Drawing Inferences about Instructors: The Inter-Class Reliability of Student Ratings of Instruction

Gerald M. Gillmore
February, 2000

OVERVIEW

The question addressed in this report is whether there is sufficient consistency in student ratings of instructors to support the use of data aggregated over classes for personnel decisions. Instructional Assessment System (IAS) data from over 2,800 instructors teaching over 23,000 classes were analyzed. Results showed adequate instructor-level reliability of ratings when aggregating across about seven classes and especially strong instructor-level reliability when aggregating across 15 or more classes. However, these results assume certain conditions of decision-making and are limited to similar conditions of measurement.

INTRODUCTION

One purpose of student ratings of instruction is to make inferences about the quality of an instructor's teaching for administrative reasons, such as merit pay increases, retention, promotion, and especially tenure. In making these kinds of decisions, data are typically considered from all of the courses that were rated within the specified period of time. The question addressed in this report is whether there is sufficient consistency in the ratings of instructors over classes to support the use of aggregated data for personnel decisions. Secondly, assuming a positive answer to the question above, from how many courses should data be collected in order to arrive at a reliable estimate of teaching quality, as perceived by students? To address these questions, we must view the reliability, or consistency, of ratings at the **instructor level**.

The reliability of student ratings data at the **class level** is well-established. In this report, we use the term *class* to refer to a specific section of a course taught by a specific instructor. Kane, Gillmore and Crooks[1] and Lowell and Gillmore[2] have argued elsewhere that inter-rater reliability is the appropriate measure for determining the class-level reliability, where the students are the raters. For inter-rater reliability, universe or true score variance derives from differences among classes, while error variance derives from differences among students' ratings within classes. For these designs, students (or raters) are almost always nested within classes. For items of the University of Washington Instructional Assessment System, reasonable reliability is generally achieved for classes of 15 or more students (see IAS General Description - Reliability).

Class-level reliability is a necessary condition for instructor-level reliability, but it is possible to have adequate class-level reliability without adequate instructor-level reliability. For example, if the course one teaches has a powerful effect on the ratings one receives, high class-level reliability could be accompanied by low instructor-level reliability, assuming an instructor is rated in a number of different

courses. In such a case, the particular mix of courses an instructor taught would have a powerful effect on the aggregated ratings she received, independent of the instructor herself. However, using a generalizability theory framework, Gillmore, Kane and Naccarato[3] found a substantial teacher variance component, leading them to conclude that adequate reliability for discriminating among teachers was achieved by sampling five or more courses per instructor. However, the sample size for their study was relatively small, 42 faculty each teaching two courses. Other studies have correlated the results of pairs of courses taught by the same instructor and have found moderately high positive relationships, thus implying adequate consistency among ratings of different courses.

This study differs from those above by using a very large data set of classes rated at the University of Washington over four years. Medians for select items from all instructors who were rated in five or more classes are analyzed. The ratings of individual students are not included. The decision-making scenario to which this design relates is one in which all class ratings are averaged, irrespective of class size or course designation, and decisions are normatively based relative to the entire campus. The results to be presented below do not necessarily model decisions made normatively at the departmental level, for example. Because the method fundamentally assesses the extent to which class averages on student ratings can discriminate among instructors, we will label our measure as inter-class reliability, analogous to inter-rater reliability at the class level.

METHODOLOGY

The Instructional Assessment System (IAS) database used in this study contains ratings of UW instructors of all academic ranks from Fall Quarter 1995 to Spring Quarter 1999. The entire database contains ratings of 7,102 distinct instructors teaching 36,424 distinct classes. For purposes of this study, the data were limited to those instructors who were rated in at least five classes. The design was not balanced, however, in that all classes were used when more than five were available. This restriction reduced the data set to 2,801 to 2,860 instructors teaching 23,466 to 27,457 classes, depending upon the analysis.

