BMC Medical Education

RESEARCH ARTICLE Open Access

Psychometric validation of the Laval developmental benchmarks scale for family medicine

Jean-Sébastien Renaud[1,2,3*], Miriam Lacasse[1,4], Luc Côté[1], Johanne Théorêt[1], Christian Rheault[1] and Caroline Simard[4]

Abstract

Background: With the implementation of competency-based education in family medicine, there is a need for summative end-of-rotation assessments that are criterion-referenced rather than normative. Laval University's family residency program therefore developed the Laval Developmental Benchmarks Scale for Family Medicine (DBS-FM), based on competency milestones. This psychometric validation study investigates its internal structure and its relation to another variable, two sources of validity evidence.

Methods: We used assessment data from a cohort of residents (n = 1432 assessments) and the Rasch Rating Scale Model to investigate its reliability, dimensionality, rating scale functioning, targeting of items to residents' competency levels, biases (differential item functioning), items hierarchy (adequacy of milestones ordering), and score responsiveness. Convergent validity was estimated by its correlation with the clinical rotation decision (pass, in difficulty/fail).

Results: The DBS-FM can be considered as a unidimensional scale with good reliability for non-extreme scores (.83). The correlation between expected and empirical items hierarchies was of .78, p < .0001.Year 2 residents achieved higher scores than year 1 residents. It was associated with the clinical rotation decision.

Conclusion: Advancing its validation, this study found that the DBS-FM has a sound internal structure and demonstrates convergent validity.

Keywords: Criterion-referenced assessment, Validation, Family medicine

Background

Medical schools around the world are moving towards competency-based education [1], which presents assessment challenges as competencies are constructs that are difficult to operationalize [2]. Among these challenges, are the fact that competencies must be operationalized by increasing level in order to define the expected performance objectives at each stage of training [2, 3]. This need has been recognized in a number of countries, such as the United States [2] Canada [4], the United-Kingdom [5], and Australia [6]. Different terms are used to refer to these expected levels of performance such as performance levels, performance indicators, performance criteria, or benchmarks [7]. In North America, "milestones" is the commonly used term used in post-graduate medical education and it is defined as a "defined, observable marker of an individual's ability along a developmental continuum" [8].

* Correspondence: jean-sebastien.renaud@fmed.ulaval.ca
[1]Department of Family and Emergency Medicine, Laval University, 1050, avenue de la Médecine, Université Laval, Québec G1V 0A6, Canada
[2]Office of Education and Continuing Professional Development, Laval University, 1050, avenue de la Médecine, Université Laval, Québec G1V 0A6, Canada
Full list of author information is available at the end of the article

Milestones can be assessed at the end of each rotation [9] and it has been demonstrated that end-of-rotation assessments conducted by clinical teachers are one of the best methods for assessing the attainment of targeted competencies [10–12]. Milestones are best assessed using a criterion-referenced rather than a more traditional norm-referenced approach to assessment [13]. In the normative approach, the resident's performance is assessed by situating it relative to that of others in the group. In contrast, in the criterion-referenced approach, performance (or level of independence) is assessed using a descriptive scale, using multiple authentic assessments situations [7]. Thus, to monitor residents' progression, their assessment should be done using descriptive scales defining milestones, which specify the expectations at various important stages of training for several domains or contexts of practice [9]. These scales should be provided to supervisors (through residency programs) as the basis for their judgment [10].

In Canadian family medicine residency programs, there are no specific milestones defined for different levels of training, or for the end of each rotation (e.g., end of postgraduate year 1). The competency framework, the CanMEDS-FM [11] specifies the key and enabling competencies, which are what the residents are required to demonstrate at the end of their program. However, the milestones defining the expected progress at each rotation of the program have not been defined. There is therefore a need not only to develop family medicine residency milestones based on the CanMEDS-FM, but also tools to assess them at each level of training.

To address this issue, the Laval University family medicine residency program developed the Laval Developmental Benchmarks Scale for Family Medicine (DBS-FM) [12, 14] (see Additional file 1). Based on the CanMEDS-FM competency framework, the DBS-FM is an assessment tool that provides milestones and sets expectations for the development of 34 key and enabling competencies during the 26 training periods of the program. This tool focuses on a specific set of relevant competencies to be assessed at each clinical rotation.

The DBS-FM was incorporated into the family medicine residency program in 2016 as part of a gradual implementation of a competency-based curriculum that begun in the early 2010s. The introduction of this new assessment tool required training and coaching clinical teachers on how to use it. To this end, an online tutorial was offered to clinical teachers as well as on-demand coaching provided by the program director. This investment was quickly offset by the fact that clinical teachers and residents appreciate that this tool clarifies the level of competency residents are expected to attain at each training period. For this reason, results of the DBS-FM are now an essential component of residents' progress reports.

The development and validation of the DBS-FM was informed by modern validity theory [15]. A first study insured its content validity using a Delphi methodology to identify the most salient key and enabling competencies from the CanMEDS-FM and their associated milestones [14]. A second study investigated validity evidence based on the response process upon which improvements were made to the DBS-FM [16].

The aim of this paper is to present the third validation study of the DBS-FM, which focused on the investigation of its psychometric properties. This study is important because the DBS-FM is the first milestone-based assessment tool for the CanMEDS-FM competency framework that has undergone an extensive validation process. It could therefore serve as a model for other milestone-based assessment tools in Canada and in other countries using CanMEDS as a basis for their medical competency framework [17]. In addition, we still have very little evidence about the psychometric quality of the tools developed to assess competency milestones in medical education. Studies presenting those tools, developed for other competency frameworks, provide very limited evidence on their psychometric qualities (e.g., [3, 18–20]).

Methods

Sample and procedures

We selected the first cohort (2016–2018) of family medicine residents assessed with the Laval DBS-FM ($n = 106$) for all the clinical rotations of their two-year program. Clinical teachers used the DBS-FM to assess their competencies at the end of each clinical rotation, totaling 1432 assessments.

