DOCUMENT RESUME

ED 335 399 TM 017 061

AUTHOR	Barnes, Laura L. B.; Barnes, Michael W.
TITLE	The Effects of Academic Discipline on Generalizability of Student Evaluations of Instruction.
PUB DATE	Apr 91
NOTE	25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*College Students; *Context Effect; *Course Content; *Generalizability Theory; Higher Education; *Professors; Rating Scales; Research Problems; *Student Evaluation of Teacher Performance

ABSTRACT

Although student evaluation data provide a reasonable basis for making decisions about instructors when generalizing across courses and students, when the course is the object of measurement (OM), data are less generalizable. This finding may be due to the type of evaluation items used or to academic discipline differences in the type of courses selected for study. This study addressed this problem by using A. Biglan's (1973) model for classifying disciplines along the dimensions of paradigmatic/preparadigmatic (hard/soft) and pure/applied. A nested sampling procedure yielded two sample types: courses within teachers, in which individual instructors taught more than one course; and teachers within courses, in which individual courses were taught by more than one instructor. For each sample type, evaluation forms for 20 courses within each discipline classification were sought. Thirty items from a survey, which required students at a private doctoral-granting institution in the Southwest to rate their instructors on a 0-5 scale, were used. The evaluation items for this study measured six dimensions of instruction: organization, breadth of coverage, group interaction, enthusiasm, grading, and individual rapport. Generalizability and decision studies were conducted in which, for one sample, teacher was the OM, and for the second sample, course was the OM. Reliable decisions about instructors could reasonably be made from all six of the evaluation dimensions. However, reliability for course decisions varied greatly with the evaluation dimension, being highest for breadth of coverage and lowest for grading. The same general pattern was noted for the paradigmatic disciplines and the preparadigmatic-applied disciplines, but not for the preparadigmatic-pure disciplines. Two figures and five tables present study data. (Author/RLC)

The effects of academic discipline on generalizability of student evaluations of instruction

Laura L. B. Barnes

Oklahoma State University

and

Michael W. Barnes

The University of Tulsa

Paper presented to the annual meeting of the National Council on Measurement in Education, April, 1991, Chicago.

Running head: G-theory and discipline

Abstract

Previous research on the generalizability of student ratings of instruction has raised questions about the effects of academic discipline and item types on the generalizability of these data for making relative decisions about instructors and about courses. In particular, although student evaluation data appear to provide a reasonable basis for making decisions about instructors when generalizing across courses and students, when course is the object of measurement, the data appear to be less generalizable. It was suggested in the literature that this may be due to the type of evaluation items used or it may be due to academic discipline differences in the type of courses selected for study. This study used Biglan's (1973) model for classifying disciplines along the dimensions of paradigmatic/preparadigmatic (hard/soft) and pure/applied. A nested sampling procedure yielded two sample types: courses within teachers, in which individual instructors taught more than one course; and teachers within courses, in which individual courses were taught by more than one instructor. For each sample type, evaluation forms for twenty courses within each discipline classification were sought. The evaluation items for this study were classified as measuring six dimensions of instruction: organization, breadth of coverage, group interaction, enthusiasm, grading, and individual rapport. Generalizability and decision studies were conducted in which, for one sample, teacher was the object of measurement, and for the second sample, course was the

object of measurement. Results indicated that reliable decisions about instructors could reasonably be made from all six of the evaluation dimensions; however, reliability for course decisions varied greatly with the evaluation dimension, being highest for breadth of coverage and lowest for grading. The same general pattern was noted for the paradigmatic disicplines and the preparadigmatic-applied disciplines but not for the preparadigmatic-pure disciplines.

The issue of reliability is of great concern in using student evaluations of instruction for making comparative decisions about faculty and courses. In this regard, Generalizability Theory has had demonstrated utility because it requires the researcher to explicitly identify the sources of variability that are to be considered error, distinct from those sources that are to be treated as universe score variance. The latter term is analogous to the true score variance in classical test theory (Brennan, 1983). In place of the classical reliability coefficient, this methodology yields a generalizability coefficient which can be interpreted in roughly the same manner, but is much more versatile in its application. Generalizability Theory can accommodate a variety of research designs and therefore has been particularly useful when the desire is to isolate and test specific sources of variability. In applications to student evaluations of instruction these sources of variability typically may include students, courses, occasions, and items.