The IAS is comprised of twelve forms designated A to J, and X. Form I is designed to be used in distance education classes and was not included in this study. The remaining forms consist of a set of common items (Items 1 through 4, and 23 through 30) and a set of idiosyncratic items. Only the common items are analyzed here.[4] The text for each of the common items is given in Table 1. In the IAS system, adjusted medians are presented for Items 1 through 4. These items are adjusted for the rating given Item 23 (see Table 1 for text), the log of the class size, and the percentage of students taking the course in their major or as an elective.

Items 1 through 4 use a six point scale: *Excellent*, *Very Good*, *Good*, *Fair*, *Poor*, and *Very Poor*. Items 23 through 27 use a seven-point scale from *Much Higher* to *Much Lower*. Items 28 and 29 use twelve-point scales from *Under 2* to *22 or more*. Item 30 uses a twelve-point scale from *A (3.9-4.0)* to *E (0.0)* (plus *Pass*, *Credit*, and *No Credit*, which are not included in the analyses). For Items 28 and 29, average ratings are divided by course credits for the analyses. The inter-class reliabilities were calculated using intraclass correlations.[5]

A word about the underlying formulation. The total variance of average ratings for a given item can be partitioned into two sources, that which is attributable to classes taught by the same instructor and that which is attributable to differences among instructors when averaged across classes. The latter is measured by the variance across simple, unweighted means computed for each instructor by averaging over the set of class ratings. Within instructor variance is computed for each instructor across the classes in which she was rated, and these variances are added across instructors. Computationally, the method begins with a one-way analysis of variance, with the item under analysis as the dependent variable and the set of instructors as the independent variable. Simply dividing the Mean Square Between minus Mean Square Within by Mean Square Between, or (F-1)/F yields an instructor-level reliability.

The reliability coefficient resulting from this computation corresponds to the average number of classes per instructor in the dataset. In fact, the reliability of any specific number of classes rated can be determined by applying the general form of the Spearman-Brown Prophecy Formula,[6] which is based on the premise that error reduces as a square root of the number of observations, or, in the present application, the number of classes rated.

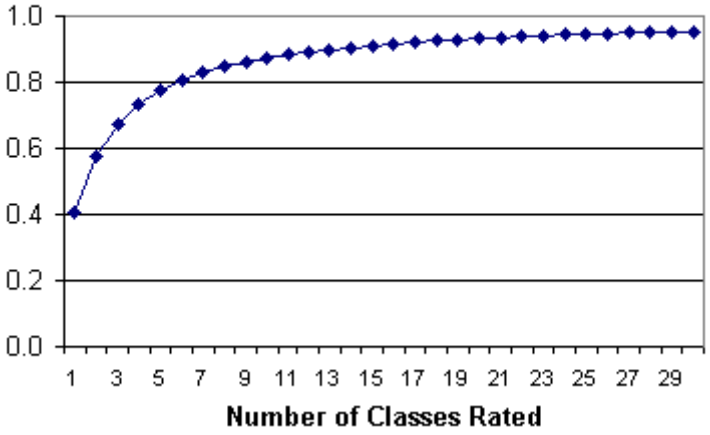
With this formulation, the inter-class reliability coefficients index the extent to which the ratings show both differences across instructors and consistency within instructors. A highly reliable set of measures would be one that highly differentiated among instructors and for which each instructor received very similar ratings for each class rated. In contrast, an unreliable set of measures would result if the differences among instructors were no greater than the differences among classes taught by the same instructor. As is the case for all reliability coefficients, values can range from 0.0 to 1.0, with a value of 0.0 connoting no reliability or consistency, and a value of 1.0 connoting perfect reliability.

