Evaluation of Score Quality From an Assessment of Student Competencies

Mary Roduta Roberts, Karin Werther, Cecilia B. Alves

Department of Occupational Therapy, Faculty of Rehabilitation Medicine, Edmonton, Alberta, Canada

INTRODUCTION

- Health science professions have identified and documented competencies for practice in their respective areas (e.g., Canadian Association of Occupational Therapists, 2012)
- Performance-based assessments are one method for assessing students' competencies within a simulated practice context in a relatively direct manner (Lane & Stone, 2006)
- Assessments, such as practical skills examinations, typically require judgments to be made on aspects of a student's performance; judgments often involve using rubrics to rate performance on demonstrations of specified competencies
- Scoring rubrics are defined as a scoring tool for qualitative ratings of authentic or complex student work
- Generalizability Theory (Brennan, 1992) is an analytical framework that can be used to investigate the extent to which scores measure the intended competency and vary due to other factors
- It is important to know the extent to which rater scores are reliable and reflect more about a student's competencies than they do lack of quality in the rubric or inconsistencies across raters; the quality of an assessment task has direct impact on the quality of the evidence generated, and consequently, the strength of inferences made about the student's proficiencies.
- Examination of a professional program's assessment practice can serve multiple purposes, from ensuring the evidence generated supports inferences and decisions made about students' competencies to supporting decision making processes related to administrations of the assessment.

OBJECTIVES

To evaluate score quality by examining the

1) reliability of scores from an assessment of student competencies, and
2) consistency across raters during the rating task

METHOD

DATA SAMPLE

- 99 Year 1 students in a course-based Masters program in occupational therapy
- Each student was assigned to a pair of raters who assessed the student independently
- There were 7 pairs of raters assessing between 11-17 students each

SCORING RUBRICS

Holistic	Analytic	Total Score
• A single score is assigned • Rater scores the overall process as a whole, without judging different categories or dimensions separately • Score range: 0-100	• Breaks the objective into different categories or dimensions • Each portion is scored independently • 6 dimensions with score ranges of 1-4	• Total score is made up of adding each category/ dimension scores • Score range: 4-24

The analytic rubric comprised six domains (4-point scale):

1) Professionalism
2) Communication
3) Theory, models, and frames of reference
4) Knowledge of client
5) Clinical reasoning
6) Evidence-based practice

DATA ANALYSIS

- Three Generalizability studies (G-studies) were conducted to estimate the variability among raters, and to assess the reliability of holistic, analytic, and total scores
- The number of raters was varied in the Decision study (D-study) to examine the effect on reliability

Holistic	Analytic	Total Score
• a one-facet G-study with students fully crossed with raters using overall holistic scores • (i.e., *p* x *r*)	• a two-facet G-study with students fully crossed with domains and raters • (i.e., *p* x *d* x *r*)	• a one-facet G-study with students fully crossed with raters using total scores derived from analytic scores • (i.e., *p* x *r*)

RESULTS

GENERALIZABILITY ANALYSES

- Variance components represent estimates of variability within the observed scores accounted by factors (i.e., facets) and by the object of measurement
- Variance components are often reported as a percentage of the variance accounted for by each component to the total variance within the measurement system
- For optimal measurement, the percentage of variance associated with students should be high relative to the percentage of variance attributable to the other facets
- The number of raters required to achieve a reliability of 0.80 was examined in the D-study for both relative and absolute decision making contexts

Analysis with holistic scores

G-study		D-study		
Students	76%	1 rater	0.80	G = 0.87
Raters	7%	2 raters	0.87	Φ = 0.83
		3 raters	0.90	* Relative decisions (G)

Analysis with analytic scores

G-study		D-study		
Students	39%	1 rater	0.67	G = 0.77
Domain	10%	2 raters	0.77	Φ = 0.72
Raters	1%	3 raters	0.82	* Relative decisions (G)

Analysis with total scores

G-study		D-study		
Students	62%	1 rater	0.64	G = 0.76
Raters	5%	2 raters	0.76	Φ = 0.74
		3 raters	0.82	* Relative decisions (G)

SUMMARY

- Raters performed consistently when assessing students producing highly reliable holistic scores (G = 0.87, Φ = 0.83), moderate-highly reliable competency-specific scores (G = 0.77, Φ = 0.72), and total scores (G = 0.76, Φ = 0.74)

CONCLUSION

- On average, rater pairs produced more reliable holistic scores in comparison to total and analytic scores; this information can be used to support decisions on which score to report
- Holistic scores were more reliable but may not be what is best for the context; analytic scores have the potential to provide more detailed feedback
- Across scoring scenarios, the largest percentage of variance was due to differences between students and the least percentage of variance was due to raters; this is a favourable outcome
- Likely other systematic sources of error (e.g., difficulty of the task and exam conditions) that have not been accounted for in the design
- The Generalizability Theory approach produces information useful for supporting existing assessment practices and decisions made about students
- Differences in reliability between analytic and holistic scoring provide insight into potential differences in assessor cognition which can be explored in future research
- Developing high quality assessments of student competencies is important when achievement of these competencies determine, in part, whether someone is fit to deliver safe and competent care

REFERENCES

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. http://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Canadian Association of Occupational Therapists. (2012). Profile of occupational therapy practice in Canada. Ottawa, ON: CAPT Publications ACE.

Lane, S., & Stone, C. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). Washington, DC: American Council on Education.

Author contact: mroberts@ualberta.ca

UNIVERSITY OF ALBERTA
FACULTY OF REHABILITATION MEDICINE
Department of Occupational Therapy