JAHE | The Journal of Assessment in Higher Education

Volume 2 | Number 1 January 2021

https://journals.flvc.org/assessment/ ISSN: 2642-0694

Rating Scale Data and its Utility: Additional Evidence to be Explored in the Medical Education Continuum

Chosang Tendhar

Hackensack Meridian School of Medicine

Abstract

The purposes of this study are to assess the utility of self-appraisals and ratings of program directors (PDs) and to introduce new ways to use self-assessed and rating scale data. The data for this study was collected from graduates of Baylor College of Medicine (BCM) who were enrolled in different residency programs around the country. The interns and PDs completed a similar set of questions. The correlation between the ratings of the two group was .21. The Cronbach's alpha of interns and PD surveys were .89 and .97, respectively. The interns consistently rated themselves lower compared to ratings the PDs assigned them. The two groups agreed on the areas of strengths and weaknesses based on their mean ratings and rank-ordering of competencies. This study proposes that lowest mean ratings of measures that appear at the bottom in the rank-ordering be considered as areas that deserve special attention. The results of this study brought validity evidence to the utility of self-appraisals and PD's ratings of interns.

Keywords: Construct Validity, Intern Survey, PD Survey, Rating scale

* Corresponding author: Chosang Tendhar, Office of Institutional Effectiveness and Assessment, Hackensack Meridian School of Medicine, 123 Metro Blvd., Nutley, NJ 07110. Chosang.tendhar@hmhn.org

Recommended citation: Tendhar, C. (2021). Rating scale data and its utility: Additional evidence to be explored in the medical education continuum. *Journal of Assessment in Higher Education, 2*(1), 54-62. doi:10.32473/jahe.v2i1.121588

©Journal of Assessment in Higher Education. Published under Creative Commons License CC BY- 4.0 International.

Introduction

The Liaison Committee on Medical Education (LCME) stipulates that medical schools collect relevant data from their recent graduates and program directors (PDs) to gauge graduates' attainment of competencies in a variety of areas, such as medical knowledge, patient care, and system-based practice. This requirement reflects on the significance of programmatic evaluation. The evaluation of medical schools' graduates in different residency programs serves three important purposes: (1) to help a program determine how effective it has been in imparting quality education; (2) to ascertain if graduates of the program are ready for their residency and beyond; and, (3) to help continuous quality improvement endeavors of the program.

The school has an obligation to achieve these three purposes. It is the societal responsibility of the medical schools to graduate students who are equipped with all the major competencies at an expected level of proficiency. In other words, it is the duty of the program to produce physicians who are "practice-ready." We collect data from our recent graduates and their residency PDs every year through online surveys to assess how the school has prepared its graduates. The two groups are surveyed before the end of the first year of the interns' residency programs.

Findings show that not every graduate fully meets the Medical School Objectives Project (MSOP) standards (Lypson, Frohna, Gruppen, & Woolliscroft, 2004). Similarly, not all residents are reasonably equipped with the expected competencies of the Accreditation Council for Graduate Medical Education (ACGME). The gaps in expected versus observed skills of medical students globally have been established.

The interns and PDs completed similar questionnaires. The interns were asked to rate themselves on 15 different measures, specifically by comparing themselves to interns in the same residency programs but who are from different schools. Similarly, the PDs were asked to rate graduates of this program on those same 15 measures, comparing them to interns from other schools. Example items included a participant's medical knowledge critical to the care of patients (Intern Survey) and the PD's assessment of an intern's ability to perform a physical examination (PD Survey).

More often than not, some kind of rating scales would be used to assess graduates' competencies because a binary response choice (e.g., yes or no) may not fully capture the information researchers are seeking and would be too simplistic an assessment for these purposes. It is presumed that most medical school graduates would possess and demonstrate certain levels of competency, but the bigger question has always been if their competencies are below, equivalent to, or beyond the expected level that demonstrates graduate proficiency.

