Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 34

A Multivariate Generalizability Analysis of Portfolio Assessments in Dental Education

Robert L. Brennan[1]

April 15, 2013

[1]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA). The author thanks Cyndie Amyot of the University of Missouri–Kansas City for for providing the data analyzed in this report, and for the opportunity to work with her on portfolio assessment in dentistry.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

Abstract

This report provides results for a multivariate generalizability theory study of portfolio assessment data collected in 2012 at two dental schools. The primary purpose of this report is to document the study results and how they were obtained, rather than describe in detail the portfolio assessments and their use. Many discussions of details about the analyses reported here assume the reader is familiar with at least the basic features of multivariate generalizability theory as discussed by Brennan (2001a, especially chaps. 9–10). A noteworthy result is that operational use of the portfolio assessments would benefit greatly from two raters evaluating each portfolio independently; one rater seems to be only marginally adequate.

This report provides results for a multivariate generalizability theory study of portfolio assessment data collected in 2012 at two dental schools. Details about the portfolio assessments themselves and their use in dental education are provided in other publications that can be obtained from Cindy Amyot (AmyotC@umkc.edu), Professor and Associate Dean, Instructional Technology and Faculty Development, University of Missouri–Kansas City, School of Dentistry, 650 E. 25th Street, Kansas City, MO 64108 (Tel: 816-235-2054). The primary purpose of this report is to document the study results and how they were obtained. It is assumed here that the reader is familiar with at least the basic features of multivariate generalizability theory as discussed by Brennan (2001a, especially chaps. 9–10).

Complete data were available for 45 students from School 1 and 22 students from School 2. The instruments used by the two schools had the following five Portfolio Primary Traits (PPTs):

- PPT1: Critical Thinking,
- PPT2: Professionalism and Ethics,
- PPT3: Communication and Interpersonal Skills,
- PPT4: Health Promotion and Service, and
- PPT5: Patient Care.

The components (or criteria) for each PPT at the two schools differed somewhat, with the common components numbering 3, 3, 2, 3, and 2 for the five PPTs, respectively. This report uses data for the common components, only. At each school, the portfolios for all students (or persons) were evaluated by the same four raters, with each rater evaluating each component of each PPT using a four-point scale (1, 2, 3, and 4). The same types of analyses were run at both schools. Since results for the two schools were quite similar, detailed results are provided here for school 1, only.

Virtually all results reported here were obtained using the computer program mGENOVA (Brennan, 2001b). The Appendix provides mGENOVA control cards that generated most of the School 1 results. The data files for both School 1 and School 2 are available under the "Data Files" tab on the CASMA website (http://www.education.uiowa.edu/centers/casma/).

G Study Results

Table 1 provides G study estimated variance and covariance components for School 1 for the $p^\bullet \times c^\circ \times r^\bullet$ design, where p stands for persons, c stands for components, and r stands for raters.[2] This design can be characterized in the following manner.

[2]In Table 1 and elsewhere in this report, the notation does not distinguish between parameters and estimates.

Table 1: G Study Variance and Covariance Components for School 1

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} .15898 & \textit{.77126} & \textit{.77398} & \textit{.82221} & \textit{.76373} \\ .11222 & .13318 & \textit{.91812} & \textit{.83091} & \textit{.77555} \\ .09133 & .09916 & .08758 & \textit{1.01099} & \textit{1.08165} \\ .11465 & .10605 & .10464 & .12231 & \textit{.87021} \\ .11779 & .10948 & .12382 & .11772 & .14962 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} .00146 & & & & \\ & .05852 & & & \\ & & .05795 & & \\ & & & -.00370 & \\ & & & & .00240 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_r = \begin{bmatrix} .01651 & & & & \\ -.01800 & .00400 & & \text{symmetric} & \\ -.00253 & .01432 & -.00442 & & \\ .00733 & -.00871 & .00804 & .01166 & \\ .05039 & -.03223 & .00862 & .02998 & .10715 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{pc} = \begin{bmatrix} .00410 & & & & \\ & .00630 & & & \\ & & .04019 & & \\ & & & .00741 & \\ & & & & .00130 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{pr} = \begin{bmatrix} .20015 & & & & \\ .09969 & .22440 & & \text{symmetric} & \\ .11271 & .12950 & .15812 & & \\ .06428 & .09822 & .10153 & .18464 & \\ .05610 & .07020 & .07148 & .07311 & .17803 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{cr} = \begin{bmatrix} .00818 & & & & \\ & .06792 & & & \\ & & .03611 & & \\ & & & .01014 & \\ & & & & .01338 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{pcr} = \begin{bmatrix} .13256 & & & & \\ & .16542 & & & \\ & & .25185 & & \\ & & & .18801 & \\ & & & & .07180 \end{bmatrix}$$

Note. Italicized upper diagonal elements of $\boldsymbol{\Sigma}_p$ are disattenuated correlations. Sample sizes are: $n_p = 45$ students, $n_r = 4$ raters, and n_c =3, 3, 2, 3, 2.