One particular question to which Generalizability Theory has been applied is that of how student evaluation data can dependably be used to make comparative decisions about instructors independently of the courses they teach. In other words, how much of the variability in students' ratings among instructors is actually due to differences in courses as opposed to differences in instructors. Likewise, if the data are to be used to make decisions about courses (e.g., How do students rate Psych 101 relative to other Psych courses?), it is necessary to separate the course effects from the effects of different instructors teaching the course.

To this end, Gillmore, Kane, and Naccarato (1976) drew two separate random samples of courses. In one sample, they selected evaluations from instructors who had taught two different courses and in the second sample, they selected evaluations for courses that had been taught by two different instructors. When teachers were the objects of measurement (i.e., relative decisions were to be made about teachers), they found generalizability coefficients to be quite satisfactory. However, with courses as objects of measurement (i.e., relative decisions were to be made about courses) the dependability of the measures across samples of teachers, students, and items was low. In response, Smith (1979) suggested that the evaluation items utilized by Gillmore et al. (1976) were not equally useful for making decisions about instructors and courses. Essentially he said that if decisions are to be made about courses, they should be based on items that solicit students' perceptions of the course, not the instructor and similarly, that decisions about instructors should be based on instructor-related items. Employing a similar sampling design, Smith (1979) found that with course as the object of measurement, generalizabiltiy coefficients were small when based on instructor-related items; however with course-related items the coefficients indicated that reasonably dependable judgments could be made about courses. Likewise, generalizability coefficients were much higher for making decisions about instructors with instructor-related than course-related items.

Gillmore (1980) suggested that the discrepancies between his (1976) study and Smith's (1979) study were not totally resolved by the use of different type items. In particular, he noted that whereas Gillmore et al. (1976) drew their sample from a variety of discipline areas, Smith drew his sample only from an Educational Psychology department. Although he attempted to resolve the issue by replicating the study with samples drawn from three disparate discipline areas, he reported being unsuccessful due to the presence of negative estimates of variance components for important main effects in the model.

Marsh (1981), although not utilizing Generalizability Theory, reported a similar sampling design in addressing the issue. He utilized an evaluation instrument (Students' Evaluations of Educational Quality, SEEQ) that contained much more specific questions measuring nine different components of instruction. He reported larger correlations between courses with teacher as the object of measurement than between teachers with course as the object of measurement. Although apparently for this instrument, the instructor effect overshadowed the course effect, the differences in magnitudes of correlations depended to some extent on the component of evaluation. For example, components related to assignments, workload/difficulty, and group interaction had relatively higher correlations between two teachers with course as the object of measurement than did other components (e.g., examinations/grading).

Thus, these two issues remain unresolved in the literature - that is, what are the effects of academic discipline differences and items on the generalizability of student evaluations of instruction for decisions about courses and instructors? Biglan (1973) presented a theoretical model for studying academic discipline differences based on a three-dimensional classification system.

According to this model academic disciplines may be characterized by the presence (or absence) of a single predominant paradigm (paradigmatic versus preparadigmatic). Examples of paradigmatic disciplines are engineering, and the physical and life sciences. Preparadigmatic disciplines include the humanities and the social and behavioral sciences. Paradigmatic disciplines are often referred to as hard disciplines, preparadigmatic disciplines, as soft disciplines. The second dimension is whether the discipline is oriented to application - the pure/applied dimension. An example of a pure-hard discipline is mathematics; applied-hard, mechanical engineering; soft-pure, sociology; and soft-applied, educational administration. The third dimension is whether the discipline is oriented to the study of life (e.g., biology) or nonlife (e.g., computer science).

