Construct-Aligned Rating Scales Improve the Reliability of Program Evaluation Data

Shayna Rusticus, Kevin Eva, and Linda Peterson
University of British Columbia

Abstract: *In workplace-based assessment, research has suggested that aligning rating scales with how clinical supervisors naturally conceptualize trainee performance improves reliability and makes assessment more efficient. This study examined the generalizability of those findings for program evaluation by determining if construct alignment improves the reliability with which competencies are ranked as having been achieved in a medical education program. These results extend previous research into the benefits of construct-aligned scales by suggesting that aggregating students' judgments of their abilities can be used to evaluate the relative successes of a program more efficiently when the scales used are aligned with the constructs of independence and sophistication rather than being phrased in terms of students' performance expectations.*

Keywords: *construct-aligned ratings scales, construct alignment, medical education, program evaluation, Readiness for Clerkship survey, reliability*

Résumé : *Les écrits scientifiques indiquent que l'appréciation de la performance est plus juste et plus efficiente si les échelles de notation utilisées correspondent à la représentation que les formateurs ont de la performance des personnes en formation. Notre étude vise à tester cette observation dans le contexte d'un programme d'éducation médicale en analysant de quelle façon la correspondance des construits avec les représentations sur la performance améliorent la fiabilité des résultats concernant les compétences acquises. Nos résultats suggérent qu'intégrer les appréciations des étudiants sur leurs propres compétences permet d'évaluer le succès d'un programme de façon plus efficiente lorsque les échelles utilisées sont alignées avec les concepts d'indépendance et d'aptitudes plutôt que d'être formulées d'une manière qui reflète les attentes liées à leur performance.*

Mots clés : *échelles de notation alignées sur les concepts, alignement sur les concepts, formation médicale, évaluation de programmes, questionnaire de préparation au stage, fiabilité*

Corresponding author: Shayna Rusticus, Evaluation Studies Unit, Faculty of Medicine, University of British Columbia, Purdy Pavilion, 2221 Wesbrook Mall, Vancouver, BC, V6T 1Z9; shayna.rusticus@ubc.ca

 Canadian Journal of Program Evaluation / La Revue canadienne d'évaluation de programme
33.1 (Spring / printemps), 1–20 doi: 10.3138/cjpe.31094

When evaluating whether a curriculum is producing the intended outcomes, the most feasible strategy is simply to ask trainees how much they learned from the activity in which they participated. However, survey fatigue and the costs incurred from cajoling students into completing evaluations can make it difficult to collect trustworthy information from a full cohort. Further, it is generally well accepted that self-assessment is a poor proxy measure of an individual's level of knowledge or ability (Davis, Mazmanian, Fordis, Van Harrison, Thorpe, & Perrier, 2006; Eva & Regehr, 2005; Eva, Regehr, & Gruppen, 2012). While it would be concerning if one were to use self-assessments to gauge an *individual's* ability, recent studies have demonstrated that aggregating the flawed self-assessments of many individuals can provide reliable information about a *group's* strengths and weaknesses (D'Eon, Sadownik, Harrison, & Nation, 2008; Peterson, Eva, Rusticus, & Lovato, 2012; Peterson, Rusticus, Wilson, Eva, & Lovato, 2015). These are promising findings for the sake of enabling efficient and meaningful evaluation because it is group- or cohort-level success (or lack thereof) that is usually of most interest to program evaluators as they strive to identify aspects of the formal curriculum that can be improved, rather than highlighting strengths or weaknesses that might be idiosyncratic to individuals.

Early forays into this domain focused on the evaluation of a time-limited educational workshop (D'Eon & Eva, 2009). We more recently offered proof that the same principles can be applied to the evaluation of a medical training program's pre-clinical (Peterson et al., 2012) and clinical phases (Peterson et al., 2015). By asking students to self-assess their own abilities, the aggregated group scores consistently rank-ordered the competencies presented in terms of the relative skill level possessed by the cohort, and those rank orderings aligned very well with faculty impressions ($r > 0.85$). While these results are promising, there remains much to learn about how to collect program evaluation in a valid and feasible way that offers guidance regarding what specific aspects of the program are most in need of improvement.

There are many avenues to explore, and no one strategy is likely to be sufficient, but it is noteworthy that evidence is beginning to emerge in the workplace-based assessment literature suggesting that gains can be made by more carefully aligning the response scales used with the way in which respondents naturally think about the construct being considered (e.g., trainee performance: Crossley, Johnson, Booth, & Wade, 2011; Crossley & Jolly, 2012). An effective response scale helps raters quantify their response along the scale at the place that corresponds to their perceptions. It enables multiple raters to accurately and consistently assign similar impressions to the same location on the response scale, thus reducing measurement error (Davies, 2008). In this vein, Crossley et al. cite evidence suggesting that scales designed to reflect linear gradations of performance (e.g., "unsatisfactory" to "superior") and those designed to reflect progress in relation to stage of training (e.g., "well below expectations" to "well above expectations") create considerable uncertainty and hesitation among assessors who, therefore, do not feel comfortable using such terms.

