American Journal of Educational Research, 2018, Vol. 6, No. 1, 38-42
Available online at http://pubs.sciepub.com/education/6/1/6

DOI:10.12691/education-6-1-6

Psychometric Validation of a Medical and Health Professions Course Evaluation Questionnaire

Kenneth D. Royal[1,*], Leigh Jay Temple[2], Jennifer A. Neel[3], Laura L. Nelson[1]

[1]Department of Clinical Sciences, North Carolina State University, Raleigh, NC, USA
[2]Education Support Services, North Carolina State University, Raleigh, NC, USA
[3]Department of Population Health & Pathobiology, North Carolina State University, Raleigh, NC, USA
*Corresponding author: kdroyal2@ncsu.edu

Abstract Education in the medical and health professions is considerably different than education in most higher education contexts (e.g., multiple instructors, clinical focus, student cohorts, etc.). Thus, despite a mature research literature relating to course and instructor evaluations in higher education, there largely is an absence of such literature in medical and health professions education. This study sought to contribute to the medical and health professions education literature by: 1) introducing a new instrument for evaluating course effectiveness; 2) demonstrating how to conduct a state-of-the-art psychometric validation study of an instrument's psychometric properties; and 3) providing a framework for interpreting construct validity evidence. Results of the validation study indicated a considerable amount of construct validity evidence is discernible to conclude the instrument is capable of producing valid and reliable measures of course quality and effectiveness. Other medical and health professions educators are encouraged to adopt the instrument for use at one's own campus, and/or replicate the validation procedures on other survey instruments.

Keywords: *medical education, health professions education, veterinary education, evaluation, measurement, survey validation, psychometrics, course evaluation*

Cite This Article: Kenneth D. Royal, Leigh Jay Temple, Jennifer A. Neel, and Laura L. Nelson, "Psychometric Validation of a Medical and Health Professions Course Evaluation Questionnaire." *American Journal of Educational Research*, vol. 6, no. 1 (2018): 38-42. doi: 10.12691/education-6-1-6.

1. Introduction

Course evaluations are conducted in virtually every medical and health professions program in the world. Results of these evaluations are used to inform instructors and college/university administrators about various aspects of a course that are both working well and may need modification. [1] Unlike student evaluations of teaching (SETs) in which the instructor's performance is the focus of the evaluation, course evaluations focus on those factors relating to course quality and effectiveness (e.g., organization of course content, quality of instructional materials, quality of the assessment instruments, etc.). Although course evaluation items are distinct from instructor evaluation items, many institutions combine both elements onto a common form for convenience.

Although the literature on course and instructor evaluations is well-researched in higher education, this is not quite the case for the medical and health professions. Further, many of the contextual factors common to higher education (broadly defined) do not extend to education in the medical and health professions. [2] For example, most higher education courses are taught by a single instructor, whereas courses in the medical and health professions often are taught by multiple instructors. A number of differences also exist with respect to the educational curriculum. Medical and health professions education uniquely focuses on the development of a defined professional role (clinical practice). Curricula in health professions education is much more static, with limited options for course and instructor selection. [3] Courses are typically conducted in a cohort format, concurrently enrolling all members of a class in the same courses each term. The enrollment of larger numbers (often 100 or more) of students may affect teaching formats and relationships between faculty and students. [4] Class size, tradition, and team-taught courses have each contributed to the common practice of assessment via multiple-choice examinations, which may influence students perspectives about course difficulty and fairness, particularly as it pertains to subjective versus objective grading. [5]

As with any newly developed instrument, the psychometric properties of the instrument should be thoroughly appraised for quality, as it is critical that these evaluations yield valid measures. According to the Standards for Educational and Psychological Testing, a joint publication by the American Psychological Association (APA), American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME), validity refers to the collective evidence (both theoretical and empirical) that supports the intended use and interpretation of scores. [6] However, there is a distinction between validity and validation. Royal [7] notes "Whereas validity refers to a

conceptual framework for interpreting validity evidence, validation refers to the practice of incorporating and applying validity theory to evaluate evidence" (p. 569).

