A Modification of the IRT-Based Standard Setting Method

Pilar Rodríguez and Mario Luzardo

Abstract We present a modification of the IRT-based standard setting method proposed by García, Abad, Olea & Aguado (Psicothema 25(2):238–244, 2013), which we have combined with the cloud delphi method (Yang, Zeng, & Zhang in IJUFKBS 20(1):77–97, 2012). García et al. (Psicothema 25(2):238–244, 2013) calculate the average characteristic curve of each level, to determine cutoff scores on the basis of the joint characteristic curve. In the proposed new method, the influence of each item on the average item characteristic curve is weighted according to its proximity to the next level. Performance levels are placed on a continuous scale, with each judge asked to determine an interval for each item. The cloud delphi method is used until a stable final interval is achieved. From these judgments, the weights of each item in the scale are calculated. Then, a family of weighted average characteristic curves is calculated and in the next step, joint weighted averaged ICC are calculated. The cutoff score is determined by finding the ability where the joint weighted averaged ICC reach a certain predefined probability level. This paper compares the performance of this new procedure for a math test with the classic Bookmarking method. We will show that this modification to the method improves cutoff score estimation.

Keywords Performance standard setting · Item response theory · Delphi method

1 Introduction

The setting of performance standards is a central issue in educational measurement. Therefore, the methods for setting them have undergone significant development in recent years. It has been one of the most researched topics over the last forty years

P. Rodríguez (✉) · M. Luzardo
Eastern Regional University Center, University of Republic, Maldonado, Uruguay
e-mail: prodriguez@cure.edu.uy

M. Luzardo
e-mail: mluzardo@psico.edu.uy

M. Luzardo
School of Psychology, University of Republic, Montevideo, Uruguay

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_6

and also one of the most contentious (Berk, 1986; Cizek & Bunch, 2007; Glass, 1978; Hambleton, 1978; Hambleton et al., 2000; Jaeger, 1989; Kane, 1994; Linn, 2003; Margolis & Clauser, 2014; Mousavi, Cui & Rogers, 2018). Different methods of setting cutoff scores provide different standards on the same test (Jaeger, 1989). Therefore, it is important to develop methods to set cutoff scores with precision and stability. This work is a contribution in such regard.

2 Method

We present a procedure which introduces a modification to the method for establishing cutoff scores devised by García et al. (2013), combined with the cloud delphi method. The proposed method can be applied to both a bank already built, and bank items built to match a certain performance level.

Let us assume that *k* levels of performance have been defined (for instance, level 1, level 2 and level 3).

In García et al. (2013)'s method, the bank is built to obtain a set of items that will represent each performance level; but it cannot be agreed that all the items classified or developed for each performance level represent the description of that level in the same way.

To capture the difference in the influence of each item for the determination of the cutoff scores, we resort to the cloud delphi method. To apply this method, it is necessary to obtain a continuous magnitude of the performance level of each item. Operationally, a correspondence of the levels is established with the interval [0, k + 1], with the integer values 1, 2, …, k being the lower ends of the levels expressed qualitatively. For instance, if there are three performance levels, the interval will be (0, 4), with the sub-interval (0, 1) corresponding to "does not reach level 1," interval [1, 2) to level 1, interval [2, 3) to level 2, and (3, 4) to level 3. There is a bijective function between the scale of skill and performance levels.

From a group of judges and by applying the cloud delphi method, a numerical value is obtained on the performance scale: where the item will have a subjective probability of 0.5 of being correctly responded to by a subject with that value on the scale. It is a difficult task for a judge to determine the point of the scale where the above property is fulfilled. However, the proposed method asks each judge to determine an interval on the performance scale, where he considers a subject with that performance level to have a 0.5 probability of correctly responding to the item. The width of the interval will reflect the uncertainty in the judge's response. The cloud delphi method allows us to stabilize their response, and the intervals provided by each judge can be used to determine the item score on the performance scale. This value of each item determines a position on that scale, which will then be used to weight its influence on the establishment of each cutoff score.