By way of contrast, research into the perspectives of clinical supervisors has suggested that these individuals most commonly describe their trainees in terms

that are quite distinct from the abstract competence labels that are now commonly used in the field. In describing their trainees, supervisors commonly refer to constructs such as trust (Hauer, ten Cate, Boscardin, Irby, Lobst, & O'Sullivan, 2014) and independence (Kennedy, Regehr, Baker, & Lingard, 2009) rather than fitting their impressions to generic expectations held for trainees at particular levels (Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010). An analysis of the milestones used by the Accreditation Council for Graduate Medical Education, conducted by Crossley et al. (2011), suggested that the two key constructs that determine judgments of trust are clinical sophistication (i.e., the ability to complete tasks correctly) and independence (i.e., the ability to complete tasks with little guidance). Following this logic, those authors modified rating scales by aligning the adjectives used with these two constructs and demonstrated that the reliability of workplace-based assessment scores collected in a series of medical training contexts was improved relative to scales that used more conventional descriptors. Taken together, this line of study suggests that aligning rating scales with the cognitive schemas and experience of the raters (what Crossley and Jolly call assessors' "reality map") might offer greater utility than trying to reframe the perspective of the raters to fit the scales (Crossley & Jolly, 2012). Considerably less research has been conducted on the "reality map" of trainees themselves (Eva, Armson, Holmboe, Lockyer, Loney, Mann, & Sargeant, 2012). As a result, while they are intuitively appealing and theoretically well grounded, the generalizability of these previously published empirical demonstrations, particularly for the realm of program evaluation, is unclear.

The purpose of this study, therefore, was to determine whether rating scales aligned to the construct of "independence" or "independence plus sophistication" would increase raters' capacity to consistently rank-order competencies and improve student and faculty agreement on competency ratings assigned to a cohort of students, thereby allowing program evaluation data to be collected in a more efficient manner by requiring responses from fewer individuals. In a prospective study, we compared the performance of the previously used, and more generically worded, competence-oriented scale (focused on whether trainees met expectations) to that of two more concretely operationalized and construct-aligned rating scales that were developed for this study: (1) an independence scale that focuses on the frequency with which students perceive the need for guidance; and (2) a behavioural sophistication/independence scale that focuses on the ability of students to correctly complete a task along with the frequency with which they perceive the need for guidance to do so.

METHOD

Participants

All 262 third year students enrolled in the University of British Columbia's (UBC's) MD undergraduate program in the 2012–13 academic year were eligible to participate. At that time, the MD program was delivered at three campuses around

the province, and we obtained a list of faculty members from all three campuses who served as preceptors/supervisors for third-year students. These included members of each department that provided required clerkship rotations: Family Medicine, Internal Medicine, Obstetrics & Gynecology, Pediatrics, Psychiatry, and Surgery. The number of eligible faculty identified was 612.

Materials

The Readiness for Clerkship (RfC) survey is a competency-based instrument used to evaluate the pre-clerkship years of an undergraduate MD training program (Peterson et al., 2012). It consists of 40 key physician tasks listed in the sequence of a typical patient encounter. The student version is framed as a self-assessment, which is then meant to be aggregated to create cohort scores, whereas the faculty version is framed with faculty members as assessors of the cohort of third-year students. While responses from students and faculty are independent of one another, the focus in each case is on the function of the pre-clerkship years to prepare students for full-time patient-based learning in clerkship by rank-ordering competencies based on the cohort of students' degree of achievement. The items in the survey cluster into two subscales: (1) Clinical Skills and Knowledge Application (CSKA): tasks associated with obtaining and interpreting information from or about patients while integrating this information with prior knowledge to form diagnostic and management plans; and (2) Working as a Professional (WP): tasks related to being responsible for and interacting with the patient and the healthcare team as well as tasks related to demonstrating self-care and self-directed learning. Relative to the originally published RfC survey (Peterson et al., 2012), nine items were revised in the present version of the survey because of poor factor loadings, two items were thought easily combinable into one, and five had wording changed to enhance alignment with a new Readiness for Residency survey (RfR; Peterson et al., 2015). Three new items were also added to the RfC survey to align with the RfR survey. The RfR survey was not used in the present study.

Experimental intervention

Three different versions of scale anchors were tested in this study, which are illustrated in Table 1. The first version was the original competence-oriented scale (CS) used in our prior research, anchored with statements about the degree of achievement that students reached. The second, the independence scale (IS), was the first of two rating scales that were developed by the authors to more deliberately align with the construct of independence as outlined in Crossley et al. (2011). The second construct-aligned scale, the behavioural/independence scale (BIS), was based on both the constructs of clinical sophistication and independence as defined by Crossley et al. and described in the introduction of this paper.

Procedure

Students and faculty were randomly assigned to complete one of the three survey versions. Surveys were administered anonymously in November 2012 (four

Table 1. The Terms Used to Anchor the Five Points on the Original Competence-Oriented Rating Scale and Both Modified Construct-Aligned Versions

Numerical value	Original Competence-oriented Scale	Independence Scale	Behavioural/Independence Scale
1	An unacceptable level of competence	Almost always requires guidance/assistance	Difficulty in completing the task and has numerous errors or omissions. Almost always requires guidance/assistance
2	A marginal level of competence	Frequently requires guidance/assistance	Completes the task with many errors or omissions. Frequently requires guidance/assistance
3	A satisfactory level of competence	Sometimes requires guidance/assistance	Completes the task with some errors or omissions. Sometimes requires guidance/assistance
4	A high level of competence	Rarely requires guidance/assistance	Completes the task with minimal errors or omissions. Rarely requires guidance/assistance
5	An extremely high level of competence	Almost never requires guidance/assistance	Completes the task with no errors or omissions. Almost never requires guidance/assistance
N/A	Unable to rate/not applicable	Unable to rate/not applicable	Unable to rate/not applicable

months after clerkship began, to allow students to be sufficiently immersed into the clerkship that they could have gained some perspective on what aspects of clinical training they were prepared for before entering clerkship, while avoiding such a long interval that they are likely to have forgotten how prepared they felt at the beginning of the clerkship). Students and the majority of faculty supervisors submitted their survey responses via one45® (one45 Software, Inc., Vancouver, Canada). Supervisors from Family Medicine tended not to use one45, so their surveys were disseminated and returned by fax. Both students and faculty had approximately three weeks to complete and submit the survey. Participation in this study was voluntary. The UBC Behavioural Research Ethics Board reviewed and approved the study.