As a result of recent developments with Entrustable Professional Activities (EPAs) in graduate medical education (GME), there is interest in developing similar EPAs in undergraduate medical education (UME) (Chen, Broek, Cate, 2015). EPAs are a set of activities that entering residents should be expected to perform on day one of residency without direct supervision. In the medical education continuum, students go through

different phases: from being students in UME to residents in GME, to fellowships or employment. Angus et al. (2016) emphasized the importance of sharing competencies at each transition point along the educational continuum. Such sharing of information will allow any gaps in knowledge and skills to be effectively addressed at an appropriate stage in the medical education continuum before graduates begin unsupervised practice.
It is, therefore, timely to begin a dialogue on the use of rating scale data in order to contribute to the current discussion of EPAs. Towards this goal, this study proposes a new way to examine the ubiquitous rating scale data. Many business organizations and academic institutions assess performances of their employees or different stakeholders using rating scales. For example, in many medical schools, students are asked to evaluate their instructors' effectiveness of teaching using five- or six-point Likert-type scale. The Baylor College of Medicine (BCM) Intern and PD surveys used a five-point Likert-type scale.

Unfortunately, there is a lot of skepticism about using rating scale data in general, and Intern and PD surveys data in particular, within the medical education community. Some of the well-known concerns about study participants' response bias to a scale are acquiescence, social desirability, random responding, and guessing. Furthermore, some of the concerns over the quality of PD survey data includes the perceptions that PDs do not have a sufficient amount of interactions with all of the interns under their supervision. That PDs have too many supervisees and are assigned numerous administrative roles for them to be able to accurately assess their interns' skills. Thus, these concerns lead to the perception that the ratings PDs assign their interns may not reflect their ratees' true skills, knowledge, and abilities.

According to Ross (2006), ratees may have inflated perceptions of their knowledge and skills, and similarly, their ratings may be motivated by self-interest (i.e., grades). However, ratees are less likely to overestimate their skills, and provide more accurate self-assessments when they perceive that they do not necessarily have much to gain from their self-assessments (Ross, 2006; Mabe & West, 1982). In this particular survey, the interns' overall scores were not intended to reflect on their academic performance.

Neither the PDs nor the program's graduates were asked to rank-order their competencies because it may have led to more response bias issues. However, when the graduates rated themselves on a variety of competencies, it was unlikely that every graduate would be perceived to have attained a similar level of competencies in all of the areas. Similarly, when the PDs rated graduates of this program, it was likely that there were variations in the ratings they assigned the interns on those 15 measures. Therefore, two purposes of this exploratory study were, 1) to assess the utility of self-appraisals and ratings of PD using the intern and PD surveys data; and, 2) to introduce new ways to use self-assessed and rating scale data.

Method

A survey was administered to 172 PDs and 181 interns who graduated in 2015. Eighty interns responded but 10 incomplete responses were deleted and this resulted in an effective response rate of 39% for interns. On the other hand, seventy nine PDs responded resulting in a response rate of 46%. Three reminders were sent to both groups. The author included 50 interns in the final analyses who had evaluation data from their PDs.

Both groups received the same survey but items were tailored to each group. Interns were asked to compare themselves to non-BCM interns in their residency programs on a range of competencies. The PDs were asked to rate BCM interns on those same measures to non-BCM interns. Fifteen items that appeared on the interns and PDs' surveys are presented in Appendix A and B, respectively. A five-point Likert-type scale was used ranging from 1 *(significantly below other interns)* to 5 *(significantly above other interns)* in both surveys.

The Statistical Package for the Social Sciences (SPSS) version 24 was used to compute descriptive statistics, scale reliabilities, and perform paired-samples *t*-tests.

Results

Table 1 presents the descriptive statistics based on the average of 15 items and the reliabilities of the scales. The average scores of interns' ranged from 2.53 to 4.67, while the average scores PDs assigned their interns ranged from 1.73 to 5.00. The Cronbach's alpha coefficient for each scale was .89 for the Intern survey and .97 for the PD survey.