RESULTS

Table 2 presents the inter-class reliability coefficients for various numbers of classes for the items under investigation. As described above, the coefficients presented in each row of Table 2 are computed by applying the Spearman-Brown Prophecy Formula to the reliability of the average number of classes. One can see that for any given number of classes the values for Items 1 through 4 and adjusted Items 1 through 4 do not show much variation. The small amount of variability observed for the one-class condition tends to disappear for larger numbers of classes. Generally, in spite of the fact that instructors teach a number of different courses, at different times, and to different students, their resulting ratings show considerable consistency. The reliability is fairly strong for about seven classes and is especially strong for fifteen classes.

In general, the reliability estimate rises as more classes are rated. This statistical relationship is illustrated in Figure 1, which displays the reliability of Item 1 as a function of the number of classes rated. This figure is included for illustrative purposes. Roughly equivalent graphs could be included for other items.

Figure 1. Inter-Class Reliability for Item 1 as a Function of Number of Classes Rated



One might expect Items 23 through 30 to show lower reliability because they appear to be measuring more course-specific attributes, and, in fact, they are generally lower. However, even these items reach reasonable reliability by seven classes. The item with the smallest associated reliability coefficient is *Your involvement in this course was:*. Contrary to expectations, the reliability of Item 30, *Expected Grade*, was comparable to that of the first four items.

DISCUSSION

One might assume that administrative decisions having major impacts on instructional personnel are made deliberately with careful weighting of many factors. Teaching effectiveness is a component of consequence and often student ratings are a major variable in this determination. The data provided above suggest that student ratings can provide reliable information about an instructor's teaching quality, in particular when ratings from seven or more classes are considered. However, the underlying methodology makes several assumptions that should be understood.

First, the model presented in this research is that all classes rated would be weighted equally regardless of size, existence of multiple sections of the same course, and importance to the program. It is unlikely that differential weighting would make a significant difference in the results, but that assumption is untested.

Second, the model assumes that decisions are made relative to the entire campus and relative to instructors at all ranks. Appropriate reliability calculations may yield smaller coefficients for decisions made relative to a particular academic unit or instructor rank because there may be less variability among the instructors. There is less reason to expect the other major source of variation, among classes within

instructors variance, to change significantly. Moreover, while the statistical adjustments lessen academic unit differences, the reliability of the adjusted items is essentially equal to the unadjusted items. This result suggests that the results may be quite general, but academic units would have to conduct specific analyses to assess this issue.

Third, we assume that the classes analyzed represent a reasonable and generalizable set. In other words, we analyzed all of the classes that were rated, and no attempt was made to select classes meeting certain criteria. Thus, the conclusion that the course taught does not appear to have a major effect was based on an analysis of a particular mix of courses. If course assignments were to radically change across the University, these results might not be applicable. In any case, in predicting future teaching ratings, one is safest in assuming that the given instructor will teach roughly the same mixture of sections of the same course and of different courses.

Finally, these data derive totally from the UW and may not fully generalize to other campuses.

Within these limitations, the results are generally encouraging for student ratings use in administrative decisions. The major lesson from analyses of all items is that student ratings of a given instructor are not overly dependent on specific classes but show consistency over classes.

Turning to Items 23 through 30, an additional tale is told. It appears that instructors can be reliably differentiated on the basis of their students' perception of their grade, the intellectual challenge, how much effort they put into the classes and how much effort is required to succeed, the extent of their involvement in the course, how many hours they work per credit on the course, and how many of these hours were worthwhile. These items are less direct measures of teaching quality and serve more to help instructors understand course strengths and weaknesses and for others to interpret course ratings. It was not altogether expected that instructors would get reasonably consistent ratings on these items over classes. To the contrary, it appears that some instructors are viewed as more demanding in all of their classes than others.