In applications of this model to the study of student ratings of instruction, the life/nonlife dimension has proven less useful than the other two. Neumann and Neumann (1985) reported that the predictors of overall teacher assessment differed primarily along the hard/soft dimension. For the soft disciplines, items assessing student involvement, cognitive contribution of the course, and level of instruction were all important predictors. However, only level of instruction emerged as a significant predictor of overall instructional rating for the hard or paradigmatic disciplines. Barnes and Patterson (1988) found that soft disciplines received generally higher ratings than hard disciplines, particularly on items that reflect a breadth of coverage (e.g., contrasted implications of various theories). Thus, it may be anticipated that the relative utility of evaluation components for course or teacher decisions depends on the academic discipline area. Our study was designed to address this issue by utilizing the sampling scheme discussed above, and to extend it to the four discipline areas suggested by the Biglan model. In addition, an evaluation instrument, similar to the SEEQ, was used so that partial replication of Marsh's (1981) study within a Generalizability Theory framework would be possible.

Method

Data for this study came from a private doctoral-granting institution in the Southwest. The instrument used was a 34-item survey developed by a university committee and contained items similar to the SEEQ instrument discussed above. Thirty of the items required students to specifically rate their instructors on a 0-5 scale; four of the items asked for student background data and were not included in this study. The instrument was used to evaluate faculty university-wide. A principle components analysis with oblique rotation yielded six interpretable evaluation components similar to those reported by Marsh (1984). The components listed in order of extraction and with the number of items associated with them are Organization (5), Breadth of Coverage (5), Group Interaction (3), Enthusiasm (5), Grading (3), and Individual Rapport (3). The six evaluation components (hereinafter referred to as dimensions to avoid confusion with variance components associated with Generalizability Theory) and the items associated with them are listed in Figure 1.

Prior to sample selection, all courses for which evaluation data were available were categorized by discipline area according to the hard/soft and pure/applied categories of the Biglan model.

Courses that were not clearly identifiable with one of these four Biglan classifications were not included in the population from which the sample was drawn.

Two samples were selected for this study. For the first sample, within each of the four discipline areas, instructors were identified who had taught at least two different courses (not different sections of the same course) for an academic year. Once the instructors were identified twenty instructors, for whom there were at least ten completed rating forms for at least two courses. were randomly selected from each discipline area. When an instructor taught more than two courses meeting the above criteria, two courses were randomly selected and ten forms were randomly selected from each course. This sample then, consisted of rating forms on instructors teaching two different courses and was termed the courses within teacher (C:T) sample. The second sample was obtained by identifying, within each discipline area, courses that had been taught by at least two different instructors over the same academic year. Once the courses were identified, we attempted to randomly select twenty from each discipline area subject to the stipulation that for each course there must be at least ten completed forms for each instructor teaching that course. However, we were successful in obtaining twenty each only for the soft-pure and soft-applied disciplines. Only three courses qualified from the hard-applied dimension and only 15 from the hard-pure dimensions. (Particularly in the hard-applied areas there are fewer multisection courses, and because these areas tend to be highly specialized the same faculty member teaches every instance of a course offering). These 18 courses were combined into an undifferentiated hard discipline category. For all groups, if a course was taught by more than two instructors, two instructors were randomly selected and ten forms were randomly selected from each instructor. This sample consisted of rating forms on courses taught by two different instructors and was termed the teachers within course (T:C) sample. Figure 2 displays the major course headings and our operationalization of their Biglan classifications.

The first analysis involved pooling the data across the Biglan classifications and conducting generalizability and decision studies for the separate course within teacher and teacher within course samples. The anlyses were conducted separately for each evaluation dimension. The design of the analyses were (s:c:t) x i and (s:t:c) x i, for the course within teacher and teacher within course samples, respectively, and where i stands for items. This design provided information regarding overall differences among the dimensions in terms of their usefulness for making teacher and course decisions. It also provided a useful baseline for comparing our results with those reported elsewhere in the literature. The second set of analyses involved separate (s:c:t) x i and (s:t:c) x i designs for each of the evaluation dimensions within each of the four discipline areas. These generalizability and decision studies spoke directly to the issue of discipline differences discussed earlier.

Results

The variance component estimates for the course within teacher sample are examined first. As displayed in Table 1, for all dimensions the variance components for students nested within courses and for the item by student interaction (confounded with random error) are large. The variance components for items, the item by course interaction, and the item by teacher interaction range from near zero to intermediate values. The magnitude of the item effects appears to be unrelated to the number of items. The standard errors for the variance components for items, however, tend to be smaller for Organization and Breadth of Coverage dimensions and to be larger for Individual Rapport. This is understandable because Organization was the first factor extracted and Individual Rapport was the last factor, and thus may be less stable. For all dimensions, the variance component estimates for the teacher effect are larger than the course within teacher effect. These results are consistent in pattern with those reported by Smith (1979) for both his Instructor and Course items. The pattern is not consistent with Gillmore et al. (1978) who found for undifferentiated items the course within teacher effect to be somewhat larger than the teacher effect.