Table 1

Descriptive Statistics (based on average of 15 items) and Reliabilities

	Min	Max	Mean	Std.Dev.	α
Interns	2.53	4.67	3.47	.46	.89
Program Directors	1.73	5.00	3.86	.79	.97

Table 2 presents the mean scores of individual items for each group; mean differences between the two groups; significance values of the *t*-tests; and, the correlations of the items. The mean scores of interns on the one to five scale ranged from 3.05 to 4.10. On the other hand, the mean scores PDs assigned to their interns ranged from 3.65 to 4.12. The interns consistently rated themselves lower than the PDs rated them. The mean differences were statistically significant for all of the items except for three items (#7, #11 and #12).

Table 2

Means of Individual Items, Mean Differences between Interns and PDs, Significance Values of Paired Sample T-Tests, and Correlations

Items	Interns	PDs	Mean Difference	*p values*	*Correlations Interns and PDs*
1	3.48	3.86	-0.38	0.02	0.22
2	3.52	3.82	-0.30	0.02	0.29
3	3.22	3.68	-0.46	0.01	0.28
4	3.43	3.84	-0.41	0.01	0.36
5	3.48	3.82	-0.34	0.02	0.12
6	3.28	3.68	-0.40	0.03	0.06
7	4.10	4.12	-0.02	0.88	0.25
8	3.30	3.86	-0.56	0.01	0.30
9	3.05	3.65	-0.60	0.01	-0.01
10	3.31	3.94	-0.63	0.01	0.20
11	3.69	3.73	-0.04	0.80	0.17
12	3.63	3.75	-0.12	0.44	0.18
13	3.60	4.10	-0.50	0.01	0.04
14	3.50	4.10	-0.60	0.01	0.08
15	3.48	4.12	-0.64	0.01	0.10
Average	3.47	3.86	-0.39	0.01	0.21

Of the 15 items, there was a broad agreement between the two groups on the highest and lowest rated items. The two items that received the lowest ratings from both the groups were item #9 (intern's skills to critically analyze the literature and apply to patient care) and item #3 (ability to perform physical examination). In terms of the ratings PDs assigned, items #3 and #6 (ability to provide evidence of decisions) had the same mean scores.

The two items that the interns rated highest were #7 (interpersonal and communication skills in caring for diverse patients) followed by #11 (leadership skills required for patient care). The PDs also assigned the highest rating to item number seven, in addition to item number 15 (the intern's overall level of professionalism). The second highest rated items for PDs were #13 (level of professionalism in dealing with patients) and item #14 (level of professionalism in dealing with colleagues). There was a broad agreement between the two groups on the strengths that graduates exhibited in the areas of soft skills.

The correlations between the intern and PD surveys on these 15 measures, presented in Table 2, ranged from -.01 to .36. The correlation between the two groups on the average score (average of 15 items) was .21.

Discussion

This study provided evidence of validity in using rating scale data. Furthermore, the findings of this study should minimize skepticism over the quality of intern and PD surveys data in the medical education community. Two indicators were used to ascertain agreements between interns' and PDs' ratings of competencies: (1) mean level agreement; and (2) correlation between interns and PDs ratings. Both indicators suggested a reasonable degree of agreement and consistent with the literature of intern self-rating and supervisor rating of intern performance (Harris & Schaubroeck, 1988; Heidemeier & Moser, 2009).

The rank-ordering of competencies based on ratings made it possible to ascertain mean level agreements and specifically, to determine the items that received highest and lowest mean ratings. Based on the results presented in Table 2, interns' self-evaluations of their competencies compared with the ratings PDs assigned them, yielded a similar ranking, particularly items that received highest and lowest mean scores. This was a strong indication of a broad agreement between the two types of assessment processes. Essentially, the program's graduates' perceptions of strengths and weaknesses in different areas of competencies were validated by their PDs perceptions.