For the present study, we performed a validation exercise to evaluate the psychometric properties and functioning of a newly designed instrument intended to measure course quality and effectiveness. Because a poor quality instrument likely will yield invalid results, it is critical to conduct a validation study shortly after an instrument's initial use. Findings should then be used to inform decisions about any changes that may improve instrument quality. Similarly, because a high-quality instrument likely will yield valid scores, it is important to have psychometric evidence to support instrument quality and to justify the continued use of an instrument. Using state-of-the-art item response theory (IRT) modeling, we will thoroughly inspect various properties of the instrument, address its strengths and weaknesses, and use the accumulation of validity evidence to make an evaluative judgment about the instrument's viability as a quality tool for evaluating course quality and effectiveness.

2. Methods

2.1. Participants and Setting

Participants included Doctor of Veterinary Medicine (DVM) students at a large public university located in the United States. All students in the didactic portion of the curriculum (years 1-3) were invited to complete evaluations for each course in which they were enrolled. The college's DVM program consists of a total of 300 students (100 per class cohort) and has a female to male ratio of 4:1. Of the participants who completed the course evaluations, 1,664 (80.8%) were female and 396 were male (19.2%), thus nearly perfectly resembling the population of students in the college.

A total of 29 courses were evaluated, resulting in 2,060 student ratings. Course enrollment figures ranged from 19 to 102 students, with a mean course enrollment of 92.6 (SD = 23.6, Med = 101) students. The mean number of students across all courses was 71.0 (SD = 18.1, Med = 77), resulting in a 77.10% (SD = 4.28%, Med = 78.0%) overall response rate.

2.2. Instrumentation

A team of education and veterinary medical experts created the instrument as a mean to address a growing problem of student survey fatigue within the college. The instrument was intended to be short, accessible and capture only the most salient information relating to course quality and effectiveness. The instrument begins with a section devoted to course-specific learning outcomes. That is, course coordinators are required to provide a list of course-level learning outcomes (typically 3-5) that are uniquely populated on each questionnaire. The remainder of the instrument, however, consists of seven items that appear on all course evaluations (see Table 1). The instrument utilizes a 4-point Likert-type rating scale for all quantitative item with categories: 1 = Strongly Disagree; 2 = Disagree; 3 = Agree; and 4 = Strongly Agree. The instrument concludes with an open-ended item that reads: "Please provide constructive comments (e.g., aspects of the course that should be continued, needs improved, etc.)." The instrument is administered electronically during the last two weeks of the course, but prior to assigning student grades in an effort to prevent response bias stemming from grade awareness.

Table 1. Descriptive statistics for the course evaluation items

Item	Mean	SD
Q1. The knowledge obtained was proportional to the effort I invested in the course.	3.37	0.66
Q2. The individual components of this course were part of a cohesive whole.	3.37	0.64
Q3. The course materials were well prepared and organized (e.g., website, notes, etc.)	3.31	0.69
Q4. Previous course work adequately prepared me for this course.	3.22	0.73
Q5. The assessments were a good measure of my learning.	3.21	0.74
Q6. The assessments provided appropriate measure of the course outcomes.	3.27	0.69
Q7. Overall, this course was effective.	3.36	0.63

2.3. Validation Framework

Research [8] has noted six major weaknesses associated with traditional statistical analyses of survey data: 1) ordinal level data are erroneously treated as interval level measures; 2) items are assumed to be equally important; 3) error is normally distributed; 4) scores are sample-dependent; 5) parametric data analytic approaches require normally distributed data; and 6) missing data can present a validity threat. Because of these weaknesses, Rasch measurement models have become a popular tool for analyzing survey data as this family of measurement models overcome each of the aforementioned limitations of statistical approaches.

In short, Rasch models are latent trait, probabilistic models that create linear (logit scaled), sample free measures of persons and items. [9,10] In the case of the instrument used in this study, the person measure refers to a student's tendency to endorse each item, and the item measure refers to the difficulty of each item to be endorsed. Rasch models then place person and item measures onto a common scale where the two measures can be directly compared and their interactions explored. Proponents of Rasch measurement models also note the models' quality control features which allow thorough investigation of psychometric properties.

