ED 394 984 TM 024 634

AUTHOR	Huang, Chi-yu; And Others
TITLE	A Generalizability Theory Approach To Examining Teaching Evaluation Instruments Completed by Students.
PUB DATE	Apr 95
NOTE	15p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Analysis of Variance; *College Students; *Course Evaluation; Evaluation Methods; *Generalizability Theory; Higher Education; *Student Evaluation of Teacher Performance; Test Construction; Test Reliability; *Test Use
IDENTIFIERS	*Variability

ABSTRACT

Generalizability theory is used to examine the sources of variability present in a teacher and course evaluation instrument. Two studies were conducted. In the first study, four different forms commonly used by one specific college of a large midwestern university were examined using responses of 915 students. The analysis of variance performed on each form separately indicated that one form did not generalize well across students in comparison with the other three. In the second study, the performance of a five-item form across three levels of courses (734 students in undergraduate, intermediate, and graduate courses) in one specific college were examined. The course evaluations from graduate level students were found to be more reliable. In this case, evaluations of undergraduate level courses should not be considered as reliable and generalizable as those collected in higher level courses. (Contains nine tables and nine references.) (Author/SLD)

A GENERALIZABILTY THEORY APPROACH TO EXAMINING TEACHING EVALUATION INSTRUMENTS COMPLETED BY STUDENTS

by

Chi-yu Huang
Shu-qin Guo
Cindy Druva-Roush
Joyce E. Moore

The University of Iowa
300 Jefferson Building
Iowa City, Iowa 52242-1470
(319) 335-0356

American Educational Research Association
San Francisco
April 1995

Abstract

Generalizability theory is used to examine the sources of variability present in a teacher and course evaluation instrument. Two studies were conducted. In the first study, four different forms commonly used by one specific college of a large midwestern university were examined. The analysis of variance performed on each form separately indicated that one form did not generalize well across students in comparison with the other three. In the second study, the performance of a five-item form across three levels of courses (undergraduate, intermediate, and graduate) in one specific college were examined. The course evaluations from graduate level students were found to be more reliable. In this case, evaluations of undergraduate level courses should not be considered as reliable and generalizable as those collected in higher level courses.

Introduction

With a decrease of available funding, there has been an increase in the demand for accountability. As a result, outcome assessment has become a popular topic. Outcome assessment involves three major components: student outcomes, course content, and teacher effectiveness. This paper will address the issue of assessing teacher effectiveness through student evaluation of the classroom environment. Specifically, the reliability of evaluation instruments filled out by students about their instructors will be studied through the use of generalizabilty theory.

Background

The primary system for faculty evaluation at the large mid-western university in this study consists of an item banking system where individual instructors can request a variety of items to be included on scannable forms. These forms are, for the most part, of a summative nature and are administered near the end of each academic semester. In the recent past, a need has been voiced by several departments to provide normative information to faculty. Providing norms requires creating an instrument or instruments with common items. The question then arises as to which items and what combinations should be used by a department or college; and how to determine which forms are more reliable and generalizable. Do identifiable characteristics of the students within a course affect reliability? This paper looks at the application of generalizability theory to examine the sources of variability within preprinted forms and the variance attributed to differences between forms and course level (undergraduate, intermediate, and graduate).

Generalizability theory (G theory) provides a broad framework for examining the reliability of behavioral measurements. It reflects not only the relationship between a set of observed scores and their true scores, but also the degree that a set of observed scores generalize to other situations (Shavelson and Webb, 1991). A behavioral measure is seen only as a random sample from a pool of behaviors. Facets (e.g. items, raters) and the objects of measurement (e.g. persons, classes) are first defined and included in the universe of admissible observations; and then variance components are obtained for each facet. Based on different universes of generalization, variance components are combined to produce coefficients of generalizability - commonly refered to as reliability coefficients.

Traditionally, researchers use rater stability and interrater consistency to reflect the reliability of teacher evaluations. However, instructors seldom teach more than one class of a particular course at a time. Measurement of interrator consistency using a split-half method is difficult because of small sample size and a test-retest method may be compromised by student memory. With G theory multiple sources of measurement error can be estimated in a single analysis and this information can be used to control the number in each facet to reduce the variance so that a dependable and efficient result with a sufficient level of reliability can be maintained.