The items that received lowest mean ratings were items #9, #3 and #6. These items are indicators of clinical competence and evidence-based medical practice. This finding is perhaps not surprising because completing literature reviews, meta-analyses/syntheses, rigorous research methods courses, and full-fledged research papers, are often not a significant part of the MD curriculum. Therefore, building on prior curricula, developing study materials and providing rigorous training for interns in areas covered during their residency years could make a significant difference in addressing interns' knowledge gaps in those areas of competencies. There may also be room in the UME to enhance and strengthen the curriculum in those areas. The lack of these skills, if not addressed effectively at an appropriate stage in the medical education continuum, could be detrimental to the ability of future physicians in becoming intelligent consumers of research. This, in turn, could severely impede many physicians from becoming life-long learners.

The items that received highest ratings were items #11, #15, #13, and #14. These items are essentially indicators of professionalism and leadership skills. These results suggest that BCM graduates have strong professionalism and leadership skills. The BCM graduates, therefore, may become good role models of high professionalism for future physicians. Furthermore, it can be inferred from the findings of this study that graduates of this program have high probability of assuming future leadership positions, such as departmental chairs, clerkship directors, and deans.

The mean scores of individual items could reflect on the BCM program that awards MD degrees. With polytomous data, it is rare to come across means below the mid-point of the scale, typically “neither agree nor disagree” but in this study, represented by “equal to other interns.” In the spirit of continuous quality improvement, irrespective of how high or low the mean score was, this study suggests that program leaders pay special attention to competencies that received lower ratings. Therefore, this new approach of rank-ordering of items based on the ratings would allow UME leadership to identify areas of their programs’ strengths and weaknesses.

In identifying interns whose performances are not meeting the expected standard, we propose that mean scores reflecting acceptable competencies range from 60% to 70% of the rating scale used. In this study, a five-point Likert-type scale was used. Therefore, 60% would translate into a mean score of 3.0 and 70% would translate into a mean score of 3.5. This suggestion is in keeping with cut-off scores normally used for pass in many preclinical courses and board exams. As Table 1 demonstrated, interns’ ratings ranged from 2.53 to 4.67 compared to the ratings that the PDs assigned them which ranged from 1.73 to 5.00. These mean scores suggest that there are some interns who should work harder and need additional support to become “practice-ready” physicians by the completion of their residency.

The interns’ and PDs’ ratings yielded an overall correlation of .21. This suggests that there exists a moderate degree of agreement between interns and PDs ratings. This degree of agreement is consistent with findings in multiple meta-analytic studies that demonstrated that the average correlation found between subordinate and supervisor ratings was .22 (Harris & Schaubroeck, 1988; Heidemeier & Moser, 2009). The correlation between self-evaluations and more objective measures, such as supervisor ratings, tend to hover around .20.

In addition to a typical correlation, other findings of this study are consistent with some of the existing literature. For example, Distlehorst, Dawson, & Klamne (2009) found that interns and PDs provided higher ratings to items that measure soft skills and lower ratings to items that measure clinical competence. Furthermore, the finding that PDs rated interns consistently higher than the interns rated themselves (Table 2) were echoed in other studies (Distlehorst et al., 2009; Moores & Durnning, 2007). The research continues to show that interns grade themselves lower than their PDs. These findings suggest that response biases are not of significant concern especially in the context of an intern’s self-assessment.

Park, Zar, Norcini, and Tekian (2016) examined data of Internal Medicine residents in a longitudinal study (postgraduate year 1—PGY1—through PGY3). They recommended that any milestones that did not grow in subsequent years should be investigated to identify factors that impeded growth. Similarly, Moores and Durning (2007) administered an identical survey to their graduates in a fellowship program and to the supervisors of those graduates. The items were intended to assess fellows’ preparation for a variety of tasks, for example, to perform invasive procedures, oversee a pulmonary function lab, and use the medical literature. The items that consistently received a low mean score were used as a basis for curricular reform. It is argued that the author’s proposal of rank-ordering of

competencies based on ratings and suggestion that those competencies with lowest mean scores should be considered as areas for further improvement is conceptually similar to the recommendation of Park et al., (2016) and Moores and Durning (2007).