The Rasch Rating Scale Model (RRSM) [11] was utilized for analyzing course evaluation data. The RRSM is ideal for survey data analyses in which the rating scale categories are consistent across all items. According to the RRSM, the probability of a person n responding in category x to item i, is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta - (\delta_i + \tau_j)}, x = 0,1,...,m$$

where $\tau_o = 0$ so that, $\exp\sum_{j=0}^{0}[\beta_n - (\delta_i + \tau_j)] = 1$, β_n is the person's position on the variable, δ_i is the scale value (difficulty to endorse) estimated for each item *i* and τ_1, τ_2, . . ., τ_m are the *m* response thresholds estimated for the *m* + 1 rating categories.

Winsteps [12] measurement software was used to perform the analysis using joint maximum likelihood estimation (JMLE) procedures. [13]

3. Results

Descriptive statistics indicate students generally provided positive ratings of their courses (see Table 1). Although researchers typically prefer to see greater variability among scores and often perceive high course (and/or instructor) evaluation scores as problematic, it is important to note that such instruments are not intended to differentiate instructor talent or otherwise compare faculty. [14] Thus, concerns of variability are unwarranted. More specifically, in the same manner that students who perform excellently given a particular criterion should receive high marks, so should faculty who perform excellently given the criteria presented on this instrument. [15]

3.1. Evaluation of Psychometric Properties

3.1.1. Dimensionality

Dimensionality was inspected by conducting a Rasch-based principal components analysis (PCA) of standardized residual correlations. [16] A total of 55.4% of the variance was explained with 14.9% of the explained variance attributed to the items. No secondary dimension had an Eigenvalue greater than 2.0, the minimum threshold for multidimensionality.

3.1.2. Reliability

The traditional reliability (Cronbach's α) coefficient was .943. However, Rasch-based reliability estimates were .83 (real) and .86 (model), suggesting the true reliability estimate likely is somewhere in between. Separation statistics were also calculated to determine the number of strata. Separation statistics ranged from 2.17 (real) to 2.45 (model). Using the formula for calculating strata, [17] results indicated 3.23 (real) to 3.60 (model), or about 3 statistically distinguishable levels were discernible within the data.

3.1.3. Rating Scale Effectiveness

Rating scale effectiveness was evaluated by inspecting various rating scale statistics (see Table 2). Results indicate the rating scale categories advanced in proper structure calibration order, [18] indicating students had no difficulty discerning the directionality of the scale. Infit and outfit mean square fit statistics were within the recommended range of .60 to 1.40, [19] indicating the appropriateness of each category from a content perspective. The frequency distribution of rating scale category usage indicates students made full use of the rating scale, although responses were heavily skewed with mostly favorable ratings (87%).

3.1.4. Person and Item Measure Quality

Item measure quality was investigated by examining various indices, such as difficulty, discrimination and fit statistics (see Table 3). Item difficulty measures ranged from -.62 to .74 logits, with an average standard error of .06 (SD = .00). Item fit statistics were evaluated using the .60-1.40 criteria recommended by Wright and Linacre [19]. All infit and outfit mean square values ranged between .59 and 1.49. Overall data to model fit values were 0.99 (infit) and 1.00 (outfit). Point-measure correlations ranged from .81-.89.

Person measures ranged from -6.06 to 6.48 with standard errors averaging 1.21 (SD = .56) in magnitude. Overall data to model fit values were 1.01 (infit) and 1.09 (outfit). A total of 572 ratings were removed from the sample of 2,060 total ratings (about 27.8% of the response data) due to significant underfit.

One item pair exhibited a hint of local item dependence. [20,21] Item #5 and #6 had a standardized residual correlation of .46, which exceeds the recommended value of .30. [22] Although the magnitude of the relationship was small, it warranted a qualitative review to ensure students' response to one item likely would not impact their response to the other. The item pair clearly measured two different criteria, as item 5 inquired about assessments as a good measure of student learning and item 6 inquired about assessments providing an appropriate measure of course outcomes. Thus, after a review of content we conclude the small association between these item pairs likely does not threaten the validity of these measures.