RESULTS

A total of 135 students and 185 faculty (52% and 30% of those eligible, respectively) completed the survey. The faculty sample represented a variety of programs: 14% Family Practice, 25% Internal Medicine, 11% Obstetrics and Gynecology, 10% Pediatrics, 26% Psychiatry, and 14% Surgery. Table 2 presents the number who responded to, and descriptive statistics for, each survey version.

Table 2. Descriptive Statistics for Student and Faculty Ratings on the Readiness for Clerkship Survey

Version	*n*	*M*	*SD*	Min	Max	Range[a]
CS-Student	53	3.22	0.36	2.51	3.86	1.35
CS-Faculty	59	3.12	0.30	2.45	3.73	1.28
IS-Student	44	3.61	0.64	2.48	4.55	2.07
IS-Faculty	65	3.36	0.46	2.55	4.17	1.64
BIS-Student	38	3.46	0.50	2.53	4.35	1.83
BIS-Faculty	61	3.24	0.37	2.54	3.96	1.42

Note. CS = Competence-oriented Scale; IS = Independence Scale; BIS = Behavioural/Independence Scale

[a]Range is the difference between the means of the highest rated item (Max) and the lowest rated item (Min) in the aggregated data set.

Inter-rater reliability

The most important test of our research question regarding whether different rating scales allow better or more efficient analysis of a medical program's strengths and weaknesses in terms of preparing its students for clerkship is an assessment of the extent to which raters offered consistent judgments regarding which competencies were achieved relatively well or relatively poorly by the cohort of students studied. To that end, generalizability theory was applied to the data collected from both students and (separately) faculty who rated all of the survey items. This analysis involves calculating the amount of variance attributable to rater, item, and the rater × item interaction for each survey version (Table 3) and then using those variance components to generate reliability coefficients, with item as the facet of differentiation (Table 4). The resulting coefficients indicate the extent to which raters consistently rank-ordered the aspects of competence listed in the survey and thereby provide a marker of utility indicating the effectiveness of these scales when used for group-level evaluation. This analysis confirmed that a single evaluation using the RfC survey cannot be considered trustworthy because the inter-rater reliability was low in all cases, ranging from G = 0.11 (Faculty BIS) to 0.34 (Student IS). In other words, one respondent's assessment does not provide a reliable indication of which physician tasks are best achieved within the educational program completed by the students.

That said, decision studies, which are statistical techniques to determine how the reliability of data would be altered if the number of observations collected changed, show that when data from multiple raters are aggregated into an average, reliable differentiation between items can be achieved with reasonable efficiency. Data in Table 4 illustrate that G = 0.8 can be achieved most efficiently (i.e., with 8–12 raters) in the case of students evaluating the program using the IS or BIS scales. To statistically compare the reliabilities observed, we therefore used k = 10 raters as a constant to calculate comparable reliabilities and their associated 95% CIs as per Streiner and Norman (2008). Examination of the CIs for the

Table 3. Variance in Scores Attributable to Items, Raters, and the Item x Rater Interaction, as a Function of the Rating Scale and Rater Group

Source	Competence-oriented Scale				Independence Scale				Behavioural/Independence Scale			
	Student (*n* = 43)		Faculty (*n* = 15)		Student (*n* = 27)		Faculty (*n* = 9)		Student (*n* = 23)		Faculty (*n* = 11)	
	Variance	%	Variance	%	Variance	%	Variance	%	Variance	%	Variance	%
Item (I)	0.10	15.39	0.08	14.29	0.33	33.67	0.10	12.99	0.20	25.64	0.12	11.01
Rater (R)	0.19	29.23	0.17	30.36	0.18	18.37	0.20	25.97	0.19	24.36	0.64	58.72
I x R Interaction	0.36	55.38	0.31	55.35	0.47	47.96	0.47	61.04	0.39	50.00	0.33	30.27
Total	0.65	100	0.56	100	0.98	100	0.77	100	0.78	100	1.09	100

doi: 10.3138/cjpe.31094

Table 4. Interrater Reliability for Each Survey Version for Both Students and Faculty

Group	Scale	G	k	G(10)[a]	95% CI	
					Lower Limit	Upper Limit
Students	CS (*n* = 43)	0.16	22	0.65	0.59	0.71
	IS (*n* = 27)	0.34	8	0.84	0.80	0.87
	BIS (*n* = 23)	0.26	12	0.77	0.72	0.81
Faculty	CS (*n* = 15)	0.14	25	0.62	0.56	0.68
	IS (*n* = 9)	0.14	26	0.61	0.55	0.67
	BIS (*n* = 11)	0.11	32	0.56	0.49	0.62

Note. CS = Competence-oriented Scale; IS = Independence Scale; BIS = Behavioural/Independence Scale; G = reliability based on a single rater; k = number of raters

[a]k was set to 10 to allow for a comparison in reliabilities across the different scales by virtue of a constant sample size. This value reflects the midway point between the k values for the two student construct-aligned scales.

student group showed no overlap between the construct-aligned scales and the CS, indicating that the construct-aligned scales are more reliable (or that it would require substantially more students completing the CS scale to achieve the same degree of reliability relative to the IS or BIS scales). This pattern did not hold for faculty ratings, which required k = 25–32 raters to reach G = 0.8 because there was less consistency in their ratings.

Correlations between student and faculty ratings

The second component of our research question concerned whether student–faculty agreement regarding the competencies achieved by a cohort of students who completed a pre-clerkship program was influenced by the scale used. To examine the consistency of student and faculty rankings of the items, Spearman correlations were performed on the item means for each rating scale. Correlations were high in all three cases: IS (r = 0.92), BIS (r = 0.89), and CS (r = 0.82). Re-performing these analyses on the rank ordering of the items for each rating scale yielded the same results: IS (r = 0.93), BIS (r = 0.89), and CS (r = 0.83).