The teacher within course sample provides somewhat mixed results. Again, the largest effects are for students and the item by student interaction. The pattern for the item effect is similar to that for the first sample and again, the standard errors for the estimated variance components for items are smallest for Organization and Breadth of Coverage. However, the teacher within course effect and the course effects are of at least intermediate value for almost all dimensions. When considering the difference between the teacher within course and the course effect, only for Grading could the difference in favor of the teacher within course effect be considered large given the magnitudes of the standard errors. The larger teacher effect for Grading is evident in both samples. This rating dimension would appear to be much more influenced by the instructor than by the course. On the other hand, for Breadth of Coverage, the course effect is much larger than the teacher within course effect in the T:C sample, and the reverse is true for the C:T sample although the difference is not as large. This suggests that ratings of Breadth of Coverage, while being influenced to some extent by the instructor, is largely a reflection of the particular course.

Before presenting results of the decision studies which are based on these estimates, we note that comparisons of these two samples rest upon an assumption that they are essentially similar samples of the same population, differing only in the way they were nested. As such, they should yield essentially similar estimates of generalizability for individual instructor/course combinations. Given the difficulty in obtaining the teacher within course sample described above, we were concerned that a systemmatic bias may have been introduced in the sampling procedure, so a test of this assumption seems appropriate. Following practice reported in Gillmore et al. (1978) and Smith (1979), generalizability coefficients were computed for both samples in which generalization was taken only across students and items. In the teacher within course sample, this meant the universe of generalization contained items and students randomly sampled from an infinite universe but only one teacher. Similarly, in the course within teacher sample, generalization was across items and students, and only one course. These coefficients are reported in the second and fourth columns of Table 2 and show that for samples of 5 items and 20 students, with one course and one teacher respectively, the two samples yielded quite similar results for all dimensions. Thus, there were no apparent systemmatic differences in these two samples.

When decisions are to be made about teachers generalizing over courses, students, and items, Table 2 indicates that Group Interaction, Breadth of Coverage, and Organization items provide the most reliable discriminations among teachers. Even with only two courses per teacher, generalizability coefficients for these three dimensions are above .70. With five courses, generalizability coefficients are above .80 for all dimensions except Individual Rapport. However, when course is the object of measurement, the magnitude of the coefficients for generalizing across teachers, students, and items depends greatly on the dimension being evaluated. Breadth of Coverage items provide the most dependable information for these types of decisions, and is the only dimension with a generalizability coefficient above .80 for either two or five teachers. The generalizability coefficients indicate that evaluations of Individual Rapport are as reliable for course decisions as they are for decisions about instructors. Course decisions based upon evaluations of Grading and Group Interaction, however, cannot be dependably made with five teachers.

Generally, these findings are consistent with Marsh's (1981) results in which, using a similar sampling design, he found a stronger relationship between ratings of two courses taught by the same instructor (C:T) than between ratings of two instructors teaching the same course (T:C). In light of the fact that students were specifically instructed to rate the instructor and when considering the nature of the dimensions evaluated, it is not surprising that Group Interaction, for example, would be more valuable for rating instructors than courses. Nor is it surprising that Grading provides a poor basis for evaluating courses. On the other hand, it is puzzling that Individual Rapport yielded similar results for instructor and course decisions. Breadth of Coverage also showed similar results for the two types of decisions. This is easier to understand because some courses do not lend themselves to the type of presentation suggested by these items. So, when students rate the extent to which their instructor contrasted implications of various theories, for example, this should show up not only as variability among instructors, but also as variability among courses.