Related Reseach

Previous research has examined the use of a specific set of common items across different classes. Hogan (1973) and Bausall, Schwartz, and Purohit (1975) correlated the average ratings for each of the following conditions: (1) two sections of the same course with the same teacher yielding a measure of interrater reliability for course-teacher combinations of classes, (2) two sections of different courses with the same teacher to isolate the teacher effect, and (3) two sections of the same course with different teachers to isolate the course effect. Neither study provided an independent estimate of the variance component due to teacher-course interaction. Gillmore, Kane, and Naccarato (1978) examined the generalizability of individual teachers across courses and then the generalizability of an individual course across various instructors. Gillmore, et al. summarize the findings of these three pieces of research:

> "All three studies agree in showing that generalizing just over students yields highly dependable results, generalizing over courses and students yields moderately dependable results, and generalizing over teachers and students yields non-dependable results." (Gillmore, et al., p.12)

These previous studies have determined that it is possible to arrive at a dependable measure of teaching effectiveness with a sample of five to ten average-sized courses (Gillmore, et al, 1978) over successive semesters. In these previous studies with common item sets, no examination of item/class interaction was made. When forms are constructed by individual faculty using a cafeteria-style system it is difficult to isolate common item sets. If common items can be identified and included on preprinted forms the task for faculty becomes one of identifying the appropriate form. Do forms vary in their dependability? Does a single form vary in dependability based upon characteristics of the course? In particular, does dependability of an instrument vary with the course level in which it is used? In this paper, generalizability theory will be used as a tool to answer these questions.

Purpose

The purpose of this paper is to:

1. Measure the reliability of teaching evaluation using G theory which generalizes over students, items, or both.

2. Estimate variance components for facets of student, item, and class for each form to reflect the magnitude of error in generalization. A comparison of reliability and magnitude of variance components for four different forms used in a single college will be made.

3. Compare the reliability and magnitude of variance compoents for three levels of courses for a single form used in one college.

Methods

The Design

For each form examined, each student responded to the same set of items but a different set of students rated each class. Items are crossed with classes and students are nested within class. Using Brennan's (1992) notation, i x s:c (item x student nested within class). The model is shown below:

$$X_{sci} = \mu + (\mu_c-\mu)+(\mu_i-\mu)+(\mu_{s:c}-\mu_c)+(\mu_{ci}-\mu_c-\mu_i+\mu)+(X_{sci}-\mu_{ci}-\mu_{s:c}+\mu_c)$$

Table 1 displays the sum of squares, expected variance terms for this model:

In generalizability theory, estimates of two error variances may be calculated: $\hat{\sigma}^2(\delta)$ - relative error variance and $\hat{\sigma}^2(\Delta)$ - absolute error variance. With the two facet, partially nested design proposed, these two terms are defined below:

$$\hat{\sigma}^2(\delta) = \hat{\sigma}^2_{ic}/n_c + \hat{\sigma}^2_{is:c}/n_c n_{s:c}$$

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2_c/n_c + \hat{\sigma}^2_{s:c}/n_{s:c} + \hat{\sigma}^2_{ic}/n_c + \hat{\sigma}^2_{is:c}/n_c n_{s:c}$$

Cronbach, Glaser, Nanda, and Rajaratnam (1972) define a reliability-like coefficient called a generalizabilty coefficient which is denoted as $E\rho^2$. A generalizability coefficient can be viewed as the ratio of universe score variance to expected observed score variance (Brennan, 1992). In attempting a general assessment of teaching effectiveness, it is appropriate to generalize over both items and students (Kane, Gillmore, and Crooks, 1976) (Gillmore, et al., 1978) and define an estimate of this generalizability as:

$$\hat{E}\,\rho^2(\mathrm{si}) = \hat{\sigma}^2_c / \{\hat{\sigma}^2_c + \hat{\sigma}^2(\delta)\}$$

Stability across only items or only students is defined similarly.