Table 2. Rating scale diagnostics

Rating Category	n	%	INFIT MnSq	OUTFIT MnSq	Structure Calibration	Category Measure
1) Strongly Disagree	241	2	1.07	1.13	None	-3.98
2) Disagree	1,150	11	1.12	1.13	-2.82	-1.63
3) Agree	3,043	29	.98	.96	-.43	1.42
4) Strongly Agree	5,949	57	.88	.90	3.24	4.36

Table 3. Item quality indicators

Item	Difficulty Measure	Error	INFIT Mean Square	OUTFIT Mean Square	Point Measure Correlation
Q1	.74	.06	.83	.84	.89
Q2	.71	.06	1.44	1.49	81
Q3	.27	.06	.73	.70	.89
Q4	-.04	.07	1.28	1.33	.81
Q5	-.45	.07	.59	.57	.88
Q6	-.62	.07	1.00	.99	.82
Q7	-.62	.07	1.07	1.08	.81

3.1.5. Construct Hierarchy

The Wright Map (see Figure 1) provides a snapshot of the psychometric ruler and illustrates the construct hierarchy. Items (located on the right side of the map) at the bottom of the scale are easier for students to endorse, whereas items farther up the scale are more difficult to endorse. Conversely, persons (located on the left side of the map) at the bottom of the scale had a lesser tendency to endorse each of the items, whereas persons farther up the scale had a greater tendency to endorse each of the items.

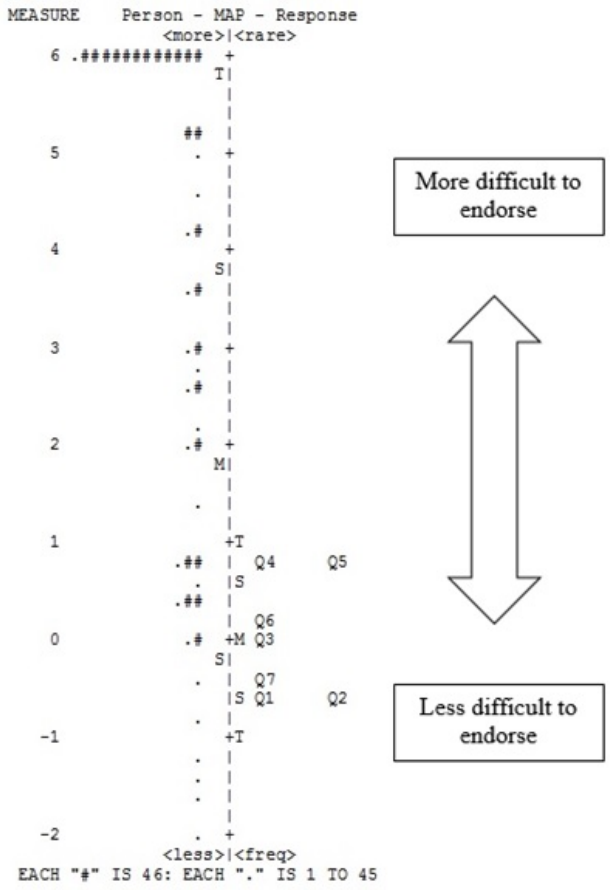

Figure 1. Construct Hierarchy

4. Discussion

4.1. Validity Framework

Many validity theorists have posited validity frameworks throughout the years, but a consensus has never been achieved. For the present work, we chose to use Messick's [23] unified conceptualization of validity which suggested fragmented 'types' of validity are not necessary as all validity evidence is construct validity.

4.2. Psychometric Properties

A Rasch-based principal components analysis (PCA) of standardized residual correlations indicated the data were sufficiently unidimensional, thus confirming the RRSM was an acceptable model for analyzing the data. Confirmation of unidimensionality addresses the substantive aspect of validity. Reliability measures ranged from .83-.86 (Rasch-based) to .943 (Cronbach's alpha), indicating scores are highly reproducible. [24] These values speak to the generalizability aspect of validity. Although the data were highly skewed with overwhelmingly positive ratings, rating scale diagnostics confirmed the appropriateness of the rating scale categories and the clarity for which the ordinal ratings are interpreted by student respondents. Further, fit statistics adequately fit the model's expectations. These findings speak to the structural aspect of validity. Person and item measures were deemed appropriate via an analysis of difficulty, discrimination and fit statistics. Further, a qualitative review of items flagged as potentially statistically dependent indicated the residual correlations exceeding .30 likely were simply a false-positive. These results speak to the content aspect of validity. Finally, as a matter of institutional policy we encourage appropriate and responsible use of course and instruction evaluation results to help ensure no unintended consequences result. This policy-driven initiative provides some evidence that speaks to the consequential aspect of validity. [25]