Tables 3 and 4 present the generalizability coefficients for the dimensions separately for the discipline areas with teacher and course as the object of measurement, respectively. Both tables are based on decision study samples of 20 students, and five items. In Table 3 coefficients are given for samples of two and five instructors. Both hard pure and hard-applied disciplines reflect the same general pattern as reported for the pooled samples - that is, in general, all of the dimensions provide for reasonable discriminations to be made among instructors. However, for the soft pure disciplines, Breadth of Coverage, Grading, and Individual Rapport do not appear to provide a reliable basis for discriminating among instructors. For these courses, Group Interaction has the highest coefficient. For the soft-applied disciplines, Enthusiasm appears to be the only weak basis for making decisions about instructors.

In Table 4, both the undifferentiated hard and the soft-pure disciplines contain coefficients that are zero. When estimated variance components are calculated to be negative, values of zero are often substituted for the negative components and consequently, generalizability coefficients (if calculated) are zero (Brennan, 1983). However, in each case reported here, the zero variance components are apparently legitimate, and are not the result of negative estimates. There was evidently no variance attributable to the course effect for Grading in the Hard disciplines, nor for Organization, Group Interaction, or Grading in the soft-pure disciplines. Thus, these dimensions provide no basis for differentiating among courses for these disciplines in our sample. For the hard and the soft-pure disciplines, Breadth of Coverage seems to provide the most dependable basis for course decisions. In the soft-applied disciplines, all of the dimensions provide a reasonable basis for course decisions.

Discussion.

These results indicate that when comparative ratings of instructors are desired, the dimensions assessed by this instrument provide a dependable basis for decision-making with as few as two courses per instructor. However, when comparative decisions are to be made about courses, the dimensions are not equally informative. For example, although Breadth of Coverage was found to be a dependable basis upon which to discriminate among courses, Grading and Group Interaction were not. Although the salience of the course effect in student ratings of Individual Rapport remains a puzzle, in general the findings support the validity of student ratings as an instructor-oriented construct.

Unfortunately, less can be said about discipline differences in the constructs. Given the small sample sizes available for computing the variance component estimates, and given the lack of a course within teacher sample for the hard-applied disciplines, any interpretations of differences must be extremely tentative. It appears that the type of decision that can reliably be made may depend on discipline area. For example, Enthusiasm appears to be an instructor-related construct for hard and soft-pure disciplines, but may have more variance attributable to courses in the soft-applied areas. Although Breadth of Coverage is not useful in the soft-pure disciplines for making instructor decisions, it appears quite useful for making course decisions. Evidently, in disciplines such as the humanities and social sciences, ratings of this dimension vary among courses, even among those taught by the same instructor, but vary less among instructors. At first glance, this seems counterintuitive, in that one would expect the social sciences, and to a lesser extent the humanities, to deal with theory so it seems that this should be less a course characteristic than an instructor characteristic. On the other hand, if all instructors in this discipline area have been socialized to deal with theory in their teaching, then when students rate the extent to which their instructor did so, there ought to be little variability associated with the instructor effect. In this case, the course variability may be attributable to course level (upper versus lower division) or type of course content (e.g., theory versus methods courses).

An alternative explanation lies in our operationalization of the Biglan model as a framework for understanding discipline differences in student evaluations of teaching. Biglan's original (1973) classification system contained a life/nonlife dimension that, had we utilized it, would have separated the humanities and social sciences. This dimension would also have separated out the more quantitatively oriented soft-applied disciplines. It is possible that some of the course variability in the soft disciplines may be attributable to the life/nonlife distinction. As has been suggested elsewhere (e.g., Gillmore, 1980), little is known about how discipline differences affect student evaluations. Gillmore (1980) attempted to address this issue but was, in his words, "thwarted", primarily because of the lack of adequate data. He and others (e.g., Marsh, 1984) have suggested the pooling of data across universities. In particular, it would be useful to pool data from institutions that use an evaluation form of similar dimensionality.

References

Barnes, M. W., & Patterson, R. H. (1988, August). Using teaching evaluation results to plan department personnel strategies to accomplish the institutional teaching mission. Paper presented at the annual meeting of the Society for College and University Planning, Toronto.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. Journal of Applied Psychology, 57(3), 195-203.

Brennan, R. L. (1983). Elements of generalizability theory. The American College Testing Program.

Gillmore, G. M. (1980, April). Student instructional ratings: To what universe can we dependably generalize results? Paper presented at the annual meeting of the American Educational Research Association, Boston.

