OFFICE OF EDUCATIONAL ASSESSMENT
Improving Learning Through Assessment

OEA Report 02-05

Drawing Inferences about Instructors: Constructing Confidence Intervals for Student Ratings of Instruction

Debbie E. McGhee
February, 2002

OVERVIEW

This report expands upon an earlier discussion of instructor-level reliability of course ratings. Gillmore (2000) previously demonstrated that adequate instructor-level reliability may be obtained when ratings are aggregated across at least seven classes. What was left unexamined, however, was the *precision* with which one should regard mean ratings. This brief report presents confidence intervals for true scores based on Instructional Assessment System (IAS) data from approximately 4,000 instructors.

INTRODUCTION

Decision makers partially base retention, promotion, and merit raises for faculty on average ratings from student evaluations. These decision makers need to know if the difference between two average ratings is meaningful. As discussed in Gillmore (2000)[1], an important aspect of student ratings of instruction is their reliability at the instructor level. That is, how confident can one be in the stability of ratings aggregated across classes? In order to compare ratings between instructors, however, information is needed beyond the reliability coefficient. Decision makers also need to gauge the precision of mean scores, and this requires estimating *true scores and random error*.

Following Allen and Yen (1979)[2], confidence intervals for true scores may be constructed as long as one has three pieces of information: the mean observed score, the standard deviation of observed scores, and the reliability coefficient. The standard deviation and the reliability coefficient are used to compute the *standard error of measurement*, which is the standard deviation of obtained scores around a hypothetical true score or, more plainly, the amount an observed score can be expected to deviate from the true score.

The present study utilized six years of University of Washington course ratings data. As in Gillmore (2000), medians for selected items from all instructors who were rated in five or more classes were analyzed. This study expands upon the earlier one by presenting information about measurement error in the common IAS items, especially as it relates to inter-class reliability. The goal of this paper is to aid decision makers in answering questions such as, "Is this instructor's 10-class average rating of 3.1 honestly different from another instructor's 8-class rating of 3.4?"

METHODOLOGY

The Instructional Assessment System (IAS) database used in this study contained ratings of UW instructors of all academic ranks from Fall Quarter 1995 to Spring Quarter 2001. The database contained median ratings of 10,478 individual instructors teaching 56,278 distinct classes (i.e., course sections). Data were limited to those instructors who were rated in at least five classes. These parameters reduced the data set to 3,592 - 4,005 instructors rated in 43,129 - 44,408 classes, depending upon the item under analysis. Note that data from all classes for a particular instructor were used even when more than five were available.

Readers are encouraged to review Gillmore (2000) for more extensive background on methodology, including a description of IAS items and the computation of reliability coefficients. Briefly, there are ten items that are common to all IAS course evaluation forms, and these were the items analyzed. At the class level, instructors receive median ratings; thus, inter-class reliability, in contrast to inter-rater reliability, nests students within classes and does not use individual student ratings. Instructor-level reliability was computed by conducting, for each item, a one-way analysis of variance with *item* as the dependent variable and *instructor* as the independent variable and then computing the value of (F-1)/F; the result is *r*, the inter-class reliability coefficient. The Spearman-Brown Prophecy formula[3] was then applied to each reliability coefficient in order to estimate reliabilities for various numbers of classes rated.

The inter-class reliability coefficient indexes the extent to which ratings show both consistency within instructors and differences between instructors. A reliable measure would be one on which each individual instructor received very similar ratings for all classes and that also differentiated among instructors. In contrast, an unreliable measure would result if the differences among instructors were no greater than the differences among classes taught by the same instructor. Values of reliability coefficients can range from *r* = 0.0 to 1.0, with a value of 0.0 denoting no reliability or consistency, and a value of 1.0 denoting perfect reliability.

Following the computation of reliability coefficients for each IAS item and for various numbers of classes, the next step was to compute the standard error of measurement (SEM) for each item using the following formula:

$$SEM = SD * \sqrt{(1-r)}$$

RESULTS AND DISCUSSION

Table 1 presents the observed item means and reliabilities, as well as the computed inter-class reliability coefficients and SEMs for various numbers of classes. Notice that the reliability coefficients are very similar (and in many cases identical) to those computed by Gillmore (2000). Secondly, because the SEM is inversely related to reliability, more observations (i.e., classes rated) give a better fix on an instructor's true score.

Table 1 may be used to compare the ratings of two or more instructors. The two methods for comparing scores are: 1) to construct confidence intervals and see whether they overlap and 2) to compute the standard error of the difference (SEdiff) in order to determine the minimum statistically significant

difference.

Constructing confidence intervals

Confidence intervals indicate the range within which an examinee's true score is expected to fall a predetermined proportion of the time if the examinee were rated repeatedly. The prevailing convention is to create confidence intervals at the 95% level, meaning that 95% of the time properly constructed confidence intervals should contain the examinee's true score. The formula for a 95% confidence interval is:

$$T = x \pm (SEM * 1.96),$$

where T is the examinee's true score, x is the observed mean, and 1.96 is the critical value of the standard normal deviate (z) for the desired confidence level of 95%. The formula shows that, under the assumptions of classical test theory, observed scores are composed of the true score plus a margin of error.

