

ASSESSING THE RELIABILITY OF STUDENT EVALUATIONS OF TEACHING (SETS) WITH GENERALIZABILITY THEORY

Dawn M. VanLeeuwen, Associate Professor

Thomas J. Dormody, Professor

Brenda S. Seevers, Associate Professor

New Mexico State University

Abstract

This paper presents a Generalizability Theory (GT) analysis of the dependability of measures obtained using the Seevers, Dormody and VanLeeuwen (1998) SET instrument. The analysis of Seevers et al. (1998) data suggests that both student means and class means are sufficiently reliable measures for most purposes. These data are used to illustrate the GT estimation of reliabilities for both student means, involving averaging over items, and class means, involving averaging over both items and students. The use of GT to obtain standard errors for use in interpreting differences in class means is also illustrated and recommendations on the interpretation of class means are given.

Introduction

Most universities use student evaluation of teaching (SET) instruments to provide information to instructors for course and instructor improvement. Increasingly, SET data also has been used by administrators and tenure and promotion committees as one piece of information used to document instructor performance. Within this context, SETs affect both annual teaching performance and salary decisions as well as promotion and tenure decisions. Clearly, SETs can have an impact on both the manner in which courses are taught in the future as well as the recognition of instructor competence. As such, it is important that SETs be valid and reliable measures of quality teaching and course development. It also is important that those interpreting SET data have some understanding of the variability inherent in SET scores and what that variability suggests about the dependability of SET data.

Seevers et al. (1998) noted that some items on a SET instrument in use at their home institution had no basis in the research on exemplary teaching. Additionally, several

dimensions of exemplary teaching suggested by the research were not represented on the home institution's SET instrument. They undertook to develop a valid and reliable SET for their college that was rooted in a strong theoretical base. The papers of Rosenshine and Furst (1971) and Feldman (1988, 1989) synthesized the information found in many studies of the dimensions of exemplary teaching and provide the theoretical base that Seevers et al. (1998) used in developing their SET instrument. This paper presents detailed information about the dependability of measurements provided by the Seevers et al. (1998) SET instrument and uses their data to illustrate the application of Generalizability Theory (GT) to correctly assess the reliability of class means obtained from the SET instrument.

SET data often are summarized using class means obtained by averaging over both items and students. However, as O'Brien (1990) stated: "Although aggregate-level variables based on the weighted sum of individual-level characteristics are commonly used in sociology, the reliability of such measures is almost never assessed." (p. 497). This is certainly true in agricultural education where Cronbach's alpha typically would be reported for

data obtained from a teaching evaluation instrument. Cronbach's alpha is a measure of inter-item consistency and provides a reasonable estimate of reliability if student perceptions are ranked on the basis of either the sum or average of the items. However, since SET data are often averaged over both items and students within the class, inter-item consistency is just one facet affecting the reliability of instructor assessments. The consistency of student perceptions is another. Student perceptions are individual-level variables, while scores obtained for purposes of evaluating teaching are aggregate-level variables.

Generalizability Theory can be used to assess the reliability of aggregate-level measures, such as those based on SETs (O'Brien, 1990; Brennan, 1975). GT provides a unified approach to understanding the dependability of measures (Brennan, 1983; Shavelson & Webb, 1991; VanLeeuwen, Barnes, & Pase 1998; VanLeeuwen, 1997) and allows accurate assessment of the reliability of complex measures as well as measures used for either relative decisions or criterion-referenced decisions. GT can be viewed simply as an extension of classical reliability theory, or it can provide a framework for thinking about the dependability of measurements in a much broader sense (Kane, 1993; VanLeeuwen et al., 1998).

Purpose

This paper has two primary purposes. The first purpose is to present a complete GT analysis of the dependability of SET data obtained using the Seevers et al. (1998) instrument. The second is to use these data to illustrate the correct application of GT to estimate the reliability of SET class means. The versatility and utility of GT is illustrated by computing reliability coefficients for both the individual-level variable (student means taken across items only) as well as the aggregate-level variable (class means taken across both items and students). In addition, the impact of alternative scoring schemes on reliability

coefficients is discussed. The use of GT to address the same issues as classical reliability theory is demonstrated. The greater generality and flexibility of GT is illustrated, as is the understanding of the dependability of measures that GT's focus on variance components allows. Correct interpretation of the variance components and reliability assessments is emphasized in the final two sections.

