Educational Research and Review Vol. 1 (8), pp. 267-271, November 2006
Available online at http://www.academicjournals.org/ERR

Full Length Research Paper

Improving Data Driven Decision Making through Assessment Literacy for Respondents

Naomi Jeffery Petersen, Randall S. Davies and Bruce Spitzer

Indiana University South Bend, South Bend, Indiana, U.S.A.

Accepted 17 October, 2006

This study examined the use of a college course evaluation instrument in an effort to better understand and improve the assessment data derived from the instrument. The purpose of the analysis was to examine the dimensionality and reliability of the instrument but, more importantly, to understand what the data really tells us and whether assessment literacy would improve data usability. Analysis of the results suggests that the instrument tended to produce internally consistent data. However, the instrument measured predominately one aspect of course quality. In addition, based on results of the assessment literacy exercise, respondents seem to use very different criterion for rating; the constructs being measured generally did not match what was intended; and the questions being asked did not always match the scale being used. As a result the usefulness of any data interpretation and subsequent decisions was deemed suspect. The results of this study also suggest that simple assessment literacy interventions by themselves do not seem to drastically change the ability of raters to score items reliably. A much more comprehensive effort would be needed to produce results that would be beneficial for long term, data-driven decision making.

Key Words: Assessment Literacy, Course Evaluation, Data-Driven Decisions.

Introduction

The use of student surveys to evaluate university instruction has become a familiar part of the college experience for many students. Most colleges and universities administer course evaluations at the end of each semester. Often this is done as a requirement for accreditation; the information gathered from these evaluations is intended to provide feedback to instructors, and helps inform decisions about tenure and promotion. Few would dispute the importance of students' satisfaction with their learning experience (Chiu, Stewart, and Ehlert, 2003), and a large body of research evidence suggests that such measures are reasonably reliable and informative (Marsh, 1987; Overall and Marsh, 1980; Marsh and Dunkin, 1992; Theall, 2002); however, not all agree (Hake, 2005). The information obtained from a course evaluation is only as good as the instrument's ability to provide quality data. Those using the instrument must be aware of what the instrument actually measures so the decisions they make using these data are justified.

There is no single course evaluation instrument that could or should be used by all schools and institutions; each assessment instrument must reflect the intended purpose for which it was designed. Instruments are considered reliable when the data they provide are a consistent estimate of what the instrument actually measures, intended or otherwise (Linn and Miller, 2005; Nunnally, 1978). Unfortunately, many course evaluation instruments are not validated or have no documentation regarding the instruments development. The course evaluation instrument used in this study seems to be one of those; yet, it has been used exclusively for several years and the instrument is believed to be a satisfactory

*Corresponding author. E-mail: radavies@indiana.edu

tool for gathering relevant course evaluation data. Thus the question posed in this study concerns the validity of the instrument and its ability to produce meaningful feedback about a course and its instructor. More specifically, what exactly does the instrument measure? In addition, this study attempted to determine how university students define quality instruction, and to test the hypothesis that assessment literacy will decrease group response variance thereby increasing the reliability and usefulness of the group data when students evaluate the instruction they receive.

Method

The first stage of the investigation involved administering the course evaluation instruments and conducting factor analysis. The second stage involved an assessment literacy exercise asking students to consider the questions in each construct (or subscale) and provide a descriptive label for each before completing the course evaluation a second time.

Participants

Five classes in a university School of Education (SOE) setting were recruited, resulting in a convenient sample of 92 student participants. Sixteen (16) cases of missing data were eliminated listwise, thus reducing the final sample to 75. The sample size was deemed adequate for the purpose of quantifying and interpreting a group of students representing the population that would be expected to use this particular instrument, but insufficient for establishing value beyond its local use.

Course Evaluation Instrument

The existing course evaluation form was used, altered only by using an online format that permitted the easiest transition to the companion exercise. The existing survey included a few demographic questions which were not analyzed due to the size of comparative demographic groups. Another 19 items used to assess student opinion of various aspects of the course. Participants responded to items using the existing 4-point Likert scale (1 = strongly agree; 2=agree; 3 = disagree; 4 = strongly disagree). The instrument was administered before and after an assessment literacy exercise designed by the researchers and structured according to dimensionality revealed by an exploratory factor analysis of the pre-survey.

Assessment Literacy Exercise

An assessment literacy exercise was developed based on the 3 most prominent components revealed by the preliminary factor analysis. Participants were asked to consider the questions in each subscale. For each of the three sets of items (i.e., constructs), participants chose one item which, in their opinion, was 'most influential.' Students were then guided through a rubric development exercise of describing what factors they consider when deciding the degree to which they agreed with the item. Participants also identified the item they found least important in each group of questions. Students were asked to explain how they rated the item they felt was most important in each construct. Open-ended responses allowed students to label the constructs being addressed. The post survey included the same 19 items as the pre-survey.