- There are seven matrices corresponding to the seven effects (main effects and interactions) in the design.
- Each matrix is of size 5×5 with the five rows and columns corresponding to the five PPTs.
- $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_r$, and $\mathbf{\Sigma}_{pr}$ are full, symmetric matrices. They are full because each person and each rater provides data associated with each PPT.
- The other four matrices involving the components, c, are diagonal because the components are different for each PPT.
- Persons (or students) are the objects of measurement. Components and raters are random facets, in the sense that the actual components and raters are assumed to be a sample of the possible components and raters that could be used. The five PPTs, however, are assumed to be fixed, in the sense that the same PPTs would be used if the study were replicated.

As noted above, $\mathbf{\Sigma}_p$ is a symmetric matrix, which means that the lower and upper off-diagonal elements are the same. In Table 1, however, the upper off-diagonal elements of $\mathbf{\Sigma}_p$ have been replaced with disattenuated correlations solely for the purpose of reporting these statistics. Roughly speaking, if the disattenuated correlation for, say, PPTa and PPTb is close to one, then the true (or universe) scores for PPTa and PPTb are highly related. Technically, disattenuated correlations cannot be greater than one, but estimates of them can be quite unstable leading to values that are indeed greater than one, on occasion, as is the case in Table 1 for PPT3 and PPT4, as well as for PPT3 and PPT5. The other disattenuated correlations, however, are notably less than one, suggesting that the different PPTs are measuring different constructs, for the most part.

D Study Results

The variance and covariance components in Table 1 are the building blocks used to estimate results for making decisions about students' reported scores. The two principal types of statistics used for this purpose are coefficients (e.g., generalizability coefficients) and standard errors of measurement (SEMs). The studies conducted to estimate these statistics are called D (Decision) studies.

Coefficients for School 1

In the G study, $n_r = 4$ and $n_c = 3,\ 3,\ 2,\ 3,\ 2$ for each of the five PPTs, respectively. When the same sample sizes are used for the D (Decision) study, for composite scores the estimated generalizability coefficient (sometimes called a reliability coefficient) is $\boldsymbol{E}\rho^2 = .78$, assuming the individual PPT scores are weighted proportionately to the number of components in each of them. This is the weighting that occurs "naturally" if the composite for each person is

obtained as a simple sum of the four ratings associated with the 13 component scores. (We return to this matter later.)

In general, we can estimate a generalizability coefficient for any number of raters and components. (Note that D study sample sizes are denoted n' to differentiate them from G study sample sizes denoted n.) When portfolio assessments are used operationally in dental education, it is highly unlikely that as many as four raters will be used to evaluate all components of every student's portfolio. Also, different schools will probably use different numbers of components, with the total number or components likely being in the range of 15–20. Given these speculative comments, the following table provides estimated generalizability coefficients for $n'_c = 3$ and $n'_c = 4$ for each PPT, as well as for $n'_r = 1$, 2, 3, and 4.

$\boldsymbol{E}\rho^2$	$n'_r = 1$	$n'_r = 2$	$n'_r = 3$	$n'_r = 4$
$n'_c = 3$	.49	.65	.74	.79
$n'_c = 4$	.49	.66	.74	.79

Obviously, $\boldsymbol{E}\rho^2$ does not not differ much for different values of n'_c, which is to be expected given that $\boldsymbol{E}\rho^2$ involves $(1/n'_c)\boldsymbol{\Sigma}_{pc}$, which has values that are all very small (see Table 1).[3]

The $\boldsymbol{E}\rho^2$ results are for the D study $p^\bullet \times C^\circ \times R^\bullet$ design, which means that the *same* n'_r raters evaluate all components for all students. This means, for example, that when $n'_r = 1$, there is a single rater who evaluates all components for all students. Clearly, in operational environments, this is not a very likely scenario. Much more likely scenarios involve a group of raters each of whom evaluates: (a) all the components for a subset of the students; (b) a subset of the components for all of the students; or (c) a subset of the components for a subset of the students. Under certain assumptions, it is theoretically possible to estimate $\boldsymbol{E}\rho^2$ for each of these scenarios, but it is not very practical (and probably too speculative) to do so with the currently available data.

