Psychometric validation of the Laval Developmental Benchmarks Scale for Family Medicine

Jean-Sebastien Renaud (✉ jean-sebastien.renaud@fmed.ulaval.ca)

Universite Laval <https://orcid.org/0000-0002-2816-0773>

Miriam Lacasse

Universite Laval

Luc Côté

Universite Laval

Johanne Théorêt

Universite Laval

Christian Rheault

Universite Laval

Caroline Simard

Universite Laval

Research article

Keywords: Criterion-referenced assessment, validation, family medicine

Posted Date: August 12th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-51118/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Medical Education on June 27th, 2021. See the published version at <https://doi.org/10.1186/s12909-021-02797-3>.

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 the assessment data from a cohort of residents ($n = 1\,432$ 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.

Background

In Canada, summative end-of-rotation assessments of residents' clinical performance are based on a number of behaviors reflecting CanMEDS roles observed by clinical teachers (Kassam, Donnon, & Rigby, 2014). To date, they have been recognized as one of the best tools for evaluating all seven CanMEDS roles (1, 2). Still, the absence of well-defined performance expectations for each end-of-rotation assessment throughout the residency program remains a challenge. According to the Triple C Part 2 Report (3), these criteria or milestones should be provided to supervisors (through residency programs) as the basis for their judgment of a resident's progress, thereby assessing the behavior itself rather than comparing with other learners (4). This criterion-referenced approach is also much more consistent with

the paradigm of competency-based assessment than the traditional normative-referenced assessment. In the competency-based paradigm, rather than interpreting the resident's performance normatively by situating it in a group, the assessment is done using a criterion-based approach to assess level of performance (or level of independence) according to a descriptive scale, using multiple measures from authentic situations (5). Thus, to monitor residents' progression, their assessment should be done using descriptive scales defining different levels of performance, also known as development indicators, milestones, or rubrics, which specify the expectations at various important stages of training for several domains or contexts of practice (6).

In Canadian family medicine residency programs, there are no prescribed criterion-referenced standards defined for specific training periods or for the end of each rotation (e.g., end of postgraduate year 1). The competency framework (7) specify the key and enabling competencies, which are what the residents have to be able to demonstrate at the end of their program. However, the milestones reflecting the evolution of the competency level of residents for each rotation throughout the program are not defined and yet are essential to resident assessment. Indeed, performance expectations are not the same at the beginning and at the end of the residency program because residents develop competencies as future family doctors over time. Therefore, a performance that is acceptable at the beginning of the residency program can be considered insufficient at the end of the program.

Consequently, the Laval University family residency program developed a criterion-referenced assessment tool, the Laval Developmental Benchmarks Scale for Family Medicine (DBS-FM) (8, 9) (Appendix). Using milestones, the DBS-FM sets expectations for the development of 34 key and enabling competencies during the 26 training periods of the program. This tool focuses only on a specific set of relevant competencies to be assessed at each clinical rotation.

Development and validation of the DBS-FM was based on modern validity theory (10). 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 (9). A second study investigated validity evidence related to the response process upon which improvements were made to the DBS-FM (11). This psychometric validation study aims at investigating two other sources of validity evidence: internal structure and relations to other variables. This is an important step as it provides empirical evidence about the DBS-FM's reliability, dimensionality, rating-scale functioning, items targeting residents' competency levels, biases (differential item functioning), items hierarchy (adequacy of milestones ordering), score responsiveness, and convergent validity.

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. 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 0–100%, 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 (12) and of Linacre (13). 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 (9) 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 (13) 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 (9). 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" (13). 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 1 432 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 (14) 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 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.

Table 1
Analysis of the rating scale structure using Linacre's (14) eight guidelines

Linacre's (2004) guidelines	Result
1. 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 22023 in the third).
2. 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. 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. 4. Outfit mean-squares less than 2.0	Infit and outfit indices were acceptable, all comprised between .99 and 1.30.
5. 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. 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. 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. 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.

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 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 probability of being rated as "1" or higher is 50%. The "TOP P = 50%" column shows the Rasch-Thurstone thresholds for the highest rating (option 3) on each item, where the probability 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 (9). This ordering was highly reliable, both the Generalizability coefficient (15) and the Dependability index (16) 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 two 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.