The Data

Details about the dimensions of exemplary teaching used for purposes of developing this instrument are given in Seevers et al. (1998). The final instrument consisted of 27 items. Data were collected during the fall semester 1997 in the College of Agriculture and Home Economics at New Mexico State University. Thirty classes were selected using a purposive sample that represented a broad spectrum of courses in the college. All departments and course levels (lower, upper, and graduate) were represented. There was no duplication of instructor or course so that instructor and course were completely confounded. 531 students in 26 classes completed the SET. The number of students in individual classes ranged from 7 to 47 per class. All available data were used in estimating variance components and reliabilities. Thus data from completed forms and forms with missing items were used.

Analysis

The following three phases are involved in deriving reliability coefficients using GT. First, observations must be modeled using a linear model. Second, the data and the model are used to estimate variance components. Third, estimated variance components are interpreted in terms of their impact on the reliability of the assessment. A variance component's impact on the reliability of assessment depends both on how the measurement is formed (i.e., measurement structure) and how the measurement is used in the decision-making process.

It should be acknowledged that the Seevers et al. (1998) instrument has two sections; one for the course and one for the instructor. Separate analyses for each section would be appropriate if separate scores were to be obtained. This paper focuses on an analysis to assess the reliability of means, particularly the class means, obtained using all 27 items.

The Model

To ascertain the correct model, sampling structure and components contributing to data variability must be identified. For SET data,

classes (c), items (i), and students (s) all contribute to variability in responses. While all items are administered to all students in all classes, students are nested within class. Thus the design is denoted by (s:c)xi to indicate the nesting of students within class and the crossing of items by students within class. This design and its application to assessing reliability of class means has been considered extensively by Brennan (1975), Kane, Gillmore, and Crooks (1976), and Kane and Brennan (1977). Both Brennan (1975) and Kane and Brennan (1977) write out the linear model in detail.

Table 1. Estimated Variance Components (Original Five-Point Likert Scoring)

Source of Variation	<u>df</u>	Estimated Variance Component	Percentage of Total Variance
Classes	25	0.16028	20.20%
Items	26	0.02569	3.24%
Students within classes (s,sc)	565	0.25784	32.50%
ci	650	0.03511	4.43%
error ^a	14366	0.3 1438	39.63%

Note. c = classes, i = items, s = students

^aerror combines a pure error component and the item-by-student interaction and class- by-item-by- student components

Estimates of Variance Components

GT does not specify any particular method for estimating variance components. While ANOVA estimates are commonly used, Restricted Maximum Likelihood Estimators (REMLs) often are preferred, particularly in the presence of unbalanced data (Marcoulides, 1990). Estimates computed using SAS[®] Proc Mixed (SAS Institute Inc., 1996) are presented (Table 1).

GT provides a broader view of the dependability of measurements through its emphasis on estimated variance components. A

starting point for GT is simply to consider and compare magnitudes of estimated variance components. For these data, the variance component for error is the largest component and accounts for 39.63% of the total variability. The student within class variability is the second largest variance component, accounting for 32.50% of the total variability. The variance component for class is substantially smaller than either of these.

Interpreting Variance Components

The estimated variance components' impact on a measure's reliability (i.e., a decision

made using a measure) depends on the following three considerations:

1. The universe of generalizability and whether facet conditions represented in the study are a sample of conditions or are the entire population of conditions;
2. Measurement structure or how the measurement is to be formed; and
3. The type of decision rule to be used.

GT considers two general types of decisions: relative decisions and absolute or criterion-referenced decisions. (See Brennan, 1983, for reliability coefficients when a specific single cut-off score is used.)

Regarding the generalizability of class means, Kane and Brennan (1977) write, "The intention to generalize to some larger universe of students is quite explicit whenever variation among students is used to estimate sampling error. Also, it is usually inappropriate to restrict generalization over items to the particular finite set of items used in some study" (p.289).

Even though the constructs proposed by Feldman (1988, 1989) and Rosenshine and Furst (1971) were used as a basis for generating instrument items, it is reasonable to consider the particular items chosen to be an exchangeable sample from some much larger universe of items that might be used and to which generalization is desired. Thus the model is a random model and generalizability (reliability) coefficients are calculated in a manner consistent with generalizing conclusions over a much larger (i.e., infinite universe of both items and students).