Study 1

Selection of sample

Four different forms commonly used by one specific college at a large midwestern university were examined. This data was collected in the 1994 Fall semester. Based on the proportion of classes administering each form, a random sample of classes was chosen for each form. To ensure a balanced design, fifteen students were randomly selected for each class if the number of students in the class was greater than fifteen. A description of each form's data is displayed in Table 2.

Results

The results of the analyses of variance performed for each form are displayed in Tables 3 through 6.

Findings

1. The percentage of total variance is largest on four forms for the student facet nested w'thin class.
2. Foım #0430 has the smallest generalizability coefficient.
3. The generalizability coefficients over items for all four forms are all very similar in magnitude.
4. The generalizability coefficient over students and items for form #0430 is discernably smaller that the coefficients for the other three forms.

Study 2

Sample Selection

An examination of the performance of the five-item form across three levels of courses in one specific college was performed. This data was collected during the 1993 Fall semester. The three levels of courses are defined by course number: undergraduate (course number less than 100), intermediate (course numbers 100 to 200) including both undergraduate and graduate students, and graduate (course numbers greater than 200). Based upon the proportion of classes offered at these three levels, varying sample sizes were chosen for each level - 20 for undergraduate, 14 for intermediate, and 9 for graduate. To ensure a balanced design, twenty students were selected from each class at the undergraduate and intermediate levels. Six students were randomly selected from each graduate level course. A separate analysis of variance was performed on each level with both item and student considered random and the class considered as the object of measurement. The results are displayed in tables 7 through 9.

Findings

1. The generalizability coefficients for undergraduate level courses are lower than the intermediate and graduate level courses. The generalizability coefficient increases substantially when the students are held fixed indicating that the undergraduate level students are less reliable judges of courses (i.e. results do not generalize well over different groups of students).
2. The generalizability coefficients for the intermediate level courses are similar to those of the graduate level courses when items are considered to be random. Both increase noticebly when students are considered fixed. However, when items are considered fixed, the generalizability coefficient for graduate level courses jumps above that of intermediate level courses. The graduate level students seem to be more reliable judges of course/instructor quality.

3. At both the undergraduate and intermediate levels, the major source of variability lies with students within a class. With the graduate level courses, some variability is shifted to the class facet. There is more consistent ratings within a class than between classes.

Discussion

Student evaluation of instruction information has long been considered a valuable data source for both the improvement of teaching and for inclusion in reviews associated with personnel decisions. It is believed that students, because of their exposure to professors, should know best whether teaching is adequate and whether they are learning (Cruse, 1987). However, little has been done in assessing the goodness of the evaluation instruments used. Generalizability theory offers a means to do this. Multiple sources of variation can be investigated and the results used to select instrument based upon the dependability of the item group and the characteristics of the population from which the evaluations are taken.

When various forms, differing in the number of items per form were compared within a single college, various sources of reliability could be investigated. Although a classical measure of internal consistency (Cronbach's alpha) suggested very little difference in the performance of the four forms (#2647 $\alpha = .926$, #2555 $\alpha = .934$, #2450 $\alpha = 924$, #0430 $\alpha = .910$), a generalizability analysis of the data results in somewhat different findings. The analyses of variance indicated that class was not a major sourse of variance. Ratings do not differ that much between classes. The evaluation findings generalized across a different set of items (high generalizability coefficients when items were considered random). However, the generalizability coefficient for form #0430 dropped dramatically when students were considered random and items fixed. That form does not generalize well across students. Although this form contains the least number of items (10), its coefficient of genealizability ($\hat{E}\rho^2(s)$ =.780) is quite a bit lower than the 11-item form (#2450) ($\hat{E}\rho^2(s) = .898$). This finding might suggest substituting form #2450 for form #0430.

When a single form was examined across three levels of courses, the graduate level students were found to be more reliable; and, their results more generalizable across students. Evaluations of undergraduate level courses should not be considered as reliable and generalizable as those collected in higher level courses.

References

Aubrecht, J.D. (1981). Reliability, validity and generalizability of student ratings of instruction. Idea Paper No. 6, Center for Faculty Evaluation and Development, Kansas State University.