After the items have been calibrated, the ICC of each item can be used to calculate the weighted ICC in relation to each cutoff score, which we will note as $WP_k(\theta)$.This curve connects the performance level scale with the ability scale, and represents the probability that a subject will correctly respond to a typical item of cutoff score *k*.

From the $WP_k(\theta)$ we can find the joint probability of correctly responding to a prototype item of cutoff score *k* and the previous cutoff scores. We will note this curve as $JWP_k(\theta)$.The cutoff score will be determined as the value of the ability that causes the joint probability to reach a predetermined value μ (for instance, 0.5); that is, it solves the equation $JWP_k(\theta) = \mu$.

2.1 Cloud Delphi Method

The cloud model relates a qualitative concept with quantitative data based on probability and the fuzzy set theory. The most important model here is the normal cloud model, based on the normal distribution and the Gaussian membership function. In particular, the normal cloud model makes it possible to measure the deviation of a random phenomenon from a normal distribution, when the former does not strictly satisfy the latter (Wang, Xu, & Li, 2014).

This model uses three numerical concepts: expectation (*Ex*); entropy (*En*), which represents the degree of cloudiness of the concept; and hyper entropy (*He*), which represents the variability of the concept in the cloud (Yang, Zeng, & Zhang, 2012).

Formally, let us denote U as the universe of discourse, which is made up of numbers, and let *T* be a qualitative concept. Let us assume that concept *T* is determined in U by its expectation, entropy and hyper entropy; in other words, by the triple *(Ex, En, He)*. Let $x \in U$ be a random realization of concept *T,* such that *x* has normal distribution of mean *Ex* and variance σ_x^2. In addition, we assume that σ_x^2 is a random variable with a normal distribution of mean *En* and variance He^2. Let $\mu_T(x) \in [0, 1]$ be the certainty degree of *x* belonging to *T*. We will say that the distribution of *x* over *U* is a normal cloud if

$$\mu(x) = e^{\frac{(x-Ex)^2}{2(y)^2}}\ with\ y \sim N\left(En, He^2\right) \tag{1}$$

Then, the distribution of *x* in universe U is defined as a cloud and *x* is called cloud drop. This definition establishes that drop $x \in U$ is an extension of concept T. Mapping $\mu_T(x)$ establishes that the certainty degree of *x* belonging to concept *T* is a probability distribution (Yang et al., 2012).

The procedure for applying the cloud delphi method was developed by Yang et al. (2012), and we applied it by following the procedure explained in the previous section to obtain the level of each item.

A set of *n* judges was asked to determine the interval on the performance level scale in which they think a subject has a 0.5 probability of correctly responding to the item. The procedure involves the following steps:

Step 1 : Set the iteration counter *j* equal to one.

Step 2 : In iteration *j*, each judge provides the requested interval. The following intervals are thus obtained $\left[l_i^{(j)}, u_i^{(j)}\right]$, where *i* indicas the *i*-th judge.

Step 3 : The interval provided by each judge is expressed in terms of the normal cloud model, determined by the triple $C_i^{(j)} = \left(Ex_i^{(j)}, En_i^{(j)}, He_i^{(j)}\right) i = 1, \ldots, n.$

Cloud parameters can be calculated as follows for $i = 1, \ldots, n$:

$$\begin{gathered} Ex_i^{(j)} = \frac{l_i^{(j)} + u_i^{(j)}}{2} \\ En_i^{(j)} = \frac{u_i^{(j)} - l_i^{(j)}}{6} \\ He_i^{(j)} = \frac{\max\{u_i^{(j)} - u_i^{(j-1)}, 0\} + \max\{l_i^{(j-1)} - l_i^{(j)}, 0\}}{6} \; and \; He_i^{(1)} = \frac{En_i^{(1)}}{6} \end{gathered} \tag{2}$$

Step 4 : Generate the feedback information for the next iteration by using cloud aggregation algorithms described by Yang et al. (2012).