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

Marsh, H. W. (1981). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. Applied Psychological Measurement, 6, 47-60.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

Neumann, L., & Neumann, Y. (1985). Determinants of students' instructional evaluation: A comparison of four levels of academic areas. Journal of Educational Research, 78 (3), 152-158.

Smith, P. L. (1979). The generalizability of student ratings of courses: Asking the right questions. Journal of Educational Measurement, 16(2), 77-87).

Figure 1. Summary of evaluation items.

Organization
Paced course appropriately
Set reasonable course requirements
Communicated expectations near beginning of course
Explained how course grade would be determined
Returned assignments promptly

Breadth of coverage
Presented historical origins of ideas and concepts
Discussed recent developments in the field
Contrasted implications of various theories
Discussed points of view other than his/her own
Gave references for more interesting and involved points

Group interaction
Encouraged students to ask questions
Encouraged students to express own ideas
Attempted to determine student understanding of material

Enthusiasm
Held students' attention in class
Presentations were thought provoking/stimulating
Used examples and illustrations to clarify
Presented material coherently
Was well-prepared for lectures/discussion

Grading
Tests allowed students to demonstrate learning
Test questions were clearly written
Had sufficient evidence to evaluate achievement

Individual Rapport
Was available outside of class
Respected students as individuals
Commented individually on students' work

Figure 2. Courses by Biglan Classification.

	Course within Teacher		Teacher within Course	
	Hard	**Soft**	**Hard**	**Soft**
Pure	Biology Math Chemistry Physics Geology	History Sociology English Political Science Anthropology Communications	Biology Math Chemistry	Spanish Sociology English Political Science Anthropology Communications Philosophy Psychology
Applied	Chemical Engineering Electrical Engineering Engineering Science Petroleum Engineering Computer Science Mechanical Engineering	Education Music Theater Accounting Economics Marketing Management Nursing Communicative Disorders Finance Physical Education	Computer Science Engineering Science Mechanical Engineering	Education Music Theater Accounting Economics Marketing Management Nursing

Table 1. Variance components for full samples.

Courses within teachers

Source	Variance Components	Standard Error
Organization		
t	.131	.032
c:t	.043	.019
s:c:t	.525	.025
i	.010	.006
ti	.025	.009
ci:t	.053	.010
si:c:t	.742	.014
Breadth of Coverage		
t	.271	.064
c:t	.110	.035
s:c:t	.758	.035
i	.001	.002
ti	.062	.014
ci:t	.085	.014
si:c:t	.889	.017
Group Interaction		
t	.170	.042
c:t	.044	.024
s:c:t	.690	.035
i	.047	.034
ti	.045	.012
ci:t	.032	.011
si:c:t	.681	.018
Enthusiasm		
t	.146	.040
c:t	.098	.028
s:c:t	.642	.029
i	.069	.040
ti	.030	.006
ci:t	.007	.006
si:c:t	.640	.012
Grading		
t	.218	.063
c:t	.138	.045
s:c:t	.766	.038
i	.045	.033
ti	.049	.020
ci:t	.130	.023
si:c:t	.752	.020
Individual Rapport		
t	.115	.040
c:t	.093	.030
s:c:t	.563	.033
i	.088	.063
ti	.081	.017
ci:t	.028	.013
si:c:t	.908	.024

Teachers within courses

Source	Variance Components	Standard Error
Organization		
c	.080	.034
t:c	.084	.031
s:t:c	.581	.032
i	.003	.002
ci	.008	.009
ti:c	.059	.013
si:t:c	.773	.017
Breadth of Coverage		
c	.393	.095
t:c	.089	.035
s:t:c	.742	.040
i	.005	.004
ci	.081	.016
ti:c	.054	.013
si:t:c	.863	.019
Group Interaction		
c	.073	.050
t:c	.169	.050
s:t:c	.745	.043
i	.037	.028
ci	.074	.015
ti:c	.011	.010
si:t:c	.650	.020
Enthusiasm		
c	.100	.050
t:c	.183	.048
s:t:c	.602	.032
i	.069	.041
ci	.010	.006
ti:c	.022	.008
si:t:c	.610	.013
Grading		
c	.042	.055
t:c	.217	.063
s:t:c	.749	.046
i	.058	.043
ci	.091	.024
ti:c	.060	.019
si:t:c	.857	.027
Individual Rapport		
c	.098	.040
t:c	.075	.034
s:t:c	.630	.040
i	.042	.031
ci	.026	.015
ti:c	.057	.018
si:t:c	.817	.025

Table 2. Generalizability coefficients for full samples.