The two grading items are especially interesting. It appears that some instructors are consistent in giving higher grades than are others, at least according to student perceptions when they are completing the forms. There are several competing explanations for this consistency. Perhaps, the students of some instructors consistently learn more regardless of the course and, hence, expect higher grades. One might think that the consistency of Item 28, total hours per credit, suggests that some faculty consistently work students harder than others and through this work attain more learning. However, the correlation between the grading items and Item 28 is negligible and thus the Item 28 result adds no credence to this explanation. Alternatively, perhaps some instructors are naturally more lenient in grading than others.[7] It is this possibility, independent of the current study, that led to the adjusted ratings. Finally, some classes may be offered within departmental cultures that grade more leniently, whereas other classes are offered in departmental cultures that tend to grade more harshly. Since students take the majority of their classes within the department of their major, this interpretation is consistent with students' rating of their expected grades. It also predicts the higher reliability for expected grade than relative grade.

TABLES

Table 1. Item Text

1.	*The course as a whole was:*
2.	*The course content was:*
3.	*The instructor's contribution to the course was:*
4.	*The instructor's effectiveness in teaching the subject matter was:*
Relative to other college courses you have taken:	
23.	*Do you expect your grade in this course to be:*
24.	*The intellectual challenge was:*
25.	*The amount of effort you put into this course was:*
26.	*The amount of effort to succeed in this course was:*
27.	*Your involvement in this course was:*
28.	*On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, writing papers, and any other course related work?*
29.	*From the total average hours above, how many do you consider valuable in advancing your education?*
30.	*What grade do you expect in this course?*

return to text

Table 2. Inter-Instructor Reliability Coefficients for Various Numbers of Classes Taught

# Classes	1	2	3	5	7	10	15	20	30
Item 1	0.41	0.58	0.67	0.77	0.83	0.87	0.91	0.93	0.95
Item 2	0.40	0.57	0.66	0.77	0.82	0.87	0.91	0.93	0.95
Item 3	0.43	0.60	0.69	0.79	0.84	0.88	0.92	0.94	0.96
Item 4	0.44	0.61	0.71	0.80	0.85	0.89	0.92	0.94	0.96
Adjusted Item 1	0.40	0.57	0.67	0.77	0.82	0.87	0.91	0.93	0.95
Adjusted Item 2	0.39	0.56	0.66	0.76	0.82	0.87	0.91	0.93	0.95
Adjusted Item 3	0.43	0.61	0.70	0.79	0.84	0.88	0.92	0.94	0.96
Adjusted Item 4	0.44	0.61	0.70	0.80	0.85	0.89	0.92	0.94	0.96
Item 23	0.33	0.49	0.59	0.71	0.77	0.83	0.88	0.91	0.94
Item 24	0.35	0.52	0.62	0.73	0.79	0.84	0.89	0.91	0.94
Item 25	0.33	0.50	0.60	0.71	0.78	0.83	0.88	0.91	0.94
Item 26	0.33	0.50	0.60	0.72	0.78	0.83	0.88	0.91	0.94
Item 27	0.27	0.42	0.52	0.65	0.72	0.79	0.85	0.88	0.92
Item 28/Credit	0.38	0.55	0.65	0.75	0.81	0.86	0.90	0.92	0.95
Item 29/Credit	0.34	0.51	0.61	0.72	0.79	0.84	0.89	0.91	0.94
Item 30	0.41	0.59	0.68	0.78	0.83	0.88	0.91	0.93	0.95

return to text

[1] Kane, M. T., Gillmore, G. M., and Crooks. T. J. (1976) Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, *13*, 171-184. return to text

[2] Lowell, N. and Gillmore, G. M. (1991) Reliability of the Items of the Instructional Assessment System: Forms A-G. *OEA Reports*, *91-1*. return to text

[3] 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*, *14*, 1-21. return to text

[4] Items 2 to 4 on Form G are idiosyncratic, and were excluded from these analyses. return to text

[5] Ebel, R. L. (1951) Estimation of the reliability of ratings. *Psychometrika*, *16*, 407-424. return to text

[6] The general form of this formula is rk = (kr1)/[1 + (k-1)r1], where r1 is the reliability of one observation and rk is the reliability of k observations. return to text

[7] Greenwald, A. G.and Gillmore, G. M. (1997) Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *53*, 1209-1217. return to text