Laval developmental benchmarks scale for family medicine

The Laval DBS-FM can assess 34 enabling competencies, including 13 key (mandatory achievement) competencies, with progression milestones specified for each of them. A variable set of relevant competencies is assessed during each clinical rotation. For each of them, clinical teachers assess the level of self-directedness of residents using the following three-point scale: Supervision by direct observation / Supervision by case discussion / Independent, with specific rubrics defined for each level. Assessing the level of self-directedness can initially be challenging for clinical teachers. Indeed, our experience shows that they are used to judging residents' performance but less so residents' self-directedness, even if those two concepts are related. In other words, using the DBS-FM required clinical teachers to change the focus of their assessment. Depending on the competency and time period, those levels of self-directedness are considered as one of the following: early achievement, achievement at expected timing, limit for achievement of competency, or late competency achievement. In order to suggest the rotation decision (pass, in difficulty, or failure) to the evaluator, the computerized system performs a calculation based on the proportion of unachieved competencies (i.e. limit or late). This calculation takes six parameters into account: 1) a late score for one key competency or more results in a *failure*; 2) three or more late scores for non-key competencies result in a *failure*; 3) limit scores for all competencies result in an *in difficulty* decision; 4) a maximum of one late score for a non-key competency without any other late or limit results in a *pass*; 5) limit scores for all competencies, with at most one late non-key competency, lead to an *in difficulty* decision; and 6) limit scores for all key competencies only or all non-key competencies only result in a *pass*. However, the final decision as to the outcome of the rotation remains in the hands of the evaluator, who may or may not accept the system's proposal. A competency achievement score (CAS) is also calculated, ranging from 0 to 100%, and is interpreted as the proportion of competencies for which the developmental level was assessed as "Independent" relative to the total number of competencies assessed during the clinical rotation. This score helps to keep track of residents' progress. It is also considered in the selection process for advanced residency programs in family medicine, as a high CAS in the first year of residency is an indication of a high achievement on enabling competencies.

Analyses

The internal structure of the DBS-FM was assessed using three sets of analyses. First, we analyzed data from the 1432 assessments with the Rasch Rating Scale Model (Andrich, 1978) in Winsteps 3.81. This model was chosen because it allows for missing data in the analysis. Therefore, it was possible to analyze the 34 items (i.e. 34 competencies) in a single model even if only item subsets were used for each clinical rotation. The Rasch analysis process was inspired by the guidelines of Tennant and Conaghan [21] and of Linacre [22]. After investigating model fit, we analyzed rating-scale functioning, dimensionality and local independence, reliability, differential-item functioning, and item targeting. Secondly, we estimated the correlation between expected and empirical item hierarchies. In fact, competencies that should be acquired early in the program according to experts consulted in a previous Delphi study [14] should be the easier items on the DBS-FM, and conversely, competencies that should be acquired late in the program according to experts should be harder items on the DBS-FM. To estimate this correlation, 31 out of the 34 competencies were used because 3 of them were modified between the Delphi study and the final version of the DBS-FM. Thirdly, to test the responsiveness of the CAS on the DBS-FM, we compared the residents' average score for their first and second years with a paired sample t-test. Finally, we estimated the DBS-FM convergent validity with a point-biserial correlation between residents' CAS and a dichotomous variable indicating the decision for the clinical rotation (fail /in difficulty/ pass).

Results

Internal structure

Model fit

The 34 items showed an acceptable fit to the Rasch Rating Scale Model, based on Linacre's [22] guidelines. All items had an infit mean-square statistic between .79 and 1.49 (M = 1.03, SD = .15), and 32 had an outfit mean-square statistic between .75 and 1.43 (M = 1.07, SD = .29), with two items exceeding 1.50. Items 11 and 6 had respectively outfit mean-square values of 1.59 and 1.93. We decided nevertheless to keep both items for two reasons. First, removing them would negatively affect content validity, as these are the 34 items retained from a larger set of competencies to better represent the CanMEDS-FM framework [14]. Second, because items with infit or outfit mean-square statistics between 1.5 and 2.0 are considered "unproductive for construction of measurement, but not degrading" [22]. Infit and outfit mean-square statistics for persons had a mean of 0.97 (SD = .42) and of .98 (SD = 1.16), respectively. Out of the 1432 persons observed, 43 (3%) had a statistically significant infit or outfit value at a .01 level of significance (i.e., standardized value greater than |2.58|). They were removed from subsequent analyses. Upon removal, mean item and person fit statistics improved slightly. Items infit and outfit mean-square values were thereafter respectively 1.01 (SD = 0.12) and 1.00 (SD = 0.38), while person infit and outfit mean-square values were respectively 0.98 (SD = 0.36) and 0.90 (SD = 0.89).

Rating scale functioning

Option characteristic curves are illustrated in Fig. 1. Analysis of the rating scale structure was carried out using Linacre's [23] eight guidelines, summarized in Table 1. Guidelines 1, 3, 4, 5, 7 were respected, while guidelines 2, 6, and 8 were not. Non-respect of the second guideline (Regular observation distribution) reflected the fact that only 0.2% of the observations received the lowest rating (1 = Supervision by direct supervision), while the majority (85.7%) of the observations received the highest rating (3 = Independent). Regarding the sixth guideline (Ratings imply measures, and measures imply ratings), the low congruence between