Despite the abundance of validity evidence, this validation effort also possessed a number of minor limitations. First, an investigation of differential item functioning (DIF) across relevant subpopulations (e.g., race, gender, etc.) was not performed due to a lack of ability to link these variables to ratings. DIF analyses offer an important element of validation, as these analyses can specify the degree to which item measures remain invariant across various subpopulations. Thus, we present no evidence to speak to the systematic aspect of validity. Further, given the unique purpose of course evaluations we have not correlated these results to any external data sources to examine score convergence/divergence. Thus, we present no evidence relating to the external aspect of validity. Finally, approximately 28% of the data were treated as missing because of significant underfit. These data are indicative of responses that were highly predictable given the model's expectations. For persons unfamiliar with Rasch measurement, one may find the discarding of data to be problematic. However, it is important to note that misfitting data routinely are discarded for Rasch measurement analyses as the focus is on fitting data to the model (as opposed to fitting a model to the data as traditional statistical approaches would dictate). [26,27] Collectively, however, adequate validity evidence is discernible to support the psychometric quality of the instrument.

4.3. Other Considerations

As noted previously, the seven items appearing on the course evaluation form are not exhaustive. In accordance with best practices in assessment, the instrument also includes several items (typically 3-5) that pertain to course-specific learning outcomes. These items are particularly valuable for informing instructors of the degree to which they achieved their course-specific educational mission. We encourage others interested in adopting this instrument to also include several items pertaining to course-specific learning outcomes.

One complaint we have received from students in previous years is that end-of-semester evaluations provide feedback that may help future students, but does not

necessarily affect students currently enrolled in the course. This point was well-taken, thus we adopted a practice of also offering mid-term evaluations. These evaluations are much shorter and serve as a "temperature check" in which students answer three questions relating to the course to this point. The questions ask, "To date, the course has been well-organized", "To date, the pace of the course has been appropriate", and "I have a clear understanding of how well I am doing in the course." The same aforementioned Likert-type rating scale is provided for these items. Students also are presented one open-ended item that reads: "Please provide specific, constructive comments about the course to this point. Please note aspects that were conducted well, and aspects that could be improved."

All course evaluation results are shared with course coordinators and the respective instructors from each course. Instructors are provided guidance for interpreting the results and asked to generate a list of improvements they intend to make to the course before it is offered again. Accountability for course improvements largely is a three-pronged approach that involves oversight by the Office of Academic Affairs, the college's curriculum committee, and department chairs.

Given many current students had completed the former instrument, we informally polled students to learn how the new instrument fared in comparison. Students were given time to complete the new instrument during mandatory lunchtime meetings at which refreshments were provided and the importance of their feedback to curricular improvement was emphasized. Afterward, students were very positive and appreciative of the efforts to streamline the instrument and offer it electronically. In fact, the response rate for the college increased from approximately 15-35% for a typical course to nearly 80% given the course evaluation instrument and process overhaul, though the relative effect of protected time and improvement in the instrument is uncertain. Thus, there is indirect evidence to support the continued use of the instrument from the students' perspective. We have yet to investigate course coordinators and instructors' perspectives about the new instrument and its utility, but will include this information as part of our comprehensive efforts to evaluate the instrument's quality and effectiveness.

4.4. Conclusions

The present study employed the Rasch Rating Scale Model to evaluate the psychometric properties of a newly developed instrument intended to measure the quality and effectiveness of courses in medical and health professions programs. Results indicate the instrument is psychometrically sound and capable of yielding valid and reliable measures. The authors encourage others to consider adopting the instrument for use at one's own campus, and/or to replicate the procedures presented within this text to evaluate the psychometric properties of other survey instruments.