In conclusion, the findings of this study should ameliorate concerns over the quality of rating scale data. Many concerns arise from psychometric perspectives. However, the two indicators used to ascertain agreement between the raters established construct validity. The mean level agreement between interns and PDs demonstrated that mean scores can be meaningfully used to infer strengths and weaknesses of programs as well as interns. There was also a moderate degree of correlation between the two raters. Therefore, scores from these scales can be confidently used in making judgments about competencies of future physicians along the educational continuum in addition to other assessment data. Heeding the recommendation of Angus et. al., (2016), these findings should be shared with graduate medical education so that those interns with unsatisfactory mean scores could be provided with additional help.

Limitations and Future Research

This study is not without its limitations. First, a larger sample size would permit additional statistical analyses such as comparing and contrasting degree of agreement among different residency programs. Second, demographic data (e.g., sex and race) were not collected, which prevented analyses of response patterns among different groups. Third, following career trajectories of the program's graduates to investigate whether or not they assumed leadership roles would establish further validity evidence, or lack thereof, to the inference drawn from high mean scores. Fourth, the level of agreement between raters was assessed at the individual item level. Future studies should test the level of agreement at the factor level. Fifth, non-respondents were not followed up with, therefore, it is hard to determine whether or not the findings of this study generalize to them. Sixth, this study should be cross-validated on a different set of sample, preferably with a larger sample size. If the order of the competencies, mean level agreement, and correlation found in this study are supported by similar results in future studies, that essentially brings evidence of construct validity. Finally, addressing these issues in future research has the potential to advance knowledge in the areas of utility of rating scale data and strengthening communication along the medical education continuum.

Acknowledgments: The author wishes to acknowledge contributions to an early version of this work by Dr. Joel Purkiss, Assistant Dean of Evaluation, Assessment, and Education Research at Baylor College of Medicine.

References

Angus, S. V., Vu, T. R., Willet, L. L., Call, S., Halvorsen, A. J., & Chaudhry, S. (2016). Internal medicine residency program directors' views of the core entrustable professional activities for entering residency: An opportunity to enhance communication of competency along the continuum. *Academic Medicine*, Advanced online publication.

Chen, H. C., Broek, S., & Cate, O. (2015). The case for use of entrustable professional activities in undergraduate medical education. *Academic Medicine, 90*(4), 431-436. doi: 10.1097/ACM.0000000000000586

Distlehorst, L. H., Dawson, B. K., & Klament, D. L. (2009). Supervisor and self-ratings of graduates from a medical school with a problem-based learning and standard curriculum track. *Teaching and Learning in Medicine, 21*(4), 291-298. doi:10.1080/10401330903228364

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*(1), 43-62. doi:10.1111/j.1744-6570.1988.tb00631.x

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology, 94*(2), 353-370.

Lypson, M. L., Frohna, J. G., Gruppen, L. D., & Woolliscroft, J. O. (2004). Assessing residents' competencies at baseline: Identifying the gaps. *Academic Medicine, 79*(6), 564-570.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280-296.

Moores, L., & Durning, S. (2007). The feasibility and construct validity of graduate and supervisor survey in a pulmonary fellowship training program. *Teaching and Learning in Medicine, 19*(1), 70-74. doi:10.1080/10401330709336627

Park, Y. S., Zar, F. A., Norcini, J. J., & Tekian, A. (2016). Competency evaluations in the next accreditation system: Contributing to guidelines and implications. *Teaching and Learning in Medicine, 28*(2), 135-145. doi:10.1080/10401334.2016.1146607

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation, 11*(10), 1-13.