The synthetic cloud and weighted cloud of each item are shown graphically: to each judge for the purpose of making a new estimate of the interval. These clouds are determined by means of the following equations:

Synthetic Cloud

Let us assume we have *n* clouds $C_i = (Ex_i, En_i, He_i) i = 1, \ldots, n$. The parameters of synthetic cloud $C_s(Ex_s, En_s, He_s)$ are defined by

$$\begin{gathered} Ex_s = \frac{1}{n} \sum_{i=1}^{n} Ex_i \\ En_s = \frac{1}{6}\left[\max_i\{Ex_i + 3En_i\} - \min_i\{Ex_i - 3En_i\}\right] \\ He_s = \frac{1}{n} \sum_{i=1}^{n} He_i \end{gathered} \tag{3}$$

Weighted Cloud

The parameters of weighted cloud $C_{wa}(Ex_{wa}, En_{wa}, He_{wa})$ are defined by

$$\begin{gathered} Ex_{wa} = \sum_{i=1}^{n} w_i Ex_i \\ En_{wa} = \sqrt{\sum_{i=1}^{n} (w_i En_i)^2} \\ He_{wa} = \sqrt{\sum_{i=1}^{n} (w_i He_i)^2} \end{gathered} \tag{4}$$

The relative importance of each judge in the *j-th* step is:

$$r_i^{(j)} = \frac{1}{\left|\frac{\left(Ex_i^{(j)} - Ex_s^{(j)}\right)}{Ex_s^{(j)}}\right| + En_i^{(j)} + He_i^{(j)}} \quad i = 1, 2, \ldots, n \tag{5}$$

Finally, the weights are:

$$w_i^{(j)} = \frac{r_i^{(j)}}{\sum_{i=1}^{n} r_i^{(j)}} \quad i = 1, 2, \ldots, n \tag{6}$$

Step 5 : The relative difference of the entropy with respect to the previous iteration, which we will denote as ΔEn; and the *Unc* ratio of hyper entropy with respect to the entropy, are calculated for the *j-th* iteration.

$$\begin{gathered} \Delta En_i^{(j)} = \frac{\left|En_i^{(j-1)} - En_i^{(j)}\right|}{En_i^{(j-1)}} \quad y \ \Delta En_i^{(1)} = En_i^{(1)} \\ Unc_i^{(j)} = \frac{He_i^{(j)}}{En_i^{(j)}} i = 1, \ldots, n \end{gathered} \tag{7}$$

Step 6 : If $Unc_i^{(j)} = 0$ and for δ > 0 prefixed $\Delta En_i^{(j)} \leq \delta \quad i = 1, 2, \ldots, n$ iterations are completed.

The cloud delphi method is applied until the opinion stabilizes. Once the final intervals have been obtained, the synthetic cloud and weighted cloud are obtained. The weighted cloud of each item is considered the final decision of the judges; and its expectation, which we will denote as b_i, will be the score of the item on the performance scale.

To illustrate the information received by a judge, Fig. 1 shows the graph for item 230 of a mathematics test.

2.2 Setting Cutoff Scores

This second stage involves generalizing García et al. (2013)'s method to obtain the cutoff scores. From the ICCs of each item, the weighted average ICC is obtained at each cutoff score.

$$WP_k(\theta) = \frac{\sum_{j=1}^{N} K\left(\frac{b_j - k}{h}\right) P_j(\theta)}{\sum_{j=1}^{N} K\left(\frac{b_j - k}{h}\right)} \tag{8}$$

where b_j is the item score estimation on the performance scale, h is the bandwidth and K is a kernel. Kernels are used to determine the weight of each item in the weighted average ICC estimation.