Findings

Reliability and Statistical Assumptions

Chronbach's alpha was calculated to determine the internal consistency of the sample responses. The result suggested a high level of internal consistency (α = .95). The sample data was also considered adequate, with a Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) exceeding the required value of .5 to proceed with a factor analysis. The Bartlett's Test of Sphericity was statistically significant (χ= 1148.802; *df* = 19; *pcalc*= .000). Because sample size was adequate (N > 60) and internal consistency was high, an exploratory factor analysis was deemed feasible.

Dimensionality

A principal components (exploratory) factor analysis found four factors with Eigenvalues greater than one, and the skree plot supported the decision to identify three. Following Varimax rotation, the three components explained 26.8%, 17.9%, and 14.7% of the variance, respectively. In addition, at least four items were found for each factor. The fourth component did contribute 11.9% of the rotated sums of squared loadings, but there were only two items for this subscale. Convention requires at least four items for a subscale, thus the fourth subscale was not included in the assessment literacy exercise and the study was based on only the three prominent subscales. Both questions in the fourth subscale asked questions about assessment.

Student Interpretation of Subscale

The instrument used in this study seemed to be measuring primarily three aspects of the course: administrative organization, instructor ability, and the instructor's disposition. By far the majority of items asked (i.e., 10 out of 17) were interpreted as administrative or organizational. Table 1 includes the descriptive labels most commonly given by the students, followed by the number of items associated with each construct. As is typical in factor analyses, the factors become more difficult to summarize, so the second and third factor elicited greater variety of terms (Table 2). Other descriptors used by students for the third factor included: caring, concerned, fairness, dispositions, personality,

Table 1. Students' Subscale Descriptors and Most/Least Valued Items.

Students' Description of Subscale	Most valued item(s)	Least valued item
Course Organization (9 items)	'I learned a lot'	'prompt return of papers'
Instructor's Ability (5 items)	'knowledgeable' 'presents clearly'	'office hours'
Student-Teacher Relationship (4 items)	'enthusiastic about teaching the course'	'cares about us individually'

Table 2. Factor Analysis Subscales

Factor #1 Items	
14.	The instructor is one of the best I have had at the college level
15.	The course is well organized.
16.	The instructor is well-prepared for class meetings.
17.	The objectives of this course are stated clearly.
18.	The stated objectives of this course are consistently pursued.
19.	Assignments are related to the objectives of this course.
20.	I learned a lot from this course.
23.	The basis for assigning grades was clearly explained.
24.	The instructor returns exams, homework, papers, etc. promptly.
Factor #2 Items	
7.	The instructor is regularly available for consultation or is otherwise accessible outside of class.
8.	The instructor recognizes when students are not following him or her in class.
10.	The instructor answers questions carefully and precisely.
11.	The instructor presents material clearly.
13.	The instructor is knowledgeable about the subject.
Factor #3 Items	
5.	The instructor is fair and impartial in dealing with students.
6.	The instructor respects and welcomes student questions and comments about the subject.
9.	The instructor has a genuine interest in students as individuals.
12.	The instructor is enthusiastic about teaching this course.

people skills.

What we cannot conclude from this analysis is which of these subscales was most important to students nor whether there are other important aspects students consider when evaluating a course. However, based on the number of items in each subset, the course organization component is heavily weighted (i.e., assumed important by default) in any aggregate calculation and subsequent analysis of the course evaluation data. This finding suggests the instrument may lack overall construct validity and thus its ability to accurately measure the effectiveness of the overall course is questionable.

Student Perceptions and Values

According to this analysis, students do seem to have a wide range of criteria for evaluating their experience in taking a course. Variance in student responses can be so great at times that one might question whether they shared the same instructional experience. Expectations of what is important vary as well as how important specific aspects of the course are to students. Students appear to associate course organization with learning a lot in the course and with the instructor being a good teacher. They seemed to associate part of their satisfaction with the instructor's content knowledge and ability to present material clearly. Students also seem to value the instructor's enthusiasm for the subject.

Pre- and Post Survey Analysis

Based on the results of the pre- and post survey data, the assessment literacy intervention did not produce a different overall result in the outcomes (i.e., none of the item outcome comparisons were statistically significant). The correlation between individual pre- and post survey item results was typically between .40 and .50 suggesting the individual respondents tended not to change their ratings drastically. Noted differences in the results often depended on the specific questions. Questions that lent themselves to a dichotomous response tended to

produce an increase in the number of raters responding "strongly agree" rather than just "agree" (e.g., the basis for assigning grades was clearly explained). Items that seemed to be addressing a construct that measured a state rather than trait (or simple event occurrence) tended to fluctuate more erratically, often slightly downward (e.g., the instructor is regularly available for consultation) possibly indicating a time sensitive response (i.e., temporary or momentary changes in the state of the construct).

When asked if the intervention changed the way they rated items on the post survey, 49 students responded; 21 indicated that they had changed the way they answered survey questions as a result of the assessment literacy intervention, 38 indicated they did not change the way they responded; 4 were unsure. Several of the respondents indicated they became more critical in their assessment (i.e., they responded less favorably as a result of the assessment literacy intervention).

