003425>
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G.J. Cizek (Ed.), *Setting performance standards*. Lawrence Erlbaum Associates.
- Kara, H., & Cetin, S. (2020). Comparison of passing scores determined by the Angoff method in different item samples. *International Journal of Assessment Tools in Education*, 7(1), 80-97. <https://doi.org/10.21449/ijate.699479>
- Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4-12. <https://doi.org/10.1111/j.1745-3992.2006.00047.x>
- Keuning, J., Straat, J. H., & Feskens, R. C. W. (2017). The data-driven direct consensus (3DC) procedure: A new approach to standard setting. In S. Blömeke & J. E. Gustafsson (Eds.), *Standard setting in education* (pp. 287-305). Springer. https://doi.org/10.1007/978-3-319-50856-6_15
- Khalid, M. N. (2011). Cluster analysis: A standard setting technique in measurement and testing. *Journal of Applied Quantitative Methods*, 6(2), 46-58.
- Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.
- Lathrop, Q. N. (2014). R package cacIRT: Estimation of classification accuracy and consistency under item response theory. *Applied Psychological Measurement*, 38(7), 581-582. doi:10.1177/0146621614536465
- Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, 47(1), 1-17.
- Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. McGraw-Hill.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225-254). Routledge.
- 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). Lawrence Erlbaum.
- Muthén, L. K., & Muthén, B. O. (2024). Mplus (Version 8.11) [Computer software]. Muthén & Muthén.
- Özarkan, H. B., & Dogan, C. D. (2020). A comparison of two standard-setting methods for tests consisting of constructed-response items. *Eurasian Journal of Educational Research*, 90, 121-138. <https://doi.org/10.14689/ejer.2020.90.7>
- Peabody, M. R., & Wind, S. A. (2019). Exploring the influence of judge proficiency on standard-setting judgments. *Journal of Educational Measurement*, 56(1), 101-120. <https://doi.org/10.1111/jedm.12202>
- Phillips, G. W. (2012). The benchmark method of standard setting. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 323-346). Routledge.
- Plake, B. S., Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and Yes/No standard setting methods. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181-200). Routledge.
- Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development*, 42(3), 223-232. <https://doi.org/10.1177/0748175609344096>
- Revelle, W. (2025). psych: Procedures for Psychological, Psychometric, and Personality Research (Version 2.5.6) [Computer software]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *The Alberta Journal of Educational Research*, 52(1), 53-64. <https://doi.org/10.55016/ojs/ajer.v52i1.55111>
- Roberts, J. S. (2016). Generalized Graded Unfolding Model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp.369-390). (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences). Chapman & Hall/CRC.
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14), 1-5. <https://doi.org/10.7275/an9m-2035>

- Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, 10(13), 1-4. <https://doi.org/10.7275/56a5-6b14>
- Sayar, A., Dimitrov, D., & Sadaawi, A. (2024, August). *Standard-Setting Procedures for the National Assessment for Schools in Saudi Arabia*. Paper presented at the 2024 Annual Meeting of the American Psychological Association, Seattle, WA.
- Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, 12(3), 301-325. https://doi.org/10.1207/S15324818AME1203_5
- Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, 18(3), 233-256. https://doi.org/10.1207/s15324818ame1803_3
- Strickman, S. N. (2017). *You only live up to the standards you set: An evaluation of different approaches to standard setting* (Doctoral dissertation) James Madison University.
- Thomas, C. L., Cassady, J. C., & Finch, W. H. (2018). Identifying severity standards on the cognitive test anxiety scale: Cut score determination using latent class and cluster analysis. *Journal of Psychoeducational Assessment*, 36(5), 492-508. <https://doi.org/10.1177/0734282916686004>
- Weiss, D. J., & Minden, S. V. (2012). A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG (Technical Report). Assessment Systems Corporation.
- Wyse, A., Bunch, M., Deville, C., & Viger, S. G. (2014). A body of work standard-setting method with construct maps. *Educational and Psychological Measurement*, 74(2), 236-262. <https://doi.org/10.1177/0013164413502037>
- Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, 36(7), 602-624. <https://doi.org/10.1177/0146621612451522>
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125-145. <https://doi.org/10.1177/014662168400800201>
- Yim, M. (2018). Comparison of results between modified-Angoff and bookmark methods for estimating cut score of the Korean medical licensing examination. *Korean Journal of Medical Education*, 30(4), 347-357. <https://doi.org/10.3946/kjme.2018.110>
- Yousefi, A. M. (2021). Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests. *BMC Medical Education*, 21, 1. <https://doi.org/10.1186/s12909-020-02436-3>
- Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-51). Lawrence Erlbaum.
- Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 20(2), 15-25. <https://doi.org/10.1111/j.1745-3992.2001.tb00059.x>