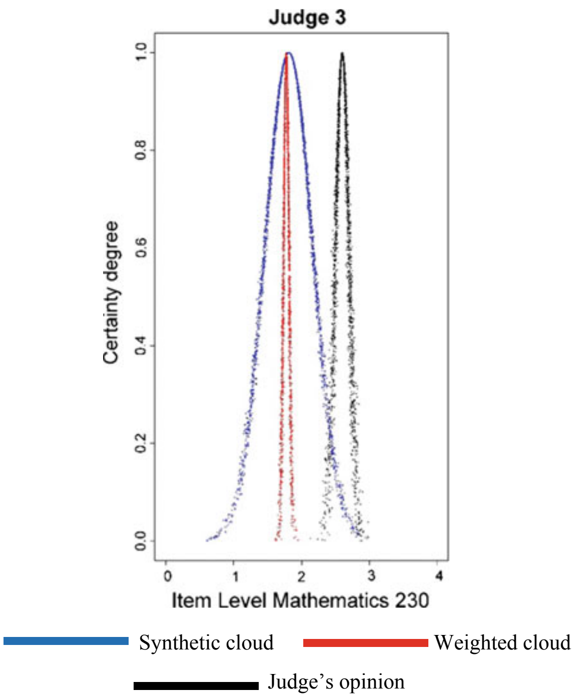

Fig. 1 Cloud model showing synthetic cloud, weighted cloud and judge's opinion for mathematics item 230

Joint averaged ICC (JWP) is calculated for the cutoff scores and represents the probability that an examinee with ability θ will respond correctly to the prototype item of cutoff score k and all previous ones. It is calculated by means of $JWP_k(\theta) = \prod_{z=1}^{k} WP_z(\theta)$.

To calculate cutoff score k, we identify the ability for which the probability of responding to the prototype item of cutoff score k and the previous ones is equal to a predetermined value. We denote probability with μ and the examinee's ability with θ_{cs}. This ability is the solution to the equation $JWP_k(\theta_{cs}) = \mu$, from which the cutoff score is obtained.

3 Results

The method was tested in a university entrance exam assessing reading and mathematics (Rodríguez, 2017). Two methods were applied for performance standard setting: bookmark and the method proposed in this paper.

The item bank has 247 items; a sample of 50 reading and 50 mathematical items was taken. Judges established three performance levels. For the proposed method, two kernels were applied: Gaussian and Epanechnikov. The Gaussian kernel is defined by $\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$; the Epanechnikov kernel by $\frac{3}{4}\left(1-u^2\right)$, with $|u| \leq 1$. Results for different methods are shown in Tables 1, 2, 3 and 4.

Reading

The final cutoff scores from the proposed method represent the average of the results in both kernels.

Mathematics

For the Mathematics test, the final cutoff scores by the proposed method represent the average of the results in both kernels.

A sample was selected of 204 students who took the exam. Their performance levels were classified using the bookmark method and proposed method. They were also classified by expert judgment. The proportions of students in each level are presented in the following graph (Fig. 2).

Table 1 Cutoff scores obtained by Bookmark method in the three performance levels for the Reading test

Levels	Bookmark
1	−0.94
2	0.12
3	1.62

Table 2 Cutoff scores obtained by the proposed method using Gaussian and Epanechnikov kernels in the three performance levels for the Reading test

Levels	Epanechnikov	Gaussian	Average
1	−1.63	−1.47	−1.55
2	−0.19	−0.07	−0.13
3	1.27	1.37	1.32

Table 3 Cutoff scores obtained by Bookmark method in the three performance levels for the Mathematics test

Levels	Bookmark
1	−1.23
2	−0.09
3	1.57

Table 4 Cutoff scores obtained by the proposed method using Gaussian and Epanechnikov kernels in the three performance levels for the Mathematics test

Levels	Epanechnikov	Gaussian	Average
1	−0.88	−0.94	−0.91
2	−0.01	0.13	0.06
3	1.39	1.43	1.41