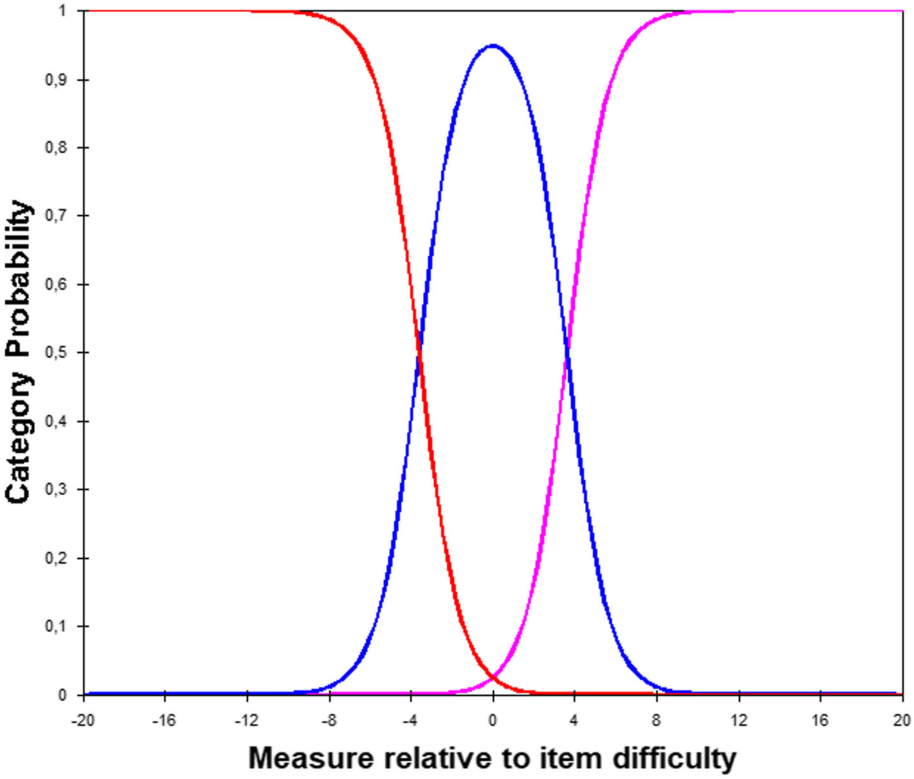


Fig. 1 Option characteristic curves for the 3-point scale

ratings and measures concerned the lowest rating (option 1) and therefore relied on only 54 observations for this estimate. Non-respect of the eighth guideline (Step difficulties advance by less than 5.0 logits) implies large steps on the latent variable between rating options and therefore less measurement precision.

Dimensionality and local Independence

A principal residuals component analysis showed that the first dimensions had an Eigenvalue of 33.3 and explained 49.5% of score variability. The second dimension had an Eigenvalue of 1.9 and explained 2.8% of score variability. The second dimension having a strength of

Table 1 Analysis of the rating scale structure using Linacre's [23] eight guidelines

Linacre's (2004) guidelines	Result
1. At least 10 observations of each category	There were at least 10 observations per response option (54 observations in the first option; 3615 in the second; and 22,023 in the third).
2. Regular observation distribution	Distribution of observations across response options was irregular, meaning that option 3 was clearly the most frequent option, followed by option 2, while option 1 was seldom chosen.
3. Average measures advance monotonically with category	Average ability estimates advanced monotonically with options going from −1.10 logits (option 1) to 2.77 logits (option 2) and then to 6.59 logits (option 3).
4. Outfit mean-squares less than 2.0	Infit and outfit indices were acceptable, all comprised between .99 and 1.30.
5. Step calibrations advance	Step calibrations advanced, indicating no disordered thresholds. The step between option 1 and 2 was estimated at −3.61 logits, and the step between option 2 and 3 was estimated at 3.61 logits.
6. Ratings imply measures, and measures imply ratings	Congruence between measures and ratings as well as between ratings and measures was generally good. It varied between 66 and 93% for options 2 and 3. For option 1, the congruence between measures and ratings was acceptable at 55%, but the congruence between ratings and measures was at 11%.
7. Step difficulties advance by at least 1.4 logit	The distance of 7.22 logits between the two steps was larger than 1.4 logits.
8. Step difficulties advance by less than 5.0 logits	The distance of 7.22 logits between the two steps was larger than 5 logits.

less than two items, the structure of the DBS-FM was considered unidimensional. Regarding local independence, the largest standardized residual correlation between the items had a value of .48 (between items 1 and 2), indicating that the maximum amount of shared variance between two items was 23%. Items were therefore considered locally independent.

Differential item functioning

We tested the invariance of the measurement scale between year 1 and year 2 observations. This was done by investigating for the presence of differential item functioning (DIF) based on residency level (year 1 versus year 2) using Welch's t-test. A Bonferroni correction was applied to guard against the inflation of type 1 error because this analysis resulted in 34 tests, i.e. one for each item. The alpha level of statistical significance was therefore set at .05/34 = .001. Two items (21 and 22) showed significant DIF, both being easier for year 2 residents. The Item 21 (Clinical expertise – Technical gestures) parameter estimate was 3.05 logits for year 1 residents and 1.84 logits for year 2 residents, with an estimated difference of 1.22 logits between the two. The Item 22 (Clinical expertise – Investigation and treatment) parameter estimate was 2.38 logits for year 1 residents and 1.39 logits for year 2 residents, with an estimated difference of .98 logits between the two. To test the impact of these DIF on ability estimates, we correlated resident ability estimated with and without these two items. The correlation between these two score sets was 0.99.

Reliability of CASs

The reliability of residents' CASs was estimated at .83 for observations not having an extreme score (n = 752) (i.e. ability parameter of 7.00 logits or lower), and at .66 (n = 1389) when including an analysis of the 637 residents having an extreme score. As can be seen in Fig. 2 below, the extreme scores, especially those at the top of the scale, have the highest standard error or, in other words, the lowest measurement precision. Classical reliability estimates for the subsets of items used in the different clinical rotations, using Cronbach's alpha, were between .76 and .93.

Item targeting

Residents' ability parameters ranged from – 4.33 to 9.45 logits (M = 6.34 logits, SD = 2.43). More precisely, as illustrated in Fig. 3, ability parameters for year 1 residents ranged from – 4.33 to 9.45 logits (M = 4.89 logits, SD = 2.46) (n = 803 assessments), and from – 0.09 to 9.45 logits (M = 7.75 logits, SD = 1.85) for year 2 residents (n = 629 assessments). In comparison, difficulty parameters for the 34 items of the DBS-FM ranged from – 4.24 to 2.72 logits (M = 0.00 logits, SD = 1.79). The Wright map (Fig. 4) shows the location of the candidates ("person" column) and items ("measure" column) relative to each other on the latent variable. The "BOTTOM P = 50%" column shows the Rasch-Thurstone thresholds for the lowest rating (option 1) on each item, where the probably of being rated as "1″ or higher is 50%. The "TOP P = 50%" column shows the Rasch-Thurstone