What can be done, however, is to estimate a lower bound for $\boldsymbol{E}\rho^2$ for such scenarios. One such lower bound is the case in which there is a single rater for each student. Formally, this is $\boldsymbol{E}\rho^2$ for the D study $(R^\bullet{:}p^\bullet) \times C^\circ$ design with $n'_r = 1$. Given the School 1 data, it can be shown that for this scenario, $\boldsymbol{E}\rho^2 = .46$, which is only slightly lower than the previously reported value of $\boldsymbol{E}\rho^2 = .49$ for the $p^\bullet \times C^\circ \times R^\bullet$ with $n'_r = 1$. Neither of these values is very impressive, however. In performance assessment contexts, that is usually the price that is paid for having only one rater evaluate products produced by a student. As indicated in the above table, there is a substantial increase in reliability going from one to two raters.

In generalizability theory there is another reliability-like coefficient, called Φ. The difference between $\boldsymbol{E}\rho^2$ and Φ is that the former involves relative error variance, whereas the latter involves absolute error variance. For the School 1

[3]Corresponding results could be obtained using mGENOVA for unbalanced situations in which each PPT had either $n'_c = 3$ or $n'_c = 4$. Such results would differ very little from the results reported in the table.

data there is very little difference between estimates of these two error variances and, therefore, very little difference between $\boldsymbol{E}\rho^2$ and $\widehat{\Phi}$.

SEMs for School 1

The principal virtue of a reliability-like statistic such as $\boldsymbol{E}\rho^2$ is that it is scale independent,[4] with values always ranging from 0 to 1. A standard error of measurement (SEM) is another statistic for quantifying the dependability of scores.[5] It is scale dependent, and in generalizability theory an SEM is usually reported in terms of the mean-score metric, although it can be obtained for any metric. Metric matters and their consequences for SEMs can be complicated; they are considered in more detail next under relatively simple assumptions.

Consider the case of $n'_c = 3$ for all five PPTs with $n'_r = 1$, and suppose some student had ratings for the five PPTs as indicated in the second row, below.

Score	v_1 = PPT1	v_2 = PPT2	v_3 = PPT3	v_4 = PPT4	v_5 = PPT5
X_{prcv}	(3,3,3)	(2,2,2)	(2,3,4)	(3,2,4)	(4,4,4)
$\overline{X}_{pv}$	3	2	3	3	4

Note that X_{prcv} means the score assigned to the person by rater r for component c in PPTv. The sum of all of these scores is 45, which is the composite total score. The last row provides the person's mean scores ($\overline{X}_{pv}$) for each of the PPTs. Note that these mean scores are necessarily scores in the 1–4 rating rubric range. The mean of these mean scores is $15/5 = 3$, which is called the composite mean score; it, too, must be in the 1–4 rating rubric range.

Alternatively, the composite mean score can be expressed as

$$\left(\frac{1}{5}\right)\left(\frac{9}{3}\right) + \left(\frac{1}{5}\right)\left(\frac{6}{3}\right) + \left(\frac{1}{5}\right)\left(\frac{9}{3}\right) + \left(\frac{1}{5}\right)\left(\frac{9}{3}\right) + \left(\frac{1}{5}\right)\left(\frac{12}{3}\right) = 3.$$

This is a specific application of a general formula for obtaining a composite mean score for a person:

$$\overline{X}_p = \sum_{v=1}^{n_v} w_v \overline{X}_{pv}, \tag{1}$$

where w_v is the user-specified weight for PPTv, the sum of the w_v is 1, and $\overline{X}_{pv}$ is the mean score for person p for PPTv. For the simple example introduced above, all the w_v are $1/5 = .2$, and the $\overline{X}_{pv}$ are the mean scores 3, 2, 3, 3, and 4. The fact that the weights are all equal implies that scores for each of the PPTs are equally important, as far as the investigator is concerned.[6]

[4] Strictly speaking, values of $\boldsymbol{E}\rho^2$ are invariant with respect to *linear* transformations of scores, not non-linear transformations.

[5] In generalizability theory, there are two types of SEMs: absolute-error SEM denoted $\sigma(\Delta)$, and relative-error SEM denoted $\sigma(\delta)$. For this study, the two types of SEMs are very similar and, therefore, the relative-absolute distinction is ignored. Strictly speaking, the reported values are absolute-error SEMs.