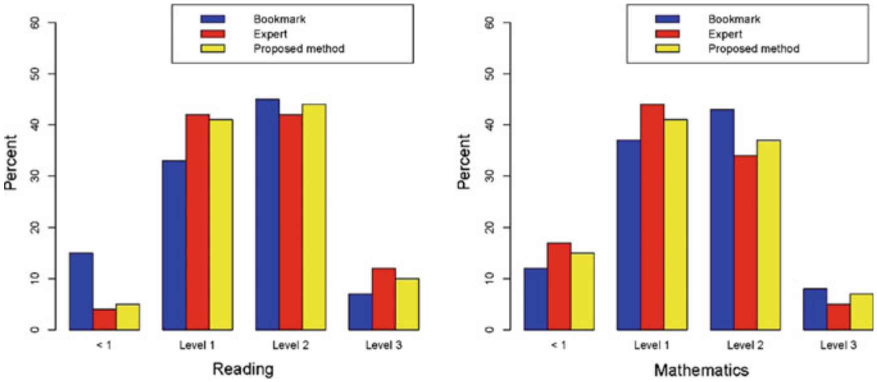

Fig. 2 Comparative graphs of the two methods and expert judgment for the cutoff scores of each level of the Mathematics and Reading tests

4 Discussion

The proposed method establishes cutoff scores closer to the expert judgment than the bookmark method. Moreover, it is better at capturing the variability of the item bank and manages to weight the qualitative judgments. It differs from the bookmark method in that all items participate in determining the cutoff score beyond its ordering by level of difficulty. In addition, it avoids the confusion and discrepancies of the bookmark method when there is no agreement between the difficulty obtained through the theory of response to the item, and a judge's perception of subjective difficulty related to the item. This method considers both the empirical difficulty and the judges' relative difficulty, with both participating in determining the cutoff scores.

This method also allows greater variability in the judges' opinion, capturing the fuzziness of the process; it does not require the determination of a score, but an interval, which makes the task simpler and more efficient.

Unlike García et al. (2013)'s original method, which requires that the items are developed for a certain performance level, it can be applied to banks of previously developed items. The proposed method is also more flexible, as the original considers the items developed for each level to contribute with the same magnitude to each cutoff score. Therefore, this approach makes it possible to obtain a more adjusted valuation of the contributions of each item in the continuum representing the performance level. These advantages make the proposed method a better alternative for the establishment of cutoff scores.

References

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion. *Review of Educational Research, 56*(1), 137–172.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oak, CA: Sage Publications.

García, P. E., Abad, F. J., Olea, J., & Aguado, D. (2013). A new IRT-based standard setting method: Application to elath-listening. *Psicothema, 25*(2), 238–244.

Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement, 15*(4), 237–261.

Hambleton, R. K. (1978). The use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15*(4), 277–290.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355–366.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485–514). New York: American Council on Education and Macmillan.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425–461.

Linn, R. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved from: http://epaa.asu.edu/epaa/v11n31/.

Margolis, M. J., & Clauser, B. E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard setting exercises. *Educational Measurement Issues and Practice, 33*(1), 15–22.

Mousavi, A., Cui, Y., & Rogers, T. (2018). An examination of different methods of setting cutoff values in person fit research. *International Journal of Testing*, 1–22. https://doi.org/10.1080/15305058.2018.1464010.

Rodríguez, P. (2017). Creación, desarrollo y resultados de la aplicación de pruebas de evaluación basadas en estándares para diagnosticar competencias en matemática y lectura al ingreso a la universidad. *Revista Iberoamericana de Evaluación Educativa, 10*(1), 89–107. https://doi.org/10.15366/riee2017.10.1.005.

Wang, G., Xu, C., & Li, D. (2014). Generic normal cloud model. *Information Sciences, 280,* 1–15.

Yang, X. J., Zeng, L., & Zhang, R. (2012). Cloud delphi method. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20*(1), 77–97.