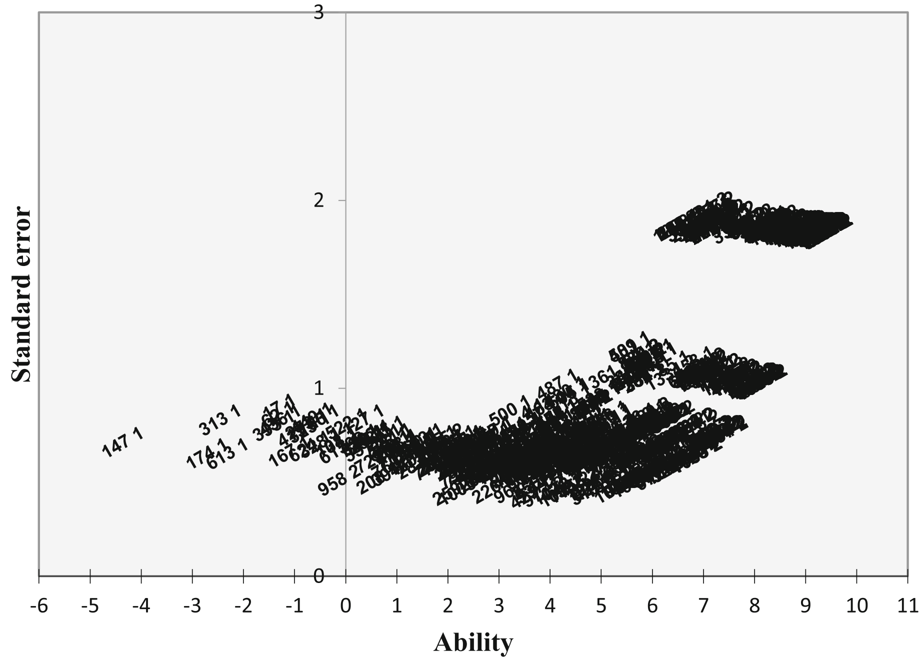


Fig. 2 Standard error of measurement relative to estimated Rasch ability level of residents

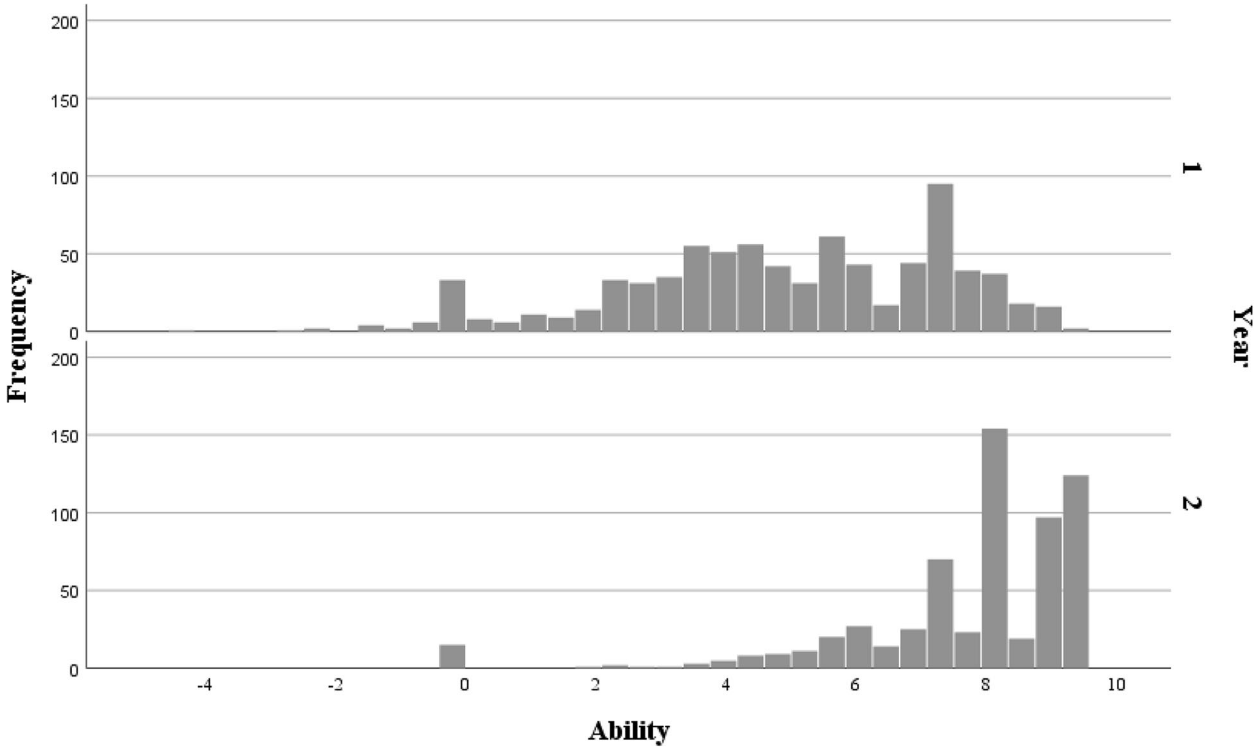


Fig. 3 Distribution of the Rasch ability parameters for year 1 (top) and year 2 (bottom) residents

thresholds for the highest rating (option 3) on each item, where the probably of being rated 3 or below is 50%. The distance between the bottom and upper Rasch-Thurstone thresholds is the operational range of the scale, in other words the latent variable range where the scale is able to discriminate between different competency levels, i.e. between approximately − 8.00 and 7.00 logits. Therefore, the scale cannot discriminate between the highest scoring residents, located between 7.00 and 9.45 logits. For year 1 residents, 232 (32%) out of the 803 assessments were higher than 7.00 logits. For year 2 residents, 489 (68%) of the 629 assessments were higher than 7.00 logits.

Item hierarchy

The expected item hierarchy corresponded to the ordering of competencies by time of expected achievement by the 28 experts at the last phase of the Delphi study [14]. This ordering was highly reliable, both the Generalizability coefficient [24] and the Dependability index [25] being .91. The empirical item hierarchy estimate was also reliable (Rasch item reliability = 0.99). The correlation between the expected item hierarchy according to experts and the empirical item hierarchy estimated by the Rasch item difficulty parameters was.78, $p < .0001$.