References

[1] Smart, D.T., C.A. Kelley, and J.S. Conant. "Mastering the art of teaching: Pursuing excellence in the new millennium." *Journal of Marketing Education*, 25(1), 71-78, 2003.

[2] Kogan, J. R., & Shea, J. A. "Course evaluation in medical education." *Teaching and Teacher Education*, 23, 251-64, 2007.

[3] Billings-Gagliardi, S., Barrett, S. V., & Mazor, K. M. "Interpreting course evaluation results: insights from thinkaloud interviews with medical students." *Medical Education*, 38, 1061-70, 2004.

[4] Haidet, P., & Stein, H. F. "The role of the student-teacher relationship in the formation of physicians. The hidden curriculum as process." *Journal of General Internal Medicine*, 21 Suppl 1, S16-20, 2006.

[5] Royal, K. D., & Stockdale, M. R. "Are Teacher Course Evaluations Biased Against Faculty That Teach Quantitative Methods Courses?" *International Journal of Higher Education*, 4(1), 217-224, 2015.

[6] American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, 2014.

[7] Royal, K. D. "Four tenets of modern validity theory for medical education assessment and evaluation." *Advances in Medical Education and Practice*, 8, 567-570, 2017.

[8] Royal, K. D. "Making meaningful measurement in survey research: A demonstration of the utility of the Rasch model." *IR Applications*, 28: 1-16, 2010.

[9] Rasch, G. *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research, 1960.

[10] Bond, T.G., & Fox, C. M. *Applying the Rasch Model. Fundamental measurement in the human sciences* (3rd ed.). Routledge, New York, 2015.

[11] Andrich, D. "A rating formulation for ordered response categories." *Psychometrika*, 43, 561-573, 1978.

[12] Linacre, J. M. WINSTEPS® (Version 3.92.0). Computer Software. Beaverton, OR: Winsteps.com, 2017.

[13] Wright, B. D., & Masters, G. N. *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press, 1982.

[14] Royal K. D., & Guskey, T. R. "On the appropriateness of norm- and criterion-referenced assessments in medical education." *Ear, Nose & Throat Journal*, 94(7):252-4, 2015.

[15] Royal, K. D., & Guskey, T. R. "The perils of prescribed grade distributions: What every medical educator should know." *Journal of Contemporary Medical Education*, 2(4):240-1, 2014.

[16] Linacre, J. M. "Dimensionality: contrasts & variances." Available at: http://www.winsteps.com/winman/principalcomponents.htm, Accessed on July 6, 2017.

[17] Wright, B. D., & Masters, G. N. "Number of person or item strata: (4*Separation + 1)/3." *Rasch Measurement Transactions*, 16(3):888, 2002.

[18] Linacre, J. M. "Optimizing rating scale category effectiveness," *Journal of Applied Measurement*, 3(1):85-106, 2002.

[19] Wright, B. D., & Linacre, J. M. "Reasonable mean-square fit values." *Rasch Measurement Transactions*, 8:370, 1994.

[20] Royal, K. D. "The impact of item sequence order on local item dependence: An item response theory perspective." *Survey Practice*, 2016. Available at: http://surveypractice.org/index.php/SurveyPractice/article/view/344/html_78. Accessed on July 6, 2017.

[21] Marais, I., & Andrich, D. "Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model." *Journal of Applied Measurement*, 9(2): 105-124, 2008.

[22] Smith, R. M. "Fit analysis in latent trait measurement models." *Journal of Applied Measurement*, 1(2): 199-218, 2000.

[23] Messick, S. "Validity," In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan, 1989.

[24] Royal, K. D., & Hecker, K. G. "Understanding reliability: A review for veterinary educators." *Journal of Veterinary Medical Education*, 43: 1-4, 2016.

[25] Royal, K. D., & Puffer, J. C. "The consequential validity of ABFM examinations." *Journal of the American Board of Family Medicine*, 27(3): 430-431, 2014.

[26] Linacre, J. M. "When to stop removing items and persons in Rasch misfit analysis?" *Rasch Measurement Transactions*, 23(4):1241, 2010.

[27] Linacre, J. M. "Removing Rasch Misfit and Cleaning Windows." *Rasch Measurement Transactions*, 23(4):1241, 2010.