Differences in mean scores

Previous research (Peterson et al., 2012, 2015) has made it clear that aggregated self-assessments should not be used to make absolute judgments about the degree of achievement obtained (i.e., that programs should use this technique only to judge the aspects of competence in which students appear to be relatively weak, thus guiding program development efforts). However, for the sake of comprehensiveness we examined whether changes in the scales used affected the mean scores observed using two-way ANOVA performed on the average score assigned by each participant while treating rater group (student, faculty) and survey version (CS, IS, BIS) as independent variables. A small but statistically significant difference was found for rater group, with students

(M = 3.42, SD = 0.49) assigning higher scores to their own performance than faculty (M = 3.22, SD = 0.57) assigned to them, $F_{1,217} = 9.60$, $p < 0.01$, $\eta^2 = 0.04$, further reinforcing the importance of not trusting absolute values. A medium-sized and statistically significant difference was found for survey version, $F_{2,217} = 7.19$, $p < 0.01$, $\eta^2 = 0.06$. Tukey post-hoc tests showed that ratings were higher when the IS (M = 3.51, SD = 0.48) was used compared to the CS (M = 3.19, SD = 0.47). The BIS (M = 3.34, SD = 0.62) did not differ from the other two scales. There was no statistically significant interaction between rater group and survey version, $F_{2,217} = 0.51$, $p = 0.60$, $\eta^2 = 0.01$.

Differences between students and faculty ratings were also examined separately for each item (see the Appendix). The proportion of items that revealed statistically significant differences was lower when the CS was used (8/40) relative to when either the IS (17/40; $\chi^2 = 4.71$, $p = 0.03$) or the BIS (16/40; $\chi^2 = 3.81$, $p = 0.05$) was used. All three numbers are greater than the number of comparisons that would be expected to be statistically different based on chance alone (2/40; $\chi^2 \geq 4.11$, $p < 0.05$ in each instance). Of the 41 significant differences observed, 40 were in the direction of students rating themselves higher than faculty.

Survey completion rates

Finally, the number of fully completed surveys and the total amount of missing data for each of the three rating scales were examined (see Table 5). A total of

Table 5. Percentages of Missing Data for Each Rating Scale for Student and Faculty on the Readiness for Clerkship Subscales and Scale Score

	Competence-oriented Scale				Independence Scale				Behavioural/ Independence Scale			
	Student		Faculty		Student		Faculty		Student		Faculty	
	(*n* = 53)		(*n* = 59)		(*n* = 44)		(*n* = 65)		(*n* = 38)		(*n* = 61)	
Parameter	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Fully complete surveys	43	81.1	15	25.4	27	61.4	9	13.8	23	60.5	11	18.0
Surveys missing > 3 items	7	9.4	35	59.3	6	13.6	45	69.2	4	10.5	38	62.3
Total items missing[a]	33	1.6	400	17.0	48	2.7	556	21.4	50	3.3	452	18.5
CSKA items missing[b]	6	0.6	152	12.9	11	1.3	214	16.5	4	0.5	172	14.1
WP items missing[b]	27	2.5	248	21.0	37	4.2	342	26.3	46	6.1	280	23.0

Note. CSKA = Clinical Skills and Knowledge Application. WP = Working as a Professional.
[a]The denominator for calculating the percentage was determined by multiplying the sample size for the scale by 40 items.
[b]The CSKA and WP subscales each had 20 items. The denominator for each was determined by multiplying the sample size for the scale by 20 items.

88% of the missing data resulted from respondents selecting an item as "not applicable" as opposed to leaving the item blank (12%). A three-way mixed ANOVA, with rater group and survey version as between-groups variables, and subscale (CSKA, WP) as a repeated variable, showed that missing data were more common for faculty than for students, $F_{1,314} = 146.44$, $p < 0.01$, $\eta^2 = 0.32$ (large effect), for both of the construct-aligned scales relative to the original competence-oriented scale, $F_{2,314} = 3.58$, $p = 0.03$, $\eta^2 = 0.02$ (small effect), and for items in the WP subscale relative to items in the CSKA subscale, $F_{1,314} = 41.52$, $p < 0.01$, $\eta^2 = 0.12$ (medium effect). Small but statistically significant interactions between subscale and rater group and between subscale and survey version showed differences between subscales for both students and faculty and for all three survey versions, with larger differences seen for students and for the two construct-aligned scales, $F_{1,314} = 6.30$, $p = 0.01$, $\eta^2 = 0.02$, and $F_{2,314} = 3.67$, $p = 0.02$, $\eta^2 = 0.02$, respectively.

DISCUSSION

The Readiness for Clerkship survey is a competency-focused instrument used to evaluate the effectiveness of a pre-clerkship medical education curriculum. It is not designed to assess individual students' performance on physician tasks. Rather, aggregated student ratings on this survey are used to differentiate between aspects of competence (the items) by ranking them to identify the relative strengths and weaknesses of a cohort of students for the sake of demonstrating points of need in pre-clerkship training programs. In an effort to determine whether students' judgments in this domain would have more utility (Van Der Vleuten, 1996) if the scales used were manipulated to align with how clinical supervisors conceptualize the performance of trainees (Crossley et al., 2011; Crossley & Jolly, 2012), we compared the data collected using two newly developed (construct-aligned) rating scales to the previously used competence-oriented scale. The fundamental purpose was to determine if shifting the focus of the scales toward aspects of performance that align with examiners' natural cognitive tendencies would reveal apparent changes in the utility of the instrument when used for the sake of program evaluation.

Did the construct-aligned rating scales improve the discrimination between items?