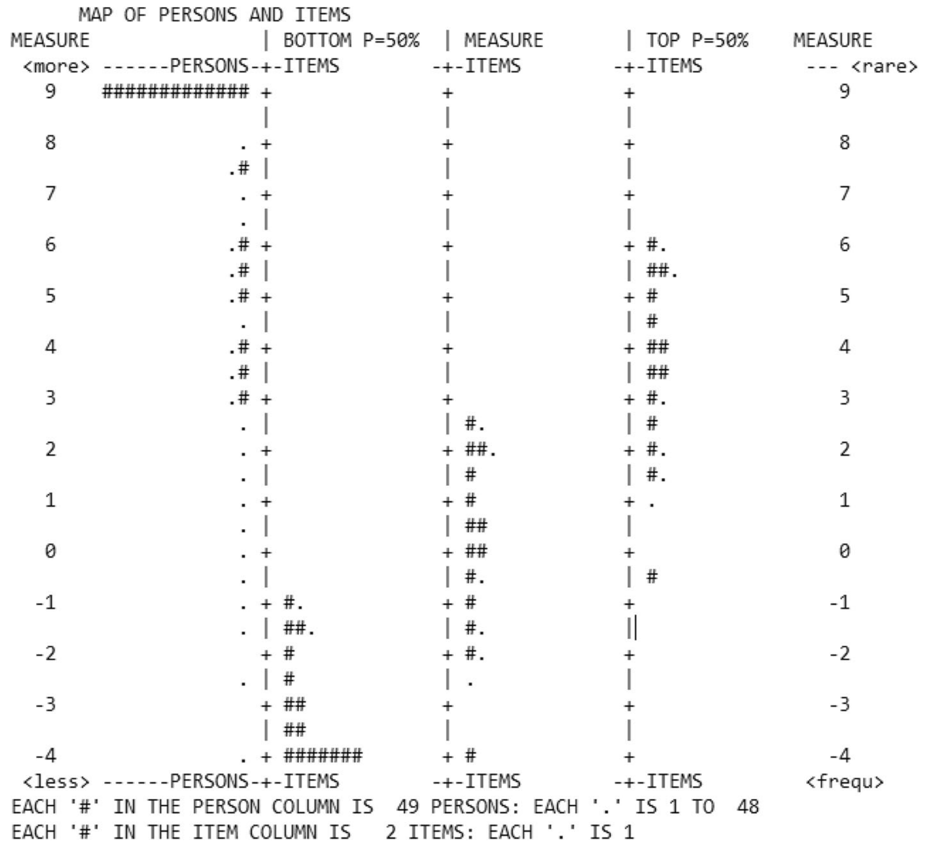
Global score responsiveness

Figure 5 shows the average CAS on the DBS-FM with 95% confidence intervals for the 26 periods of the residency program. The average CAS was .71 (SD = .18) for year 1 residents (clinical rotations 1 to 13) and .83 (SD = .10) for year 2 residents (clinical rotations 14 to 26). A paired sample t-test showed that the difference between the average CAS for year 2 and year 1 residents is statistically significant, $t(94) = -7.52$, $p < .0001$. Using the Rasch ability parameters rather than the CASs yielded similar results, t (1427.6) = − 25.00, $p < .0001$.

However, the difference between those 2 years is lower than expected. The expected CAS (Fig. 6) for the first year of residency varied between .23 and .49 for an average student, which is much lower than the observed CAS, which varied between .59 and .74. The expected CAS for year 2 residents varied between .73 and .91, which is comparable to the observed CAS that varied from .74 to .94.

Convergent validity

Results from the point-biserial correlation, $r = -.28$, $p < .0001$, show that the CAS was significantly associated with being classified as “pass” or “in difficulty or failure.” In other words, having a low CAS was associated with a higher probability of an “in difficulty or failure” decision for a clinical rotation.



```
      MAP OF PERSONS AND ITEMS
MEASURE               | BOTTOM P=50%  | MEASURE       | TOP P=50%    MEASURE
 <more> ------PERSONS-+-ITEMS        -+-ITEMS        -+-ITEMS         --- <rare>
   9    ############# +               +               +                  9
                      |               |               |
   8                . +               +               +                  8
                   .# |               |               |
   7                . +               +               +                  7
                    . |               |               |
   6               .# +               +               + #.               6
                   .# |               |               | ##.
   5               .# +               +               + #                5
                    . |               |               | #
   4               .# +               +               + ##               4
                   .# |               |               | ##
   3               .# +               +               + #.               3
                    . |               | #.            | #
   2                . +               + ##.           + #.               2
                    . |               | #             | #.
   1                . +               + #             + .                1
                    . |               | ##            |
   0                . +               + ##            +                  0
                    . |               | #.            | #
  -1                . + #.            + #             +                 -1
                    . | ##.           | #.            ||
  -2                  + #             + #.            +                 -2
                    . | #             | .             |
  -3                  + ##            +               +                 -3
                      | ##            |               |
  -4                . + #######       + #             +                 -4
 <less> ------PERSONS-+-ITEMS        -+-ITEMS        -+-ITEMS        <frequ>
EACH '#' IN THE PERSON COLUMN IS  49 PERSONS: EACH '.' IS 1 TO  48
EACH '#' IN THE ITEM COLUMN IS   2 ITEMS: EACH '.' IS 1
```

Fig. 4 Wright map of persons and items parameters

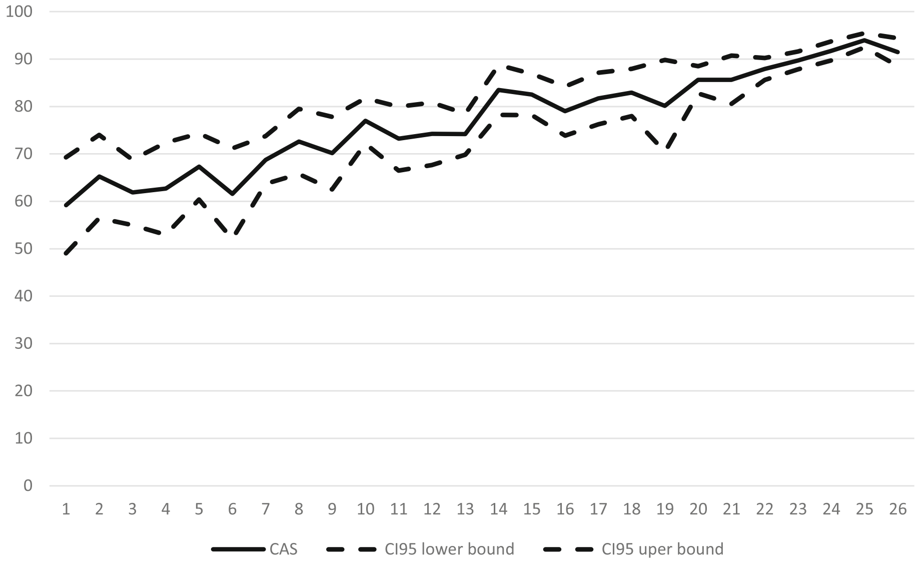


Fig. 5 Average CAS with 95% confidence intervals for the 26 periods of the residency program