Discussion

The DBS-FM is a criterion-referenced assessment tool based on the CanMEDS-FM and 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 with respect to its internal structure and relation to another variable. Results show that the DBS-FM has a sound internal structure but that it is unable to discriminate at the highest end of the latent variable, in other words between residents with the highest CAS. In addition, a low CAS on the DBS-FM is associated with the highest probability of being rated as a resident “in difficulty or failure,” showing evidence of convergent validity.

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. In addition, the empirical item hierarchy supports the adequacy of the ordering of milestones by experts consulted in a previous study. 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 (17).

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 ($\alpha = .73$) and very good ($\alpha = .93$) (18). 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 (19–21).

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 between the two steps of the rating scale suggests that there is space on the latent competency variable for a finer-grained rating scale. Such a strategy 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 1 and 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. It would make sense 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.

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. However, 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. 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. The expected and empirical CAS are comparable for year 2 students.

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.”

There are some limits to this study. First, although it seems plausible that these results should be similar for the next cohorts of family 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

Using milestones, the DBS-FM sets expectations for the development of 34 key and enabling competencies for the two-year residency in family medicine program. Advancing its validation, this study found that it has a sound internal structure and solid convergent validity. Further validation research on the DBS-FM should focus on the consequences of using this criterion-referenced assessment. For example, we want to document the decisional comfort of clinical teachers with this computerized assessment tool. We also are working on evaluating the influence on the quality of feedback of using this criterion-referenced tool over the normative assessment tool previously applied.

List Of 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.

Declarations

Ethics approval and consent to participate

The Research Ethics Board at Laval University reviewed and approved this study.

Consent for publication

Not applicable

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.

Competing interests

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

Funding

This study was funded by the Medical Council of Canada's Research in Clinical Assessment Grant.

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 were involved in the interpretation of data, revised the manuscript, and approved its final version.

Acknowledgements

Not applicable

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.

References

1. Bandiera G, Sherbino J, Frank J. The CanMEDS assessment tools handbook. An introductory guide to assessment methods for the CanMEDS competencies Ottawa: The Royal College of Physicians and Surgeons of Canada; 2006.
2. Chou S, Cole G, McLaughlin K, Lockyer J. CanMEDS evaluation in Canadian postgraduate training programmes: tools used and programme director satisfaction. *Med Educ*. 2008;42(9):879–86.
3. Oandasan I, Saucier D. Triple C. Competency-based curriculum report – Part 2: Advancing implementation. Mississauga: College of Family Physicians of Canada; 2013.
4. Saucier D. A Guide for translating the Triple C Competency-based recommendations into a residency curriculum. In: Oandasan I, Saucier D, editors. Mississauga: College of Family Physicians of Canada; 2013.
5. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting paradigms: From Flexner to competencies. *Acad Med*. 2002;77(5):361–7.
6. Tardif J. L'évaluation des compétences: documenter le parcours de développement. [Competency-based assessment: documenting learning trajectories]. Montréal: Chenelière-éducation; 2006. xviii, 363 p p. French.

7. College of Family Physicians of Canada. CanMEDS-Family Medicine 2017: A competency framework for family physicians across the continuum. Mississauga, ON; 2017.
8. 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 Médicale*. 2017;18(2):83–100. French.
9. 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.
10. AERA. APA. NCME. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.
11. 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 Médicale*. 2017;18:17–24. French.
12. 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.
13. Linacre JM. A user's guide to WINSTEPS MINISTEPS Rasch-model computer programs. Chicago, IL 2014. Available from: <http://www.winsteps.com/aftp/winsteps.pdf>.
14. 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. pp. 258–78.
15. Cronbach LJ. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley; 1972. xix, 410 p p.
16. Brennan RL, Kane MT. An index of dependability for mastery tests. *J Educ Meas*. 1977;14(3):277–89.
17. Hartz 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.
18. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*. 2018;48(6):1273–96.
19. Woodson MICE. The issue of item and test variance for criterion-referenced tests. *J Educ Meas*. 1974;11(1):63–4.
20. van der Linden WJ. Criterion-referenced measurement: Its main applications, problems and findings. *Evaluation in Education*. 1982;5(2):97–118.
21. Popham WJ, Husek TR. Implications of criterion-referenced measurement. *Journal of Educational Measurement*. 1969;6.

Figures

Figure 1

Option characteristic curves for the 3-point scale

Figure 3

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

Figure 4

Wright map of persons and items parameters

Figure 5

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

Figure 6

Expected CAS

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [developmentalbenchmarksscalefamilymedicine.pdf](#)