Reliability was evaluated by performing generalizability studies that assessed the ability of each scale to consistently discriminate between items (i.e., to generate a reliable indication of which aspects of competence may require further support within the educational program). Our hypothesis that construct alignment would increase the ability of raters to discriminate between items compared to the competence-oriented scale was supported by the student data, but not by the

faculty data. The number of raters required to reliably (G = 0.8) differentiate between items was lowest for students when the construct-aligned scales were used. For faculty raters, the number of raters required was similar for all three scale versions, leading to the conclusion that construct alignment is preferable because at the very least it does no harm.

Did the construct-aligned scales improve student and faculty agreement?

Consistent with previous findings, we found that the correlation between item averages for faculty and students was very high for each rating scale, demonstrating that student judgments aligned very well with those of faculty raters, despite their apparent overconfidence. This suggests that students can be used as a proxy for faculty opinion as long as the absolute values are not considered (Peterson et al., 2012). That is, priority should be placed on the rank ordering of the items to define the relative strengths and weaknesses of a student cohort to provide direction to continuous quality improvement efforts at the program level, regardless of which scale is adopted.

Did the construct-aligned scales yield a greater rate of survey completion?

The prevalence of unrated items was a prominent issue within the faculty responses. While this might seem at first glance to be a flaw to this approach to evaluation, it is one of the reasons that scales that do not require an extensive number of respondents are valuable. It is quite reasonable for faculty to respond "not applicable" (as was done in most of cases in which they did not provide a rating), given that it is unlikely that any faculty member will observe students performing all of the physician tasks listed in the scale. Faculty supervise students only in their area of clinical specialty, where all the tasks may not be carried out. Similarly, students may also not have had opportunities to engage in all of the tasks at the point at which the survey was completed (e.g., communicate difficult or bad news to your patient), making "not applicable" responses appropriate in some instances. Administering the survey at the end of the clerkship instead of four months into the clerkship could potentially rectify this issue (at least for student evaluations), but we were concerned that a longer time from pre-clerkship would have interfered with students' ability to remember their experiences at the beginning of the clerkship, and the RfC scale is intended as an evaluation of the pre-clerkship training program. In any case, the fact that reliable discrimination between items could be achieved with a reasonably small number of respondents (column k in Table 4) suggests that missing data should not be overly problematic, as most educational programs can feasibly sample more than 8–22 students. Even in the faculty samples, which were less consistent than the student samples, reliable data could be gathered with as few as 25–32 raters.

IMPLICATIONS

The cost of conducting an evaluation, both in terms of financial resources and time, is an important factor to consider for any educational program. The decision studies conducted in this research reveal that half to one third as many student raters are needed to achieve reliable results with the construct-aligned scales relative to the competence-oriented scale. In either case, the data suggest that only a fraction of a typical medical class needs to be recruited to generate stable estimates of relative achievement across competencies. This has important implications for reducing student survey burden, a common problem that causes many to be concerned about the thoughtfulness with which program evaluation surveys are completed. Further, the need to recruit a small sample of students rather than the entire cohort has the potential to reduce the need for systems to be set up or staff time spent to coerce or cajole students into completing the many program evaluation surveys they are often asked to complete. Larger numbers of faculty are needed, but even the upper estimate of 32 is much smaller than the pool of faculty available to most programs. While smaller samples run the risk of not being fully representative of the entire population, the consistency with which reliable measurement has been observed with relatively small numbers across studies suggests that the risk of biased samples skewing the results to be more theoretical than influential. Given the high alignment between student and faculty item rank orders and the poorer quality of the faculty data (lower response rate, greater number of unrated items, and greater number of raters needed), we believe that using student ratings only is a sufficient and more cost-effective method of data collection for the program evaluation purposes outlined in this study (i.e., the identification of aspects of competence that might require greater attention in the formal curriculum). As with any form of evaluation or assessment, no one form of data collection is likely to be sufficient, so we are not suggesting faculty input to be unimportant but rather propose that their time and energy can be saved for other forms of evaluation.

In terms of the more general issue of the value inherent in designing rating scales that are deliberately aligned with the "reality map" of raters, these findings support and extend the ideas of Crossley et al. (2011) by demonstrating empirical replication of their findings (at least within the student sample) in a highly distinct domain (program evaluation relative to workplace-based assessment) and by revealing that the same benefits exist with student raters as were seen previously with faculty raters. This latter finding suggests, consistent with previous work, that students also judge their competence in more general terms like comfort and experience with a task rather than through explicit reference to more formal definitions of competence (Eva, Armson, et al., 2012).

LIMITATIONS

There are a few limitations inherent in this study that should be considered. First, the response rates, especially within the faculty, were lower than desirable and were accompanied by several surveys with unrated items. Both of these limitations carry

the potential that the respondents were not representative of the target population and, indeed, for the faculty sample in particular, these results may not be generalizable beyond those disciplines representing the dominant portion of the sample (i.e., Internal Medicine and Psychiatry). Given the consistency of our findings (both here and in the broader literature) that reliable measurement in this domain can be achieved with fairly small samples, we are confident, however, that our primary conclusions regarding the influence of construct alignment in the student version of the questionnaire are robust (Peterson et al., 2012). Second, "competence" is technically defined as sufficiency for a particular purpose, thereby creating the potential that the labels applied to "levels of competence" in the competence-oriented scale might have been confusing to some raters. If that were true, rather than invalidating the study it would reinforce Crossley et al.'s (2011) argument that construct alignment is important for effectively positioning raters. Finally, while we can claim confidence that item differentiation is reliable using this tool, the fact that data have been collected from only one school makes it impossible to know if the rank ordering of items reflects the specific context of UBC or is representative more broadly of the relative difficulty of achieving any particular competence. We are currently undertaking a multi-institution study to address this limitation.