In the following subsections, reliability of data for two different measures having two different structures is considered. The first is a measure of student perceptions obtained by simply averaging (or summing) over items to produce a

single measure for each student. The second is a class mean obtained by summing over both items and students within the class. Measurement structure, together with the form of the decision rule, determine whether and how a variance component impacts reliability of the decision-making process. To illustrate both measurement structure and decision rule impacts, the reliabilities of relative assessment of student perceptions, relative assessment of class means, and criterion-referenced assessment of class means are calculated for these data.

Relative assessment of student perceptions

Reliability coefficients derived using GT follow the form (VanLeeuwen 1997):

$$\frac{\text{variance component for object of measurement}}{\text{variance component for object of measurement} + \text{error variance}}$$

Once variance components have been estimated, the difficulty is assessing the error variance for the measurement to be formed and the decision rule to be applied.

When assessing student perceptions, the object of measurement is students. The measurement is obtained by averaging the students' responses to all 27 items. For relative assessment of student perceptions, only those variance components affecting the rank orderings of the student means will impact decision-making. For this measure and decision rule, the only source of error is from the error variance component. Thus the error for relative assessment of student perceptions becomes $0.31438/27 = 0.011644$, and the reliability is given by $0.25784/(0.25784 + 0.011644) = 0.957$.

For these data, Cronbach's alpha is 0.970. The similarity of Cronbach's alpha and the reliability coefficient derived using GT is not surprising, since both are measures of inter-item consistency. They differ slightly because the GT

coefficient is based on variance component estimates from a more comprehensive model that includes class as a facet.

Relative assessment of classes

Measurements of classes involve averaging over both items and students within the class. The object of measurement is classes. The variance component for classes estimates how much variability is due to overall differences among the classes. Estimated variance components indicate that there is more variability among students within a class than there is from class to class. One implication of this is that using responses from only one student would provide a very unreliable class measurement.

Error for relative assessment of classes will be impacted by all interactions with classes. Thus the variance due to students within class, the class-by-item interaction, and the error all increase the error variance for relative assessment of classes. The error variance for relative assessment of classes is given by $0.25784/n_s + 0.0351/27 + 0.31438/(27 \times n_s)$, where n_s =the number of students in the class. Reliability will then be given by $0.16028/(0.16028 + \text{error variance})$.

The smallest class in this sample had only 7 students. Measurements for classes this size are estimated to have an error variance of 0.039798 and a reliability of 0.801. On the other hand, the largest class had 47 students. Measurements for classes of this size are estimated to have an error variance of 0.007034 and a reliability of 0.958. Assessments based on responses from a single student will have a reliability of only 0.372.

GT's emphasis on variance components and on calculating the decision error variance as well as the reliability coefficient provides decision-makers with information concerning the meaning that can be attributed to differences in class means. For example, for a class having only 7 students,

the standard error of the mean is 0.20 and the standard error for the difference between two class means (where both classes have 7 students) is 0.28. Class mean differences having a magnitude of 0.28 or less may not accurately indicate the correct rankings of the classes. That is, class mean differences this large may arise simply as a result of measurement error. In fact, it may be argued that differences of $2 \times 0.28 = 0.56$ can be attributed to chance variation (i.e., measurement error) and may not correctly indicate real differences in the class ranks.

Criterion-referenced assessment of classes

The error for criterion-referenced assessment is affected by all variance components affecting relative assessment. Additionally, criterion-referenced assessment also is affected by facet main effects and interactions among facets. For criterion-referenced assessment of classes, the error becomes $.02569/27 + 0.0351/27 + 0.31438/(27 \times n_s)$. As before, reliability will be given by $0.16028/(0.16028 + \text{error variance})$. Because the error variance is increased only by the inclusion of the item main effect variance component and because this component is relatively small, errors and reliabilities for criterion-referenced assessment will be very similar to those for relative assessment. If the item main effect component were relatively large, this would not be the case. For a class of size 7, the class mean will have a reliability of 0.797. A class of size 47 will have a reliability of 0.953.