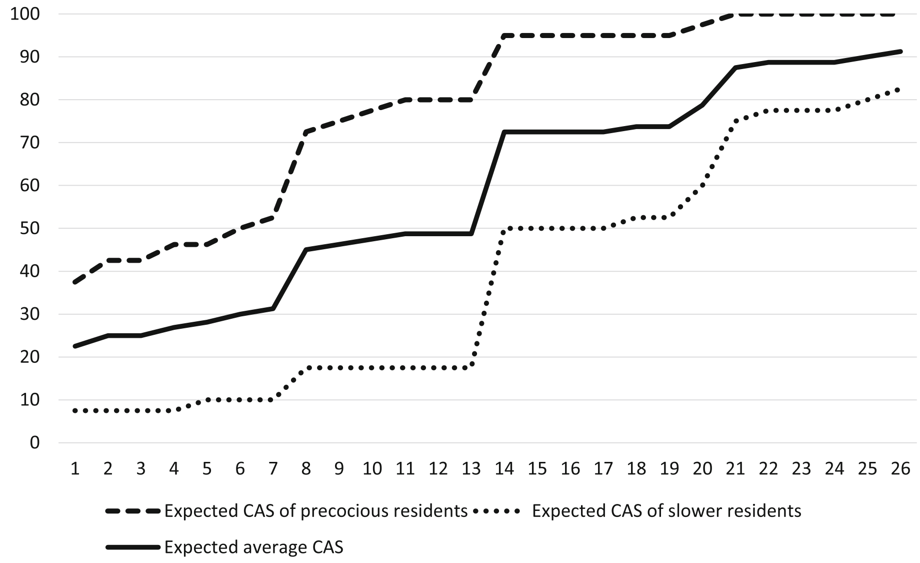


Fig. 6 Expected CAS

Discussion

The DBS-FM is a criterion-referenced, milestone-based, assessment tool based on the CanMEDS-FM used to assess family medicine residents at the end of each clinical rotation. In this validation study, we used modern and classical psychometric analyses to gather empirical evidence on its internal structure and relation to another variable. To the best of our knowledge, it is the first study to extensively explore the psychometric qualities of a milestone-based tool designed for the assessment of residents. Indeed, previous studies focused essentially on content validity, variability of scores and their capacity to show residents' progression (e.g., 3, 18, 19, 20). Results of the study show that milestone-based assessments of residents can be reliable and discriminate between competency levels and stages of residency, and can be summarized into a global latent competency score. In addition, the DBS-FM can be used by other medical schools as a model or an example of a milestone-based assessment tool that has undergone extensive validation. It could therefore contribute to filling an identified gap in the adoption and implementation of competency milestones in residency programs [2].

Analyses of the internal structure showed that the DBS-FM can be considered as unidimensional with no locally dependent items. Consequently, it is appropriate to summarize residents' competency level using a single synthetic score, and this score is sensitive enough to reflect residents' progression between their first and second year, while individual items can be used to provide more directed feedback. The ability to summarize competency levels using a single score on a latent construct is also compatible with the conceptual view that a competence is a general quality or attribute that is not directly observable [26]. In addition, the empirical item hierarchy supports the adequacy of the ordering of milestones by experts consulted in a previous study [14]. The correlation of 0.78 indicates that the expected and empirical item hierarchies share 61% of variance. We consider this to be relatively high, as experts usually struggle to guess the difficulty level of items [27].

Internal structure analyses also showed that the Classical Test Theory reliability of the subset of items used for the different clinical rotations varies between acceptable (α = .73) and very good (α = .93) [28]. The reliability of the 34 items of the DBS-FM, estimated by the Rasch model, is good (.83) for non-extreme scores (i.e. scores lower than 7.00 logits). However, reliability drops (.66) with the inclusion of extreme scores due to their large degree of measurement error. This means that the DBS-FM cannot reliably discriminate between the highest observed competency levels (i.e., 7.00 logits and higher), resulting in a large standard error of measurement for the highest scores. In other words, the item targeting is adequate for the goal of measuring low and intermediate competency levels, but not for measuring the highest levels. This is in line with the aim of the DBS-FM, which is not to discriminate among solid levels of competency, but to help the program ensure that every resident achieves the minimal competency level needed for independent professional practice and to identify those who do not meet this minimal level. It should also be mentioned that criterion-referenced assessments have long been known for having lower item variances than normative-referenced assessments because scores are more concentrated at the higher end of the notation scale [29–31].

If one needed to reliably discriminate between the highest competency levels, some solutions could be envisioned. For example, harder items (i.e. competencies achieved at the end of the two-year program or competencies achieved at the end of the two-year program only by some, but generally achieved later by most) could be added to the DBS-FM. In addition, the highest rating (option 3) on the rating scale could be split into two or three options, with the highest option going beyond "Independent." The large distance in logits between the two steps of the rating scale suggests that there is space on the latent competency variable for a finer-grained rating scale. For instance, a rating scale similar to the O-SCORE could be considered [32]. The O-SCORE has 5 levels that reflect a finer grained progression toward complete independence. Such strategies would also help to identify top performers for promotion or selection purposes.

We also observed differential item functioning for items 21 (Clinical expertise – Technical gestures) and 22 (Clinical expertise – Investigation and treatment) when comparing year 1and year 2 residents. Both items relate to clinical expertise and were harder for year 1 students than for year 2 students when the ability level remained constant. Our hypothesis is that at a similar ability level, year 1 residents are still not as good as year 2 residents when it comes to investigation and treatment as well as to technical gestures. The two differential item functionings did not have a practical impact because the correlation between the residents' ability parameters estimated with and without these items was 0.99. Therefore, differential item functioning does not pose a threat to the validity of the interpretation of residents' scores.

The CAS showed sensitivity to change and made it possible to detect a statistically significant difference between the performance of year 1 and year 2 residents. This result is consistent with that of other studies that also found that milestone-based assessments of residents can reflect residents' growth over time or distinguish between stages of training [20, 33]. However, in the present study, the difference between year 1 and year 2 residents was lower than predicted. The prediction was that year 1 residents would have much lower CAS (between .23 and .49, rather than the observed .59 to .74) and would show a relatively big increase of .24 (from .49 to .73) in their CAS between the 13th and 14th period, representing the transition between year 1 and year 2, similar to what was observed by Goldman et al. [20]. The empirical data show that year 1 residents have better CAS than expected and that the transition from year 1 to year 2 is much more gradual. This gradual increase in competency level throughout residency training was also observed in another study [32]. A possible explanation for these divergent results in the literature is that the progression of residents' competency levels could be highly dependent on the program.