CONCLUSION

The use of construct-aligned scales improved inter-rater reliability when we examined the capacity of student scores to discriminate between aspects of competence they were expected to have achieved as a result of their pre-clerkship training. As it is this ability to discriminate between items (i.e., rank them from lowest to highest) that we consider to be the most important criterion for facilitating insight in program developers regarding where greater attention might be required, we recommend use of the independence scale and have incorporated this scale version in our continued use of the Readiness for Clerkship survey and in our Readiness for Residency survey (Peterson et al., 2015).

REFERENCES

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education*, *45*(6), 560–569. https://doi.org/10.1111/j.1365-2923.2010.03913.x

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education*, *46*(1), 28–37. https://doi.org/10.1111/j.1365-2923.2011.04166.x

Davies, R.S. (2008). Designing a response scale to improve average group response reliability. *Evaluation and Research in Education*, *21*(2), 134–146. https://doi.org/10.1080/09500790802152209

Davis, D.A., Mazmanian, P.E., Fordis, M., Van Harrison, R., Thorpe, K.E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of

competence: A systematic review. *Journal of the American Medical Association*, *296*(9), 1094–1102. https://doi.org/10.1001/jama.296.9.1094

D'Eon, M., & Eva, K.W. (2009). Self-assessments for workshop evaluations. *American Journal of Evaluation*, *30*(2), 259–261. https://doi.org/10.1177/1098214009334366

D'Eon, M., Sadownik, L., Harrison, A., & Nation, J. (2008). Using self-assessments to detect workshop success: Do they work? *American Journal of Evaluation*, *29*(1), 92–98. https://doi.org/10.1177/1098214007312630

Eva, K.W., Armson, H., Holmboe, E., Lockyer, J., Loney, E., Mann, K., & Sargeant, J. (2012). Factors influencing responsiveness to feedback: On the interplay between fear, confidence, and reasoning processes. *Advances in Health Sciences Education: Theory and Practice*, *17*(1), 15–26. https://doi.org/10.1007/s10459-011-9290-7

Eva, K.W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine*, *80*(Suppl), S46–S54. https://doi.org/10.1097/00001888-200510001-00015

Eva, K.W., Regehr, G., & Gruppen, L.D. (2012). Blinded by "insight": Self-assessment and its role in performance improvement. In B.D. Hodges & L. Lingard (Eds.), *The question of competence: Reconsidering medical education in the twenty-first century* (pp. 131–154). Ithaca, NY: Cornell University Press.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine*, *85*(5), 780–786. https://doi.org/10.1097/ACM.0b013e3181d73fb6

Hauer, K.E., ten Cate, O., Boscardin, C., Irby, D.M., Lobst, W., & O'Sullivan, P.S. (2014). Understanding trust as an essential element of trainee supervision and learning in the workplace. *Advances in Health Sciences Education: Theory and Practice*, *19*, 435–456. https://doi.org/10.1007/s10459-013-9474-4

Kennedy, T.J.T., Regehr, G., Baker, G.R., & Lingard, L.A. (2009). 'It's a cultural expectation…' The pressure on medical trainees to work independently in clinical practice. *Medical Education*, *43*(7), 645–653. https://doi.org/10.1111/j.1365-2923.2009.03382.x

Peterson, L., Eva, K.W., Rusticus, S.A., & Lovato, C.Y. (2012). The readiness for clerkship survey: Can self-assessment data be used to evaluate program effectiveness? *Academic Medicine*, *87*(10), 1355–1360. https://doi.org/10.1097/ACM.0b013e3182676c76

Peterson, L., Rusticus, S.A., Wilson, D.A., Eva, K.W., & Lovato, C.Y. (2015). Readiness for Residency: A survey to evaluate undergraduate medical programs. *Academic Medicine*, *90*, S36–S42. https://doi.org/10.1097/ACM.0000000000000903

Streiner, D.L., & Norman, G.R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford, England: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199231881.001.0001.

Van Der Vleuten, C.P.M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education: Theory and Practice*, *1*(1), 41–67. https://doi.org/10.1007/BF00596229

AUTHOR INFORMATION

Shayna Rusticus is a statistical analyst in the Evaluation Studies Unit, Faculty of Medicine, at the University of British Columbia.

Kevin Eva is a senior scientist at the Centre for Health Education Scholarship, Faculty of Medicine, at the University of British Columbia.

Linda Peterson is a senior evaluator in the Evaluation Studies Unit, Faculty of Medicine, at the University of British Columbia.

Appendix: Means, Standard Deviations, and ANOVA Results for Student and Faculty Comparisons for Each Rating Scale