$n'_s = 20,\ n'_i = 5$	Courses within Teacher			Teachers within Course		
	n'_c	$\hat{\varepsilon\rho}^2_{C,S,I}$	$\hat{\varepsilon\rho}^2_{C^*,S,I}$	n'_t	$\hat{\varepsilon\rho}^2_{T,S,I}$	$\hat{\varepsilon\rho}^2_{T^*,S,I}$
Organization						
	1		.780	1		.776
	2	.729		2	.539	
	5	.853		5	.738	
Breadth						
	1		.833	1		.869
	2	.731		2	.815	
	5	.851		5	.896	
Group Interaction						
	1		.789	1		.799
	2	.756		2	.373	
	5	.861		5	.557	
Enthusiasm						
	1		.842	1		.869
	2	.662		2	.467	
	5	.814		5	.681	
Grading						
	1		.813	1		.773
	2	.654		2	.212	
	5	.807		5	.364	
Individual Rapport						
	1		.779	1		.755
	2	.577		2	.589	
	5	.726		5	.763	

Table 3. Generalizability coefficients by discipline for Course within Teacher Sample.

$n'_s=20$, $n'_i=5$ $(E\hat{\rho}^2_{C,S,I})$

	Organization	Breadth of Coverage	Group Interaction	Enthusiasm	Grading	Individual Rapport
Hard-pure						
n'_c						
2	.608	.875	.805	.888	.692	.512
5	.790	.939	.894	.943	.849	.674
Hard-applied						
n'_c						
2	.764	.693	.843	.845	.636	.764
5	.835	.829	.923	.923	.795	.842
Soft-pure						
n'_c						
2	.505	.241	.652	.475	.307	.330
5	.711	.399	.824	.676	.487	.501
Soft-applied						
n'_c						
2	.808	.660	.686	.185	.825	.687
5	.905	.820	.820	.341	.917	.821

Table 4. Generalizability coefficients by discipline for Teacher within Course Sample.

$n_s'=20$, $n_i'=5$ $(E\hat{\rho}^2_{T,S,I})$

	Organization	Breadth of Coverage	Group Interaction	Enthusiasm	Grading	Individual Rapport
Hard						
n'_t						
2	.570	.840	.291	.400	0.000	.559
5	.768	.917	.480	.613	0.000	.760
Soft-pure						
n'_t						
2	0.000	.752	0.000	.225	0.000	.237
5	0.000	.837	0.000	.419	0.000	.435
Soft-applied						
n'_t						
2	.671	.672	.628	.586	.506	.802
5	.836	.818	.799	.777	.699	.886

Table 5. Generalizability coefficients by discipline for Teacher within Course and Course within Teacher samples.

$n'_s = 20$, $n'_i = 5$

	Teacher within Course (n'_c=5)	Course within Teacher (n'_t=5)
Organization		
Hard-pure	.79	.77
Hard-applied	.83	n/a
Soft-pure	.71	0.00
Soft-applied	.90	.84
Breadth of Coverage		
Hard-pure	.94	.92
Hard-applied	.83	n/a
Soft-pure	.40	.84
Soft-applied	.82	.82
Group Interaction		
Hard-pure	.89	.48
Hard-applied	.92	n/a
Soft-pure	.82	0.00
Soft-applied	.82	.79
Enthusiasm		
Hard-pure	.94	.61
Hard-applied	.92	n/a
Soft-pure	.68	.42
Soft-applied	.34	.78
Grading		
Hard-pure	.89	0.00
Hard-applied	.79	n/a
Soft-pure	.49	0.00
Soft-applied	.92	.70
Individual Rapport		
Hard-pure	.67	.76
Hard-applied	.84	n/a
Soft-pure	.50	.43
Soft-applied	.82	.89