The DBS-FM has convergent validity when correlated with the clinical rotation decision ("pass" vs "fail/in difficulty"). A higher CAS was associated with a higher probability of being classified as "pass," while a lower CAS was associated with a higher probability of being classified as "fail" or "in difficulty." Stated differently, the CAS demonstrates decision consistency with pass/fail decisions. This is a necessary quality to ensure the credibility of the assessment.

There are some limits to this study. First, although it seems plausible that these results should be similar for the next cohorts of family medicine residents, they cannot be automatically generalized. Variations between cohorts, between assessors, or interaction effects between cohorts and assessors, for example, could lead to some variations in its psychometric properties. It will therefore be necessary to monitor the psychometric properties of the DBS-FM for future cohorts. Second, the DBS-FM can be used as a model by other family medicine programs, but it will need to be adapted to the reality of those programs to ensure its validity. Third, differential item functioning was tested for year of residency, but not for gender, due to the anonymous nature of the data. However, the milestones for the acquisition of some competencies could differ between males and females, which would result in differential item functioning. Fourth, when investigating the DBS-FM's relations to other variables, we tested its convergent validity, but not its criterion-related validity. We originally planned to test its predictive validity by comparing the mean CAS on the DBS-FM for residents who passed and those who failed the Certification Examination in Family Medicine of the College of Family Physicians of Canada. But the number of residents who failed this certification exam was too low to run a statistical analysis.

Conclusions

The DBS-FM has a sound internal structure and good convergent validity. It is the first criterion-referenced assessment tool based on the CanMEDS-FM competency framework that is used to assess milestones and that has undergone an extensive validation process. It could therefore serve as a model for other milestone-based assessment tools. Future studies are needed to investigate the validity of criterion-referenced milestone-based assessment tools in the context of formative assessment.

Abbreviations

DBS-FM: Developmental Benchmarks Scale for Family Medicine; CanMEDS-FM: The competency framework for Canadian family physicians;

CAS: Competency achievement score; M: Mean; SD: Standard deviation; Infit: Inlier-pattern-sensitive fit statistic; Outfit: Outlier-sensitive fit statistic; Logit: Log-odds unit; DIF: Differential item functioning; r: Point-biserial correlation coefficient

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-021-02797-3.

Additional file 1.

Acknowledgements

We wish to thank Ms. Selma Chipenda-Dansokho, PhD for her help with language editing.

Authors' contributions

JSR and ML designed the study and drafted the manuscript. JSR, ML, and CS worked on the analysis of data. ML, CR, and CS contributed to the acquisition of data. All authors (JSR, ML, LC, JT, CR, CS) were involved in the interpretation of data, revised the manuscript, and approved its final version.

Authors' information

Jean-Sébastien Renaud, PhD, is an associate professor in assessment in health professions education at the Faculty of Medicine, Université Laval, in Quebec (Canada). He works at the Office of Education and Continuing Professional Development and is affiliated with the Primary Care Research Centre (CERSSPL-UL). ORCID: https://orcid.org/0000-0002-2816-0773
Miriam Lacasse, MD, MSc, CCFP, FCFP (lead reviewer), is a family physician and associate professor at the Department of Family Medicine and Emergency Medicine, Université Laval (Quebec City, Canada). She co-chairs the CMA-MDM Educational Leadership Chair in Health Sciences Education and is the evaluation director for the family medicine residency program. ORCID: https://orcid.org/0000-0002-2981-0942
Luc Côté, MSW, PhD (ed.) is professor at the Department of Family Medicine and Emergency Medicine, former director of the Centre for Health Science Education and former director of Centre for Research in the Health Science Education in the Faculty of Medicine, Université Laval.
Johanne Théorêt, MD, MA, FCMFC is professor at the Department of Family and Emergency Medicine, Université Laval, Québec, and director of faculty development at this Department. She is interested with learners in difficulty and remediation.
Christian Rheault, MD, is a family physician and associate professor at the Department of Family Medicine and Emergency Medicine, Université Laval (Quebec City, Canada). He is the director of the family medicine residency program.
Caroline Simard, PhD, was a research professional at the CMA-MDM Educational Leadership Chair in Health Sciences Education at the moment of the study.

Funding

This study was funded by the Medical Council of Canada's (MCC) Research in Clinical Assessment Grant. The MCC had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets generated during and/or analyzed during the current study are not publicly available due to the fact that they contain students' assessment data that the corresponding author is not authorized to share.

Declarations

Ethics approval and consent to participate

Consent from participants was not needed for this study. The Research Ethics Board at Laval University reviewed this study and concluded that it did not constitute a research project as defined by the Canadian Tri-Council Policy Statement on the ethical conduct of research involving humans. It was considered a study for quality assurance and improvement, which do not require a research ethics board review. The data used for this study is part of the body of data that is normally collected by the department in the ordinary course of its operation. In addition, we anonymized the assessment data used for this study.

Consent for publication

Not applicable.

Competing interests

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Author details

[1]Department of Family and Emergency Medicine, Laval University, 1050, avenue de la Médecine, Université Laval, Québec G1V 0A6, Canada. [2]Office of Education and Continuing Professional Development, Laval University, 1050, avenue de la Médecine, Université Laval, Québec G1V 0A6, Canada. [3]Primary Care Research Centre affiliated with Laval University (CERSSPL-U, 1050, avenue de la Médecine, Université Laval, Québec G1V 0A6, Canada. [4]Educational Leadership Chair in Health Professions Education CMA-MDM, 1050, avenue de la Médecine, Université Laval, Québec G1V 0A6, Canada.