Item	Competency Scale				Independence Scale				Behavioural/Independence Scale			
	Student	Faculty			Student	Faculty			Student	Faculty		
	M (SD)	*M (SD)*	*F*	*η2*	*M (SD)*	*M (SD)*	*F*	*η2*	*M (SD)*	*M (SD)*	*F*	*η2*
1. Communicate respectfully and effectively with [your/their] patient and their families/support network†	3.68 (0.61)	3.59 (0.62)	0.54	0.00	4.37 (0.62)	3.60 (0.88)	24.97*	0.20	4.03 (0.75)	3.53 (0.73)	10.66*	0.10
2. Identify if [your/their] patient is seriously ill and requires immediate assessment and treatment°	3.02 (0.99)	3.07 (0.53)	0.08	0.00	3.43 (.70)	3.27 (0.84)	1.00	0.01	3.32 (0.74)	3.23 (0.79)	0.24	0.00
3. Identify when [your/their] patient may not be mentally competent and their mental status should be assessed°	3.04 (0.78)	3.07 (0.71)	0.01	0.00	3.18 (0.78)	3.28 (0.69)	0.46	0.01	3.32 (0.67)	3.33 (0.96)	0.00	0.00
4. Take a full medical history°	3.51 (0.75)	3.23 (0.73)	3.96*	0.04	3.98 (0.79)	3.49 (0.90)	8.15*	0.07	3.68 (0.90)	3.23 (0.80)	6.66*	0.07
5. Take an appropriate history of the current problem°	3.42 (0.72)	3.12 (0.77)	4.29*	0.04	3.68 (0.88)	3.36 (0.84)	3.59	0.03	3.45 (0.86)	3.23 (0.92)	1.38	0.01
6. Formulate a problem list°	2.81 (0.83)	2.89 (0.76)	0.19	0.00	3.09 (0.91)	3.03 (0.85)	0.12	0.00	2.97 (0.72)	2.95 (0.92)	0.02	0.00
7. Perform a full physical examination°	3.30 (0.64)	2.96 (0.62)	7.53*	0.07	3.32 (0.74)	3.23 (1.01)	0.20	0.00	3.42 (0.68)	3.09 (0.63)	5.46*	0.06
8. Document the history and physical exam findings°	3.23 (0.64)	3.23 (0.68)	0.00	0.00	3.75 (0.78)	3.55 (0.91)	1.29	0.01	3.47 (0.69)	3.18 (0.88)	3.04	0.03
9. Verbally present [your/their] findings to the resident or [your preceptor/you]°	2.58 (0.91)	3.13 (0.72)	11.86*	0.10	3.18 (1.00)	3.29 (0.73)	0.42	0.00	2.84 (0.79)	3.20 (0.90)	4.05*	0.04

doi: 10.3138/cjpe.31094

10.	Interpret relevant key laboratory results obtained on your patient°	2.89 (0.70)	2.94 (0.66)	0.18	0.00	2.80 (0.82)	3.06 (0.82)	2.43	0.03	2.97 (0.72)	3.02 (0.82)	0.08	0.00
11.	Interpret (explain the meaning of) relevant imaging reports for the common health problems of [your/their] patient°	2.51 (0.93)	2.76 (0.63)	2.39	0.02	2.59 (1.06)	2.98 (0.79)	3.74	0.04	2.79 (0.70)	2.73 (0.78)	0.12	0.00
12.	Explain the underlying pathology and pathophysiology of [your/their] patients' key problems°	3.30 (0.58)	2.82 (0.64)	17.01*	0.14	3.16 (0.71)	2.71 (0.91)	7.10*	0.07	3.16 (0.72)	2.81 (0.79)	4.63*	0.05
13.	Demonstrate a clear understanding of anatomy in the context of physical exams and interventions°	3.26 (0.74)	2.88 (0.81)	6.39*	0.06	3.43 (0.70)	2.98 (0.92)	6.72*	0.07	3.29 (0.65)	2.94 (.89)	4.14*	0.05
14.	Propose a differential diagnosis consisting of more than one reasonable alternative (based on Hx, PE, Lab, and other tests)°	3.00 (1.02)	2.71 (0.71)	2.92	0.03	2.95 (0.81)	2.86 (0.81)	0.31	0.00	3.05 (0.87)	2.84 (0.99)	1.23	0.01
15.	Communicate difficult or bad news to [your/their] patient†	2.85 (0.92)	2.82 (0.68)	0.02	0.00	3.16 (0.93)	2.55 (0.81)	8.27*	0.11	3.06 (1.00)	2.93 (1.07)	0.25	0.00
16.	Identify appropriate medications based on the clinical problems of [your/their] patient°	2.70 (0.89)	2.61 (0.75)	0.31	0.00	2.48 (0.79)	2.70 (0.77)	1.96	0.02	2.63 (0.97)	2.61 (0.95)	0.01	0.00
17.	Explain the choice of medication based on mechanism of action and the clinical problems of [your/their] patient°	2.74 (0.84)	2.45 (0.75)	3.29	0.03	2.52 (0.99)	2.63 (0.80)	0.31	0.00	2.53 (0.92)	2.54 (0.94)	0.01	0.00
18.	Propose a basic short-term management plan for [your/their] patient's major problems°	2.91 (0.77)	2.86 (0.62)	0.13	0.00	2.70 (0.80)	2.82 (0.81)	0.49	0.00	2.71 (0.77)	2.90 (0.97)	1.04	0.01

(Continued)

doi: 10.3138/cjpe.31094

Appendix: Continued

Item	Competency Scale				Independence Scale				Behavioural/Independence Scale			
	Student	Faculty			Student	Faculty			Student	Faculty		
	M (SD)	*M (SD)*	*F*	*η2*	*M (SD)*	*M (SD)*	*F*	*η2*	*M (SD)*	*M (SD)*	*F*	*η2*
19. Explain the short-, intermediate- and long-term management plans that were developed for [your/their] patient°	2.92 (0.74)	2.71 (0.72)	2.11	0.02	2.77 (0.77)	2.81 (0.87)	0.05	0.00	2.82 (0.77)	2.75 (0.97)	0.14	0.00
20. Identify the specific physical and psychosocial needs of [your/their] patient°	3.18 (0.74)	3.00 (0.57)	1.79	0.02	3.34 (0.81)	2.95 (0.83)	5.71*	0.05	3.30 (0.85)	3.16 (0.92)	0.50	0.01
21. Advocate for access to required health and social services based on [your/their] patient's needs†	2.82 (0.95)	2.93 (0.69)	0.36	0.00	3.02 (1.00)	2.88 (0.95)	0.51	0.01	3.11 (1.02)	3.02 (1.00)	0.15	0.00
22. Explain the concept of making patient care decisions, based upon efficient and equitable allocation of health care resources°	2.90 (0.81)	2.97 (0.69)	0.19	0.00	2.90 (0.89)	2.85 (0.94)	0.08	0.00	3.17 (0.85)	2.85 (0.95)	2.34	0.03
23. Demonstrate compassion for and interest in [your/their] patient†	3.86 (0.69)	3.73 (0.69)	1.03	0.01	4.52 (0.82)	4.07 (0.83)	7.95*	0.07	4.35 (0.72)	3.84 (0.86)	9.35*	0.09
24. Show personal commitment to honouring the choices, rights, and confidentiality of [your/their] patient†	3.81 (0.62)	3.58 (0.72)	3.17	0.03	4.47 (0.77)	3.98 (0.86)	8.38*	0.08	4.19 (0.70)	3.70 (0.88)	8.40*	0.08
25. Act only within the limits of [your/their] competence†	3.73 (0.80)	3.48 (0.68)	3.10	0.03	4.36 (0.72)	3.90 (0.84)	8.75*	0.08	4.11 (0.91)	3.72 (0.94)	4.08*	0.04
26. Disclose errors or adverse events in a full, honest, and timely manner†	3.44 (0.71)	3.28 (0.75)	0.95	0.01	4.38 (0.85)	3.75 (1.00)	7.83*	0.11	3.97 (0.76)	3.34 (1.01)	7.42*	0.10