Bausall, R.B., Schwartz.S, and Purohit, A. (1975). An examination of the conditions under which various student ratings parameters replicate across time. Journal of Educational Measurement, 12, 273-280.

Brennan, R.L. (1992). Elements of genealizability theory. ACT publications, Iowa City, Iowa. The American College Testing Program.

Crick, J.E. and Brennan, R.L. (1983). Manual for Genova: A generalized analysis of variance system. ACT Technical Bulletin, 43, Iowa City, Iowa: The American College Testing Program.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York:Wiley.

Cruse, D.B. (1987). Student evaluations and the university professor:*Caveat Professor*. Higher Education, 16, 723-737.

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

Hogan, T.P. (1973) Similarity of student ratings across instructors, courses, and time. Research in Higher Education, 1, 149-154.

Shavelson, R.J. and Webb, N.M. (1991). MMSS Generalizability theory: A primer, Vol 1. Newbury Park, California: Sage Publications, Inc.

Table 1. Analysis of variance components for completely model i x s:c design with all effects random.

Source(α)	df(α)	Sum of squares of mean squares T(α)	Sum of Squares (MS)	estimated variance $\hat{\sigma}^2(\alpha)$	estimated variability $\hat{\sigma}[\hat{\sigma}^2(\alpha)]$
item (i)	$n_i - 1$	$n_s n_c \sum \bar{X}_i^{\,2}$	$T(i) - T(\mu)$	$[MS_i - MS_{ic}]/n_s n_c$	$\{ [2(MS_i^2/\{df(i)+2\} + MS_{ic}^2/\{df(ic)+2\})]/n_s^2 n_c^2 \}^{1/2}$
class (c)	$n_c - 1$	$n_i n_s \sum \bar{X}_c^{\,2}$	$T(c) - T(\mu)$	$[MS_c - MS_{ic} - MS_{s:c} + MS_{is:c}]/n_i n_s$	$\{ [2(MS_c^2/\{df(c)+2\} + MS_{ic}^2/\{df(ic)+2\} + MS_{s:c}^2/\{df(s:c)+2\} + MS_{is:c}^2/\{df(is:c)+2\})]/n_i^2 n_s^2 \}^{1/2}$
students: class (s:c)	$n_c(n_s - 1)$	$n_i \sum \bar{X}_{s:c}^{\,2}$	$T(s:c) - T(i)$	$[MS_{s:c} - MS_{is:c}]/n_i$	$\{ [2(MS_{s:c}^2/\{df(s:c)+2\} + MS_{ic:c}^2/\{df(is:c)+2\})]/n_i^2 \}^{1/2}$
i x c	$(n_i - 1)(n_c - 1)$	$n_s \sum \bar{X}_{ic}^{\,2}$	$T(ic) - T(i) - T(c) + T(\mu)$	$[MS_i - MS_{ic}]/n_s$	$\{ [2(MS_{ic}^2/\{df(ic)+2\} - MS_{is:c}^2/\{df(is:c)+2\})]/n_s^2 \}^{1/2}$
i x s:c	$n_c(n_s - 1)(n_i - 1)$	$\sum\sum \sum \bar{X}_{is:c}^{\,2}$	$T(is:c) - T(ic) - T(s:c) + T(c)$	$[MS_{ic} - MS_{is:c}]$	$\{ 2MS_{is:c}^2/\{df(is:c)+2\}\} \}^{1/2}$
Mean (μ)		$n_i n_s n_c \bar{X}^{\,2}$			
Total	$n_i n_s n_c - 1$		$T(is:c) - T(\mu)$		

Table 2. Description of forms and sample size.

	number of items	number of classes	# of students/class
Form #2647	18	6	15
Form #2555	12	15	15
Form #2450	11	15	15
Form #0430	10	25	15

Table 3. Analysis of variance of course evaluations by students for form #2647 (both items and students are considered random)

Effect	df	mean squares	estimated variance	percentage of variance
item (i)	17	10.074	.099	9.55
class (c)	5	49.993	.166	16.01
students: class (s:c)	84	4.659	.231	22.28
i x c	85	1.146	.043	4.17
i x s:c,e	1428	.498	.498	48.02