Received: 28 July 2020 Accepted: 17 June 2021
Published online: 27 June 2021

References

1. Weggemans MM, van Dijk B, van Dooijeweert B, Veenendaal AG, Ten Cate O. The postgraduate medical education pathway: an international comparison. GMS J Med Educ. 2017;34(5):Doc63-Doc.
2. Edgar L, McLean S, Hogan SO, Hamstra SJ, Holmboe ES. The milestones guidebook: Accreditation Council for Graduate Medical Education; 2020. https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf. Accessed 1 June 2020.
3. Baglia J, Foster E, Dostal J, Keister D, Biery N, Larson D. Generating developmentally appropriate competency assessment at a family medicine residency. Fam Med. 2011;43(2):90–8.
4. Frank JR, Snell LS, Sherbino J, et al. Draft CanMEDS 2015 Milestones Guide – May 2014. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2014.
5. General Medical Council. Excellence by design: standards for postgraduate curricula. 2017.
6. Royal Australian College of General Practitioners. Curriculum for Australian General Practice 2016 - CS16 Core skills unit. Victoria: The Royal Australian College of General Practitioners; 2016.
7. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting paradigms: from Flexner to competencies. Acad Med. 2002;77(5):361–7. https://doi.org/10.1097/00001888-200205000-00003.
8. Englander R, Frank JR, Carraccio C, Sherbino J, Ross S, Snell L. Toward a shared language for competency-based medical education. Med Teach. 2017;39(6):582–7. https://doi.org/10.1080/0142159X.2017.1315066.
9. Tardif J. L'évaluation des compétences : documenter le parcours de développement [Competency-based assessment: documenting learning trajectories]. Montréal: Chenelière-éducation; 2006. p. xviii, 363. French
10. Saucier D. In: Oandasan I, Saucier D, editors. A guide for translating the triple C competency-based recommendations into a residency curriculum. Mississauga: College of Family Physicians of Canada; 2013.
11. College of Family Physicians of Canada. CanMEDS-Family Medicine 2017: A competency framework for family physicians across the continuum. Mississauga; 2017.
12. Lacasse M, Rheault C, Tremblay I, Renaud J-S, Coché F, St-Pierre A, et al. Développement, validation et implantation d'un outil novateur critérié d'évaluation de la progression des compétences des résidents en médecine familiale [Development, validation, and implementation of an innovative criterion-based tool for assessing the progression of residents' skills in family medicine]. Pédagogie Méd. 2017;18(2):83–100 French.
13. Swing SR, Cowley DS, Bentman A. Assessing resident performance on the psychiatry milestones. Acad Psychiatry. 2014;38(3):294–302. https://doi.org/10.1007/s40596-014-0114-y.
14. Lacasse M, Théorêt J, Tessier S, Arsenault L. Expectations of clinical teachers and faculty regarding development of the CanMEDS-family medicine competencies: Laval developmental benchmarks scale for family medicine

residency training. Teach Learn Med. 2014;26(3):244–51. https://doi.org/10.1080/10401334.2014.914943.

15. AERA, APA, NCME. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.
16. Simard M-L, Lacasse M, Simard C, Renaud J-S, Rheault C, Tremblay I, et al. Validation d'un outil critérié d'évaluation des compétences des résidents en médecine familiale : étude qualitative du processus de réponse [validation of a criteria-based tool for assessing the skills of residents in family medicine: qualitative study of the response process]. Pédagogie Méd. 2017; 18:17–24 French.
17. Council of the European Academy of Teachers in General Practice. The EURACT educational agenda of general practice/family medicine. Greece: WONCA-Region Europe Conference in KOS; 2005.
18. Turner TL, Bhavaraju VL, Luciw-Dubas UA, Hicks PJ, Multerer S, Osta A, et al. Validity evidence from ratings of pediatric interns and subinterns on a subset of pediatric milestones. Acad Med. 2017;92(6):809–19. https://doi.org/10.1097/ACM.0000000000001622.
19. Lomis KD, Russell RG, Davidson MA, Fleming AE, Pettepher CC, Cutrer WB, et al. Competency milestones for medical students: design, implementation, and analysis at one medical school. Med Teach. 2017;39(5):494–504. https://doi.org/10.1080/0142159X.2017.1299924.
20. Goldman RH, Tuomala RE, Bengtson JM, Stagg AR. How effective are new milestones assessments at demonstrating resident growth? 1 year of data. J Surg Educ. 2017;74(1):68–73. https://doi.org/10.1016/j.jsurg.2016.06.009.
21. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care Res (Hoboken). 2007;57(8):1358–62. https://doi.org/10.1002/art.23108.
22. Linacre JM. A user's guide to WINSTEPS MINISTEPS Rasch-model computer programs. Chicago; 2014. Available from: http://www.winsteps.com/aftp/winsteps.pdf. Accessed 1 June 2020.
23. Linacre JM. Optimizing rating scale category effectiveness. In: Smith EV, Smith RM, editors. Introduction to rasch measurement: theory, models and applications. Maple Grove: JAM Press; 2004. p. 258–78.
24. Cronbach LJ. The dependability of behavioral measurements : theory of generalizability for scores and profiles. New York: Wiley; 1972. p. xix, 410.
25. Brennan RL, Kane MT. An index of dependability for mastery tests. J Educ Meas. 1977;14(3):277–89. https://doi.org/10.1111/j.1745-3984.1977.tb00045.x.
26. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med. 2007;82(6): 542–7. https://doi.org/10.1097/ACM.0b013e31805559c7.
27. Hurtz GM, Hertz NR. How many raters should be used for establishing cutoff scores with the angoff method? A generalizability theory study. Educ Psychol Meas. 1999;59(6):885–97. https://doi.org/10.1177/00131649921970233.
28. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Educ. 2018;48(6):1273–96. https://doi.org/10.1007/s11165-016-9602-2.
29. Woodson MICE. The issue of item and test variance for criterion-referenced tests. J Educ Meas. 1974;11(1):63–4. https://doi.org/10.1111/j.1745-3984.1974.tb00973.x.
30. van der Linden WJ. Criterion-referenced measurement: its main applications, problems and findings. Eval Educ. 1982;5(2):97–118. https://doi.org/10.1016/0191-765X(82)90012-X.
31. Popham WJ, Husek TR. Implications of criterion-referenced measurement. J Educ Meas. 1969;6(1):1–9. https://doi.org/10.1111/j.1745-3984.1969.tb00654.x.
32. Keister DM, Larson D, Dostal J, Baglia J. The radar graph: the development of an educational tool to demonstrate resident competency. J Grad Med Educ. 2012;4(2):220–6. https://doi.org/10.4300/JGME-D-11-00163.1.
33. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. Fam Med. 2008;40(8):574–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.