[6] Generalizability theory does not require equal weights; the choice of weights is entirely up to the investigator. Metric matters become considerably more complicated, however, when weights differ and/or the number of components is not a constant for all PPTs.

SEMs for composite scores follow much the same logic outlined above. In particular, SEMs for the mean-score metric are different from SEMs for the total-score metric, which is to say that SEMs are metric dependent. For the portfolio assessment situation considered in this report,

$$\text{SEM}(+) = n'_r n'_c n_v \, \text{SEM}(\cdot), \tag{2}$$

where SEM(+) is the SEM for the total-score metric, and SEM($\cdot$) is the SEM for the mean-score metric.

In generalizability theory, it is traditional to report SEMs for the composite in the mean-score metric, which are the SEMs typically reported by mGENOVA. For the D study $p^\bullet \times C^\circ \times R^\bullet$ design with $n'_c = 3$, these results are provided in the second row, below. Using Equation 2, the corresponding results for the total-score metric are provided in the third row.

	$n'_r = 1$	$n'_r = 2$	$n'_r = 3$	$n'_r = 4$
SEM($\cdot$)	.36	.26	.21	.18
SEM(+)	5.47	7.80	9.64	11.22

As illustrated in the above table, for the mean-score metric SEMs *decrease* as sample sizes increase. This makes intuitive sense in that, no matter what the sample sizes are, the range of scores for the mean-score metric is 1–4, and increasing sample sizes *decreases* our uncertainly about results. By contrast, for the total-score metric SEMs *increase* as sample sizes increase. This can be understood, in part, by noting that for the total-score metric: (a) the range of possible scores increases substantially as sample sizes increase (e.g., in the case, with $n'_c = 3$ the ranges are 15–60, 30–120, 45–180, and 60–240 for $n'_r = 1$, 2, 3, and 4, respectively); and (b) SEMs increase as sample sizes increase, but at a slower rate than the range increases.[7]

As noted in the previous subsection, for the D study $p^\bullet \times C^\circ \times R^\bullet$ design, the same raters evaluate all components for all students, which is probably not very likely in an operational context. There is a large number of possible scenarios, and it is not practical to estimate SEMs for all of them. What can be done, however, is to estimate a lower bound for the SEM for such scenarios. One such lower bound is the case in which there is a single rater for each student. Formally, this is the SEM for the D study $(R^\bullet{:}p^\bullet) \times C^\circ$ design with $n'_r = 1$. Given the School 1 data, it can be shown that for this scenario, SEM = .36 for the mean-score metric and SEM = 5.47 for the total-score metric. Note that these are the same values reported above for the $p^\bullet \times C^\circ \times R^\bullet$ design with $n'_r = 1$. This equivalence of results (to two decimal places), is solely a consequence of the facts that $n'_r = 1$ and $(1/n'_c)\boldsymbol{\Sigma}_c$ is a diagonal matrix with very small values.

[7] A more technically correct explanation is that, for the total-score metric, an increase in sample size increases observed score variance and true score variance more that it increases error variance.

The typical use of an SEM is to construct confidence intervals for students' true (or universe) scores. Under typical assumptions, we can say that for students with a composite mean score of $\overline{X}_p$, 68% of the time the interval

$$(\overline{X}_p - \text{SEM}(\cdot),\ \overline{X}_p + \text{SEM}(\cdot))$$

will contain their true scores. Similarly, for students with a composite total score of X_p, 68% of the time the interval

$$(X_p - \text{SEM}(+),\ X_p + \text{SEM}(+))$$

will contain their true scores.

Suppose, for example, that $n'_c = 3$ and $n'_r = 1$, and consider students in School 1 with a composite mean score of $\overline{X}_p = 3$. For them, a 68% confidence interval in the mean-score metric is

$$(3 - .36,\ 3 + .36) = (2.64,\ 3.36).$$

These students have a composite total score of $X_p = n'_r n'_c n_v 3 = 15(3) = 45$, which implies that a 68% confidence interval for them in the total-score metric is

$$(45 - 5.47,\ 45 + 5.47) = (39.53,\ 50.47).$$

Suppose it was decided that a "true" rating of 3 in the mean-score metric constituted "proficiency" in some sense. The 68% confidence interval of (2.64, 3.36) for examinees with $\overline{X}_p = 3$ might be sufficiently narrow for a decision-maker to assert that such examinees are proficient. On the other hand, it might be noted that a 95% confidence interval would be

$$(3 - 1.96 \times .36,\ 3 + 1.96 \times .36) = (2.29,\ 3.71),$$

and this interval might be too broad for the decision-maker to be comfortable with characterizing students as proficient who have $\overline{X}_p = 3$. This is a rather typical, albeit rather ad hoc, way to use SEMs in making decisions; there are more precise ways to do so. The important point for purposes here is that SEMs provide decision makers with information about uncertainly in a metric that should be familiar to them. Given the School 1 data, it is clear that the most productive route for decreasing SEMs is to increase the number of raters per student.