Analysis Under Alternative Scoring

The above GT analysis and reliabilities are appropriate if all items have been scored appropriately (i.e., negative items reverse scored), and scoring used the original 5-point scale. However, it is not uncommon to use alternative scorings. One common alternative is to simply count the number of responses that are positive or

highly positive. This is equivalent to changing the scoring of each item to a 1 for original Likert scores of 4 or 5 and to a 0 for original Likert scores of 1, 2, or 3. Because this is a nonlinear transformation of the responses, reliabilities calculated using the data as originally scored will not be correct. Despite the fact that, for the

alternatively scored data, the percent of total variability attributable to classes has dropped from 20.20% to 13.96% (Table 2), reliabilities for the two scorings are similar (Table 3). This may not always be the case so it is important to assess reliabilities in a manner consistent with the scoring that will ultimately be used.

Table 2. Estimated Variance Components (Alternative Scoring: 1 = Positive Response, 0 = Neutral or Negative Responses)

Source of Variation	Estimated Variance Component	Percentage of Total Variance
Classes	0.01797	13.96%
Items	0.00264	2.05%
Students: Classes (s, sc)	0.02925	22.73%
ci	0.00499	3.88%
e, is, cis	0.07385	57.38%

Note. c = classes, i = items, e = error, s = students. Same df as in Table 1.

Discussion

These data provide preliminary information suggesting that the instrument developed by Seevers et al. (1998) is a good candidate for use in obtaining SET data. As noted in the next section, these data do not imply that SET class means can be interpreted as reliably ranking instructor performance. Data from more sophisticated designs need to be obtained and analyzed to assess reliability across course types. However, the standard errors derived using GT may provide some guidelines for interpreting differences in SET class means for instructors when controlling for subject matter and class clientele.

The GT analysis of the data focuses on variance components estimation and demonstrates that the reliability of student means does not imply the reliability of class means. O'Brien (1990)

stated, "It can be shown, however, both empirically and logically, that alphas based on individuals are not appropriate for assessing the reliability of aggregate-level variables. Their values may be substantially smaller or greater than the appropriate aggregate-level coefficients" (p. 475). He also notes that the conditions that lead to high reliability values of aggregate-level measures are not the same conditions that lead to high values of individual-level reliability coefficients. If, for example, all other things being equal, our variance component for classes had been 1/10 the magnitude, the reliability of student means for relative decisions would be unaffected. However, it would not be possible to obtain a reliable class mean for most class sizes in this sample. Since this situation is possible, it is important to use a process that yields correct estimates of reliability for the measurement and decision rule in use.

Table 3. Estimated Errors and Reliabilities for Alternative Scoring.

Relative assessment of student perceptions:

$$\text{error} = \frac{.07385}{27} = .002735$$

$$\text{reliability} = \frac{0.02925}{0.02925 + .002735} = .914$$

Relative assessment of classes (class size 7):

$$\text{error} = \frac{.02925}{7} + \frac{.00499}{27} + \frac{.07385}{27 \times 7} = .004754$$

$$\text{reliability} = \frac{.01797}{.01797 + .004754} = .791$$

Criterion-referenced assessment of classes (class size 7):

$$\text{error} = \frac{.00264}{27} + \frac{.02925}{7} + \frac{.00499}{27} + \frac{.07385}{7 \times 27} = .004852$$

$$\text{reliability} = \frac{.01797}{.01797 + .004852} = .787$$

Limitations

Gillmore, Kane and Naccarato (1978) recognized that in the design discussed here, having the class as the object of measurement confounds the teacher effect with the course effect. They discuss designs allowing separate estimation of teacher and course effects. Such designs then allow averaging over courses as well as items and students to obtain a measure of teaching that generalizes across courses and across students and items. Gillmore et al. 's (1978) suggestions may be improved upon by designs that recognize that every instructor does not teach

every course in a department's curriculum and almost certainly does not teach courses outside of a department's curriculum.