$\hat{E}\,\rho^2(si) = .894$ both students and items random

$\hat{E}\,\rho^2(i) = .977$ students fixed, items random

$\hat{E}\,\rho^2(s) = .907$ students random, items fixed

Table 4. Analysis of variance of course evaluations by students for form #2555 (both items and students are considered random)

Effect	df	mean squares	estimated variance	percentage of variance
item (i)	11	14.569	.057	4.34
class (c)	4	58.593	.289	21.99
students: class (s:c)	210	5.211	.394	29.98
i x c	154	1.821	.089	6.77
i x s:c,e	2310	.485	.485	36.91

$\hat{E}\,\rho^2(si) = .888$ both students and items random

$\hat{E}\,\rho^2(i) = .969$ students fixed, items random

$\hat{E}\,\rho^2(s) = .911$ students random, items fixed

Table 5. Analysis of variance of course evaluations by students for form #2450 (both items and students are considered random)

Effect	df	mean squares	estimated variance	percentage of variance
item (i)	10	35.820	.150	10.08
class (c)	14	52.202	.275	18.48
students: class (s:c)	210	5.299	.434	29.17
i x c	140	2.128	.107	7.19
i x s:c,e	2100	.522	.522	35.08

$\hat{E}\,\rho^2(si) = .868$ both students and items random

$\hat{E}\,\rho^2(i) = .959$ students fixed, items random

$\hat{E}\,\rho^2(s) = .898$ students random, items fixed

Table 6. Analysis of variance of course evaluations by students for form #0430 (both items and students are considered random)

Effect	df	mean squares	estimated varianr^	percentage of variance
item (i)	9	37.810	.097	7.55
class (c)	24	24.234	.119	9.26
students: class (s:c)	350	5.326	.481	37.43
i x c	216	1.558	.069	5.37
i x s:c,e	3150	.519	.519	40.39

$\hat{E}\,\rho^2(si) = .737$ both students and items random

$\hat{E}\,\rho^2(i) = .936$ students fixed, items random

$\hat{E}\,\rho^2(s) = .780$ students random, items fixed

Table 7. Analysis of variance of course evaluations by students for undergraduate level courses (both items and students are considered random)

Effect	df	mean squares	estimated variance	percentage of variance
item (i)	4	12.778	.030	2.97
class (c)	19	7.232	.039	3.86
students: class (s:c)	380	2.987	.515	50.99
i x c	76	.740	.017	1.68
i x s:c,e	1520	.409	.409	40.50

$\hat{E}\,\rho^2(si) = .541$ both students and items random

$\hat{E}\,\rho^2(i) = .898$ students fixed, items random

$\hat{E}\,\rho^2(s) = .587$ students random, items fixed

Table 8. Analysis of variance of course evaluations by students for intermediate level courses (both items and students are considered random)

Effect	df	mean squares	estimated variance	percentage of variance
item (i)	4	14.597	.047	2.75
class (c)	13	24.597	.181	10.60
students: class (s:c)	266	5.596	1.041	60.95
i x c	52	1.340	.047	2.75
i x s:c,e	1064	.392	.392	22.95

$\hat{E}\,\rho^2(si) = .734$ both students and items random

$\hat{E}\,\rho^2(i) = .946$ students fixed, items random

$\hat{E}\,\rho^2(s) = .772$ students random, items fixed

Table 9. Analysis of variance of course evaluations by students for graduate level courses (both items and students are considered random)

Effect	df	mean squares	estimated variance	percentage of variance
item (i)	4	2.537	.032	3.40
class (c)	8	9.004	.233	24.79
students: class (s:c)	45	1.585	.246	26.17
i x c	32	.795	.073	7.77
i x s:c,e	180	.356	.356	37.87

$\hat{E}\,\rho^2(si) = .775$ both students and items random

$\hat{E}\,\rho^2(i) = .912$ students fixed, items random

$\hat{E}\,\rho^2(s) = .824$ students random, items fixed