Summary Results for School 2

For School 2 with $n'_c = 3$, the following table provides estimates of generalizability coefficients and SEMs.

	$n'_r = 1$	$n'_r = 2$	$n'_r = 3$	$n'_r = 4$
$\boldsymbol{E}\rho^2$	.54	.70	.78	.83
SEM($\cdot$)	.35	.25	.20	.17
SEM(+)	5.19	7.35	9.02	10.44

Results for $n'_c = 4$ were very similar. These results for School 2 are quite similar to those for School 1, although School 1 has slightly higher values for $\boldsymbol{E}\rho^2$ and slightly lower values for SEMs.

Comments

It is noteworthy how little variance was attributable to components within PPTs. This suggests that perhaps holistic ratings for PPTs might work about as well as rating all components within PPTs (what might be called analytic ratings). It could be that raters truly believed there was not much basis for differentiating the components within the various PPTs. Alternatively, it could be that raters adopted a kind of "halo" effect for each PPT, perhaps because they were not trained as well as they might have been. These hypotheses would seem to be fertile areas for further research.

In any case, given the data analyzed here, there appears to be ample reason to strongly suggest that operational use of the portfolio assessments would benefit greatly from two raters evaluating each portfolio independently. More than two raters is probably not necessary, but one rater seems to be only marginally adequate.

References

Brennan, R. L. (2001a) *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2001b). *mGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa. [Computer software and manual.] Available on www.education.uiowa.edu/casma

Appendix: mGENOVA Control Cards

G and D Study Control Cards for School 1

```
GSTUDY    School-1 p(bullet) x c(circle) x r(bullet)
OPTIONS   NREC 5 "*.out"
MULT      5  PPT1  PPT2 PPT3  PPT4  PPT5
EFFECT    * p  45    45   45    45    45
EFFECT      c   3     3    2     3     2
EFFECT    # r   4     4    4     4     4
FORMAT    10 0
PROCESS   "School-1-data.txt"
DSTUDY    School-1 p(bullet) x C(circle) x R(bullet); original nc; prop wts
DOPTIONS  NEGATIVE
WWTS      .23077  .23077  .15385  .23077  .15385
DEFFECT   $ p  45    45   45    45    45
DEFFECT     c   3     3    2     3     2
DEFFECT   # R   4     4    4     4     4
ENDDSTUDY
DSTUDY    School-1 p(bullet) x C(circle) x R(bullet); nc = 3; nr = 4; = wts
DOPTIONS  NEGATIVE
WWTS      .2  .2  .2  .2  .2
DEFFECT   $ p  45    45   45    45    45
DEFFECT     c   3     3    3     3     3
DEFFECT   # R   4     4    4     4     4
ENDDSTUDY
DSTUDY    School-1 p(bullet) x C(circle) x R(bullet); nc = 3; nr = 3; =l wts
DOPTIONS  NEGATIVE
WWTS      .2  .2  .2  .2  .2
DEFFECT   $ p  45    45   45    45    45
DEFFECT     c   3     3    3     3     3
DEFFECT   # R   3     3    3     3     3
ENDDSTUDY
DSTUDY    School-1 p(bullet) x C(circle) x R(bullet); nc = 3; nr = 2; = wts
DOPTIONS  NEGATIVE
WWTS      .2  .2  .2  .2  .2
DEFFECT   $ p  45    45   45    45    45
DEFFECT     c   3     3    3     3     3
DEFFECT   # R   2     2    2     2     2
ENDDSTUDY
DSTUDY    School-1 p(bullet) x C(circle) x R(bullet); nc = 3; nr = 1; = wts
DOPTIONS  NEGATIVE
WWTS      .2  .2  .2  .2  .2
DEFFECT   $ p  45    45   45    45    45
DEFFECT     c   3     3    3     3     3
DEFFECT   # R   1     1    1     1     1
ENDDSTUDY
```