Overall, there is no evidence to suggest any major change in aggregate group response from the pre- and post-survey results; this may also have been the result of the scale used rather than a true indication of student opinion. This isolated exercise does not inform our understanding of students' competence to use this or any other instrument to evaluate their course experience, but it certainly confirms that few students are proficient in the analytical skills required to articulate a graduated range of behavior and disposition. It also suggests that a simple assessment literacy effort may not be particularly effective at training students to consistently rate a course.

Discussion

A Research Rigor Wake-Up Call

This entire study highlighted the technical problems related to established research and assessment procedures. The instrument likely generates flawed data due to the following violations:

a) Ambiguity of the items related to the response ptions;
b) Response options limited to 4 degrees of agreement (i.e., the use of a Likert scale);
c) Unbalanced number of items for each subscale;
d) Greatly variable conditions for collecting data;
e) Small sample sizes (i.e., low return rate); and
f) Methods of calculating aggregate data and thus the consequential validity of the results.

The researchers recommend the development of a less flawed instrument to collect more meaningful data that can then be used for reflection and evaluation. The technical problems for collecting valid data beg the question of what the data is intended to measure. The group recommends aligning the instrument with SOE goals and conducting a self-study of the new instrument's validity and dimensionality. This might increase the value placed on the data by instructors who question its validity. Companion recommendations include revision of the survey and refinement of the way it is administered. It is also noted that because this is a school of education and thus immersed in accountability reforms, it is important to model the data-driven decision-making and reflective practices that are not obviously flawed.

While the aggregate responses give a consistent indication of opinion, without knowing what students are responding to and what they are saying is important renders the data too weak to interpret. Training student raters is an important aspect of any course evaluation assessment strategy. The results of this study suggest that simple, one-time efforts to improve assessment literacy are not likely to produce more accurate assessments of instructional quality. Simple assessment literacy interventions by themselves do not seem to drastically change the ability of raters to score items reliably. A much more comprehensive effort would be needed to produce meaningful results, and this would include an orientation for the faculty and administrators who must interpret the results and make powerful decisions based on them.

Summary

Because the instrument currently being used has a long but vague history with no empirical or theoretical support but is nonetheless used to make decisions of great consequence to the faculty and programs, the reseachers decided to investigate its validity. Statistical analysis of student responses produced reliable (i.e., internally consistent) data, with several stable underlying constructs. However, about half the items tended to address organizational or administrative aspects of teaching. Students appeared to associate a well organized course with having 'learned a lot', and the instructor's content knowledge with the ability to present material clearly. The data however was considered flawed due to the following violations: a) Ambiguity of the items related to the response options; b) Response options limited to 4 degrees of agreement; c) Unbalanced number of items for each subscale; d) Greatly variable conditions for collecting data; e) Small sample sizes and return rate issues; and f) Methods of calculating aggregate data and thus the consequential validity of the results.

Training student raters is an important aspect of any course evaluation assessment strategy. The results of this study suggest that simple efforts to provide assessment literacy will like not produce desirable results. While the aggregate responses of students gives a consistent indication of opinion knowing what students are responding to and what they are saying is important.

Simple assessment literacy interventions by themselves do not seem to drastically change the ability of raters to score items reliably. A much more comprehensive effort would be needed to produce results that would be beneficial.

This analysis supports the contention that a call for educators to make data-driven decisions must include an expectation that we validate and understand the data collection instruments we use. There is no single course evaluation instrument that could or should be used by all schools and institutions. Each assessment instrument must reflect the intended purpose for which it was designed. It is however incumbent on those using the information gathered from these instruments to understand what the instrument measures if decisions made based on these data are to be meaningful. Thus the widespread use of student evaluations must be reviewed according to conventions of survey research.

References

Chiu K, Stewart B, Ehlert M (2003). The validation of a measurement instrument: Relationships among student demographic characteristics, student academic achievement, student satisfaction, and online business-course quality factors. Paper presented at the 2003 North America Web-Based Conference on Teaching and Learning. Retrieved July 10, 2005 from http://naweb.unb.ca/proceedings/2003/PaperChiu.html

Hake RR (2005). Re: Measuring teaching performance. Retrieved July 10, 2006, from http://lists.asu.edu/cgi-bin/wa?A2=ind0505andL=aera-landT=0andO=DandP=660

Linn R, Miller DM (2005). Measurement and assessment in teaching (9th ed.). Saddle River, NJ: Prentice Hall.

Marsh HW (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. Int. J. Edu. Res. 11. 253-388.

Marsh HW, Dunkin MJ (1992). Students' evaluations of university teaching: A multidimensional approach. In: JC Smart (Ed.), Higher education: Handbook of theory and research, Vol. 8, (pp. 143-233). New York: Agathon.

Nunnally J (1978). Psychometric Theory. NY: McGraw Hill.

Overall JU, Marsh HW (1980). Students' evaluations of instruction: A longitudinal study of their stability. J. Edu. Psychol. 72. 321-325.

Theall M (2002). Student ratings: Myths vs. research. Focus on Faculty, 10 (3), 2-3. Brigham Young University Faculty Center.