Assessing the Reliability of Performance Assessment Scores: Some Considerations in Selecting an Appropriate Framework

André F. De Champlain, PhD
Andrea Gotzmann, PhD
Sirius Qin, MS

The incorporation of learner assessments has become part and parcel of the accreditation process over the past few decades as a means of evaluating program or instructional effectiveness.[1] Given the high stakes associated with assessments not only for individual candidate-based decisions but also programs as a whole, it is critical to ensure that scores based on any tools meet certain psychometric standards. At its most elemental level, any test score is intended to reflect the competency domain(s) presumed to underlie an assessment. For example, if a candidate obtains a score of 90% on a direct observation tool, this might be interpreted as reflecting "strong" patient care, even though the latter is, in all likelihood, established on a small number of encounters. Given that high-stakes decisions may be based on such observational tools, it is critical that the sample of performance be reflective of the candidate's true ability in that competency. *Reliability* refers to the extent to which performance on any assessment (ie, in a restricted number of encounters) is indicative of the candidate's true competency level (ie, in an infinite number of encounters).[2] An "unreliable" assessment (ie, one that does not reflect the candidate's true competency level) could have dire consequences not only for the physician's medical education but also for the accreditation of the postgraduate program.

Due to restricted testing time, any assessment encompasses a limited sample of encounters that theoretically represents the domain of interest. The selection of 10 patients for inclusion into a direct observation assessment, for example, might be predicated on 3 hours of testing time. However, one could conceive of different sets of 10 patients that could have been selected. The program director who is reviewing a candidate's score of 90% with these 10 patients is not interested in restricting his or her interpretation of that "strong" performance to these 10 specific encounters, but rather *generalizes* this statement to the (theoretically infinite) pool of encounters from which the sample of 10 was selected.

Yet, several sources of measurement error can detract from the accuracy or precision with which the performance on a restricted sample of encounters generalizes to the broader domain. With performance assessments (in addition to the restricted sample of encounters), the examiners, the setting, and other factors can impede a candidate's score. *Reliability* allows us to estimate how well a score on any assessment (ie, a sample of performance) generalizes to the broader domain(s) of interest. With the previous example, how accurately does a score of 90%, in 10 patient encounters, scored by 10 examiners, generalize to all possible patient encounters and physician examiners? This generalization is quantified with a *reliability coefficient*.

Note that patients and examiners are sources of measurement error, given that any candidate's true score or ability level should not depend on the sample of patients nor the examiners encountered. A candidate's true ability level should be invariant across all these sources of measurement error or *facets*. In reality, all of these sources will detract from reliability due to the lack of representativeness of the patient encounters selected for an examination and the poor training of examiners.

Commonly, Cronbach's α coefficient is computed as the reliability estimate largely because it is readily available in most statistical software packages.[3] However, the use of Cronbach's α with examinations that are affected by several sources of measurement error, such as performance-based assessments, is ill-advised. Specifically, this coefficient does not partition all sources of measurement error in the computation of the reliability coefficient; rather, it is restricted to only 1 facet (ie, "patient encounters") in the previous example. Cronbach's α can thus yield a very misleading (spurious) reliability estimate because of its inability to incorporate (and partition out) all sources of measurement error.

DOI: http://dx.doi.org/10.4300/JGME-D-15-00751.1

Generalizability Theory (G Theory) is a reliability framework that allows us to properly quantify the impact of these error sources in regard to the extent to which we can generalize performance in a restricted sample of conditions (patient encounters, examiners) to broader domains. G Theory is an extension of Cronbach's α that allows the user to prespecify and estimate the impact of all potential sources of measurement error.[4]

G Theory uses analysis of variance modeling to estimate the amount of variability in scores due to sources of measurement error as well as their impact on the reliability coefficient, referred to as a *generalizability coefficient* (G coefficient). To use G Theory most efficiently and in a helpful manner, careful consideration must be given to all aspects of examination development (eg, the number of raters, how they are to be assigned to candidates, the number of stations, etc). For example, to estimate how much variance is due to raters, the raters need to score some common elements of the assessment (ie, either common patients or candidates).

To illustrate the application of G Theory, imagine a 9-patient encounter assessment that is completed by 90 candidates as a requirement in a given postgraduate program. The assessment targets the "patient care" Accreditation Council for Graduate Medical Education competency. Three examiners are assigned to rate different candidates: (1) examiner 1 rates candidates 1 to 30; (2) examiner 2 rates candidates 31 to 60; and (3) examiner 3 rates candidates 61 to 90. In G Theory parlance, this is a $p{:}r \times pe$ design, where p, r, and pe respectively correspond to persons (candidates), raters (examiners), and patient encounters. Persons are *nested* within raters ($p{:}r$), since not all candidates are rated by the same examiner. Furthermore, persons nested within raters are *crossed* with patient encounter ($p{:}r \times pe$) because it is assumed in this example that all candidates encounter the same 9 patients in their assessment.

The Cronbach's α value for this dataset was 0.86, which users may infer as "highly reliable" (ie, scores generalize well to domains targeted by the examination and allow us to accurately rank order candidates from low to high). However, the reliability estimate is spuriously inflated, as supported by an analysis using G Theory conducted on the same dataset using a $p{:}r \times pe$ design (TABLE).