doi: 10.3138/cjpe.31094

27. Respond to pages, patient, and team needs in a timely manner†	3.62 (0.80)	3.46 (0.78)	0.99	0.01	4.49 (0.80)	4.00 (0.80)	8.48*	0.09	4.25 (0.69)	3.74 (0.91)	8.12*	0.09
28. Arrange for care of [your/their] patient during absences to ensure continuity of care†	3.27 (0.91)	3.03 (0.98)	1.11	0.01	4.06 (0.94)	3.68 (0.90)	2.62	0.04	3.69 (0.93)	3.09 (1.15)	5.49*	0.08
29. Communicate respectfully and effectively [your/their] peers, residents, and [faculty/you]†	3.65 (0.71)	3.64 (0.79)	0.01	0.00	4.55 (0.70)	4.11 (0.82)	8.19*	0.07	4.19 (0.78)	3.93 (0.90)	2.05	0.02
30. Communicate respectfully and effectively with other health care professionals†	3.62 (0.77)	3.61 (0.73)	0.00	0.00	4.45 (0.73)	4.19 (0.82)	2.96	0.03	4.06 (0.84)	3.97 (0.83)	0.27	0.00
31. Prevent or resolve conflicts with members of the health care team†	3.34 (0.82)	3.30 (0.65)	0.05	0.00	4.11 (0.81)	3.90 (0.71)	1.22	0.02	3.80 (0.81)	3.53 (0.90)	1.46	0.02
32. Listen to and act on feedback†	3.63 (0.72)	3.40 (0.56)	3.57	0.03	4.36 (0.69)	3.95 (0.80)	7.57*	0.07	4.11 (0.84)	3.76 (0.82)	3.97*	0.04
33. Identify gaps in [your/their] knowledge and skills related to the care of your patients†	3.60 (0.66)	3.33 (0.61)	4.81*	0.04	4.20 (0.63)	3.53 (0.74)	23.60*	0.19	3.97 (0.65)	3.52 (0.80)	8.48*	0.08
34. Take steps to correct gaps in knowledge and skills†	3.58 (0.63)	3.30 (0.64)	5.25*	0.05	4.27 (0.59)	3.67 (0.77)	18.23*	0.16	3.97 (0.56)	3.56 (0.89)	6.07*	0.07
35. Pursue opportunities to learn the required technical procedures†	3.25 (0.78)	3.29 (0.65)	0.10	0.00	3.89 (0.84)	3.76 (0.85)	0.47	0.01	3.65 (0.86)	3.44 (0.99)	0.97	0.01
36. Retrieve and critically evaluate relevant information for the purposes of patient care, scholarly inquiry, and self-directed learning†	3.43 (0.61)	3.22 (0.88)	2.08	0.02	3.86 (0.77)	3.41 (0.90)	7.10*	0.07	3.57 (0.65)	3.17 (1.13)	3.63	0.04

(Continued)

doi: 10.3138/cjpe.31094

Appendix: Continued

Item	Competency Scale				Independence Scale				Behavioural/Independence Scale			
	Student	Faculty			Student	Faculty			Student	Faculty		
	M (SD)	*M (SD)*	*F*	*η2*	*M (SD)*	*M (SD)*	*F*	*η2*	*M (SD)*	*M (SD)*	*F*	*η2*
37. Educate [your/their] patient and their families/support network about their clinical problems and management plan[†]	3.17 (0.83)	3.02 (0.75)	0.89	0.01	3.57 (0.93)	3.13 (0.73)	6.54*	0.07	3.51 (0.69)	3.16 (0.91)	3.75	0.05
38. Facilitate the learning of [your/their] peers[†]	3.30 (0.76)	3.19 (0.57)	0.55	0.01	3.83 (0.77)	3.52 (0.77)	2.92	0.04	3.57 (0.82)	3.13 (0.92)	4.28*	0.06
39. Manage [your/their] time effectively in a clinical setting[†]	2.89 (0.87)	3.15 (0.75)	2.83	0.03	3.56 (0.88)	3.26 (0.91)	2.75	0.03	3.19 (0.91)	3.09 (0.97)	0.26	0.00
40. Maintain [your/their] health and well-being, and know when and how to seek help[†]	3.00 (0.86)	3.15 (0.61)	2.58	0.03	3.67 (0.84)	3.68 (0.67)	0.00	0.00	3.33 (0.89)	3.68 (0.75)	3.16	0.04

Note. Words in brackets differentiate the student and faculty survey versions, respectively.
[†]Working as a Professional subscale.
[°]Clinical Skills and Knowledge Application subscale.
* $p < 0.05$

doi: 10.3138/cjpe.31094