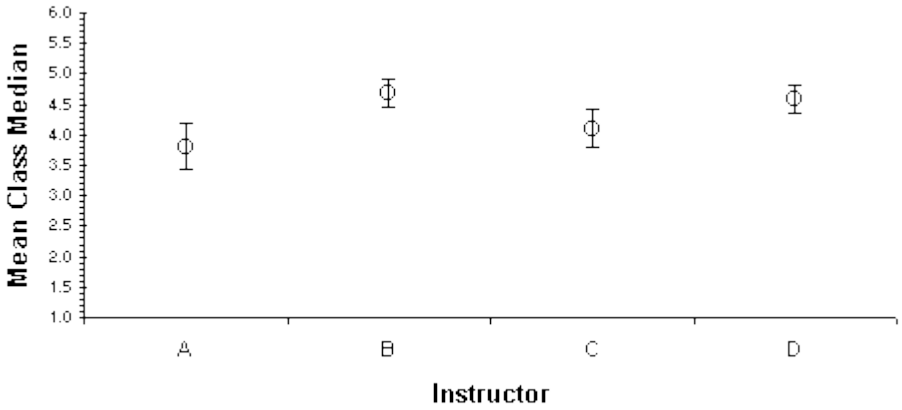
Suppose Instructor A has been rated in 10 classes and has received a mean class median on Item 32 of 3.8. Table 1 indicates that the SEM associated with 10 classes is .19. Meanwhile, Instructor B has been rated in 30 classes and has a mean class median of 4.7; the SEM for Item 32 and 30 classes is .12. As shown in Table 2 and Figure 1, because the two confidence intervals do not overlap, we can assert with 95% confidence that the true scores of Instructors A and B are significantly different. By contrast, Instructor C has received a mean rating over 15 classes of 4.1, resulting in a 95% confidence interval of 3.8 - 4.4. We cannot reject the hypothesis that the true median ratings of Instructor A and Instructor C are equal.

Table 2. Item 32 confidence intervals for four hypothetical instructors

Instructor	Item 32 Mean	Number of Classes	SEM	Margin of Error	95% CI Bounds	
					Lower	*Upper*
A	3.8	10	.19	.38	3.4	4.2
B	4.7	30	.12	.23	4.5	4.9
C	4.1	15	.16	.32	3.8	4.4
D	4.6	30	.12	.23	4.4	4.8

Note. The margin of error is computed as SEM*1.96.

Figure 1. Mean Class Medians with 95% Confidence Intervals for Four Hypothetical



Instructors

Using the standard error of the difference

When confidence intervals do not overlap, we can comfortably assert that two (or more) scores are significantly different from one another. The converse is not true. For example, there are times when intervals do overlap slightly but the means are, nevertheless, significantly different from one another (Estes, 1997; Tyron, 2001)[4] . The standard error of the difference (SEdiff) provides another way to compare the scores of multiple examinees. SEdiff may be computed from the SEM as below:

$$SEdiff = \sqrt{(SEM1^2 + SEM2^2)},$$

where SEM1 and SEM2 are the standard errors associated with the first and second observation, respectively.

In order to determine the *minimum significant difference (MSD)*, multiply the obtained SEdiff by the critical value of *z* at the desired significance level. Again, the *z* at *p*=.05 is 1.96.

Table 3. SEdiffs for comparisons among four hypothetical instructors

Instructors	Number of Classes	SEM1	SEM2	SEdiff	MSD
A vs. B/D	10 v. 30	.19	.12	.23	.45
A vs. C	10 v. 15	.19	.16	.25	.50
C vs. B/D	15 v. 30	.16	.12	.20	.39

Note. MSD is the minimum significant difference and is computed as SEDiff*1.96.

A careful comparison of Table 2 with Table 3 would show that, in all cases save one, the confidence interval and SEdiff methods would lead to the same conclusion about whether or not pairs of mean ratings are significantly different. In the case of Instructor C vs. Instructor D, however, the intervals overlap but the difference between the means exceeds the MSD. Using the SEdiff as an index is generally more accurate because it takes into account the error estimates for both cases conjointly.

CONCLUSION

Constructing confidence intervals for the mean of class median ratings is helpful in determining whether there are true differences between instructors. The confidence interval method is an accepted practice, and it provides a consistent and straightforward way for users of student evaluation data to make decisions based on those data. Often, mean ratings may appear to be different, but once the number of classes, inter-class reliability, and SEM are taken into account, another picture may emerge. Our UW data indicate that, even under the best of circumstances (i.e., having ratings for 30 or more classes) a difference of at least .3 points is necessary before one can presume that the mean ratings of two instructors may be honestly different from one another. When confidence intervals overlap only slightly, users may also wish to calculate the standard error of the difference.

---

[1] Gillmore, G. M. (2000). Drawing Inferences about Instructors: The Inter-Class Reliability of Student Ratings of Instruction. *OEA Reports 00-2*.

[2] Allen, M.J., Yen, W.M. (1979). Introduction to Measurement Theory. Brooks/Coles Publishing Company. Monterey, CA.

[3] 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.

[4] Estes, W.K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin and Review, 4*, 330-341.
Tyron, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.