The variance component associated with $p{:}r$ is akin to true score variance, as it provides an estimate of the amount of score variability due to true differences in ability among candidates. Specifically, 7% of total score variance is due to true difference in ability among candidates, suggesting some modest spread and consequently some capability to differentiate candidates (rank order). The variance component due to *patient encounter* reflects difficult differences. The small percentage of variance accounted for by this source suggests that encounters were highly comparable in terms of difficulty. The $r \times pe$ variance component, which is virtually nil, indicates that the stringency level of the examiners did not differ as a function of the patient encounter. Finally, the $p{:}r \times pe$, e component is a residual term, which reflects the amount of error in generalizing due to all other sources not specified in the design.

TABLE
Nine Patient Encounter Assessments: $p{:}r \times s$ Generalizability Analysis Results

Source of Variation	Variance Component	% of Total Variance
Persons:raters ($p{:}r$)	0.056	7.2[a]
Stations (s)	0.025	3.2[b]
Raters (r)	0.309	39.5[c]
$r \times s$	0.006	0.8
$p{:}r \times s$, e	0.386	49.3

[a] This value, though modest, does suggest that there is some variability among candidates' true level of patient care, as would be expected. This modest amount of variance among candidates is perhaps unsurprising given the relative homogeneity of residents with regard to this and other competencies.

[b] The careful selection and balancing of patient encounters, based on a clearly defined blueprint, should allow us to assemble assessments that are relatively comparable, in terms of their overall difficulty level. The small amount of variance due to patient encounters suggests that the assessments provided to candidates are equitable with respect to difficulty level, which is a key fairness consideration.

[c] The large amount of variance due to raters would not be totally unexpected in a situation where physician examiners receive little to no training. In that instance, physicians are more likely to inject personal biases in the rating task that are unrelated to the competencies targeted by the assessment.

Of particular interest in this example is the large amount of variance due to raters (39.5%). Nearly 40% of the variance in the assessment scores is due to differences in stringency between the 3 raters. Note that this effect is completely independent from the abilities of the 90 candidates. A G coefficient of 0.57 was computed—a significantly lower value than Cronbach's α.

What might account for the large difference in reliability estimates obtained with the assessment scores? In calculating Cronbach's α, the large differences in candidate scores due to the high variability among examiners (a source of measurement error) gets "confounded," as true score variance which artificially inflates the reliability coefficient. The highly divergent examiners are "injecting" a high level of score variance due to their own variability as examiners rather than being reflective of differences in candidate ability levels. Since there is no mechanism in the calculation of Cronbach's α to account for

examiner variability, this gets incorrectly partitioned as true score variance or true differences among candidate abilities. In G Theory, the error variance due to examiners is correctly partitioned out of true score variance and treated as a source of measurement error, which appropriately lowers the G coefficient value.

This example illustrates the pitfalls that can result from the sole use of Cronbach's α coefficient in estimating the reliability of scores with highly complex assessments, such as those commonly used in postgraduate medical education. It is important to point out, however, that Cronbach's α is appropriate in instances where a single rater is involved in the assessment. Also, for those assessments, such as simulations, which may involve clear scoring keys with little to no rater input, reliability can be confidently estimated with Cronbach's α given that there is only 1 source of measurement error (scenario).

However, in the example used to illustrate the concept, the high Cronbach's α value (0.86) could lead the medical educator to commit erroneous high-stakes decisions (promotion, graduation, etc), given the limitations of the reliability coefficient and its inability to properly account for the large amount of variability due to examiners. For any assessment that involves several facets (multi-source feedback, direct observation–based rating scales, etc), it is highly recommended that the practitioner complete a generalizability analysis not only to properly estimate reliability, but also to garner information that might be beneficial in improving the assessment for future uses.

References

1. Accreditation Council for Graduate Medical Education, American Board of Medical Specialties. Toolbox of assessment methods. 2000. http://njms.rutgers.edu/culweb/medical/documents/ToolboxofAssessmentMethods.pdf. Accessed April 20, 2016.
2. Traub RE, Rowley GL. Understanding reliability. *Educ Meas Issues Pract*. 1991;10(1):37–45.
3. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297–334. http://kttm.hoasen.edu.vn/sites/default/files/2011/12/22/cronbach_1951_coefficient_alpha.pdf. Accessed April 20, 2016.
4. Brennan RL. *Generalizability Theory*. New York, NY: Springer-Verlag; 2001.

All authors are with Psychometrics and Assessment Services, Medical Council of Canada, Ottawa, Ontario. **André F. De Champlain, PhD,** is Director; **Andrea Gotzmann, PhD,** is Senior Research Psychometrician; and **Sirius Qin, MS,** is Statistical Analyst.

Corresponding author: André F. De Champlain, PhD, Medical Council of Canada, Psychometrics and Assessment Services, 2283 Saint Laurent Boulevard, Suite 100, Ottawa, Ontario K1G 5A2 Canada, 613.521.6012, ext 2541, adechamplain@mcc.ca