If the current assessment is viewed as an assessment of instructors, then course is an implicit facet and is confounded with the object of measurement (instructors) since neither courses nor instructors were replicated in these data (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; VanLeeuwen et al., 1998). If, in fact, it is desired to generalize across courses to come up with a truly dependable assessment of instructors, the error contributed by the course or course type

should be assessed and considered in estimates of reliability. That is, course or course type should appear as a facet in a G study (Shavelson & Webb, 1991). When this is not done, it is not clear whether differences in class means can be attributed to the instructor, the course, the subject matter in general, the level of the course or some combination of instructor, course, subject matter, and level. GT, however, does provide the tools to both plan data-gathering and model and analyze data gathered in such a way that any number of facets and their impact on teaching measurements can be considered. This is, in fact, an underused strength of GT. Kane (1993) notes that GT places the responsibility for formulating questions concerning the generalizability of measurements on the decision-maker. It is the decision-maker's responsibility to identify all facets to which it is desired to generalize and to consider those facets in sampling and modeling. In other words, the following questions should be asked about SET class means:

- Might the subject matter have an impact on the variability of class means?
- Might the course level have an impact?
- What about courses that cater to majors versus service courses that have primarily non-majors enrolled?

If the answers to any of these questions is yes, and if the decision-maker wants to compare instructors across subject matter, course level, and courses catering to different student populations, then a G study incorporating these variables as facets is needed to assess the reliability of class means to compare instructors. In fact, all possible levels of some of these facets may need to be included so they can be treated and analyzed as fixed facets.

Recommendations

The Seevers et al.(1998) instrument is

recommended for obtaining SET data. It is founded on a strong theoretical base and provides reliable class means and student means. Furthermore, it is recommended that GT be used to assess the reliability of class means generated from SET data, since Cronbachs alpha does not accurately estimate this reliability.

Until the impact of course type on class means is better understood, it is recommended that SET data not be used to compare instructors teaching different subject areas or different course levels. Furthermore, even when controlling for subject matter and course level, care should be taken in interpreting differences in class means when comparing performance of instructors. In particular, meaning should not be attributed to differences that are too small. That is, a difference in two means should be compared to the magnitude of the standard error for the difference to ascertain whether or not the observed difference is large enough to be meaningful.

As noted in the limitations, this study's data confounded course with instructor. Generalizing assessments of instructors across subject area and course level will require more detailed understanding of the impact of these facets on class means. It is recommended that GT's sophisticated framework be used to design additional studies that provide a firmer basis for comparing instructor performance. Within the framework supplied by GT, a number of approaches may be considered. One approach might be to consider a finite set of possible course types and to adjust class means for fixed course effects. The problem of obtaining generalizable assessments of teachers is by no means solved.

References

Brennan, R. L. (1983). Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Brennan, R. L. (1975). The calculation of

reliability from a split-plot factorial design. Educational and Psychological Measurement 35, 779-788.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: Wiley.

Feldman, K. A. (1988). Effective college teaching, from the students' and faculty's view: Matched or mismatched priorities. Research in Higher Education, 28 (4), 291-344.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multi-section validity studies. Research in Higher Education, 30 (6), 583-632.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of instruction: Estimation of the teacher and course components. Journal of Educational Measurements, 15 (1), 1-13.

Kane, M. T. (1993). Review of the book Generalizability Theory: A primer. Journal of Educational Measurement, 30 (3), 269-272.

Kane, M. T. & Brennan, Robert L. (1977). The generalizability of class means. Review of Educational Research, 47 (1), 267-292.

Kane, M. T., Gillmore, Gerald M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 13 (3), 171-183.

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. Psychological Reports, 66, 379-386.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. Sociological Methods and Research, 18, 473-504.

Rosenshine, B., & Furst, M. (1971). Research on teacher performance criteria. In B. O. Smith (ed.) Research in teaching education, 27-72. Englewood Cliffs, NJ: Prentice Hall.

SAS Institute Inc. (1996). SAS/STAT® Software: Changes and Enhancements through Release 6.11, Cary, NC: SAS Institute Inc.

Seevers, B. S., Dormody, T. J., & VanLeeuwen, D. M. (in press). Developing a valid and reliable student evaluation of teaching (SET) instrument, NACTA Journal.

Shavelson, R., & Webb, N. (1991). Generalizability theory: A primer. Newbury Park, CA: SAGE.

VanLeeuwen, D. M. (1997). Assessing reliability of measurements with generalizability theory: an application to inter-rater reliability. Journal of Agricultural Education, 38, 36-42.

VanLeeuwen, D. M., Barnes, M. D., & Pase, M. (1998). Generalizability Theory: A unified approach to assessing the dependability (reliability) of measurements in the health sciences. Journal of Outcome Measurement, 2, 302-325.