A Strategy for Addressing the Validity of a Teacher Effectiveness Instrument

Dale Shaw
University of Northern Colorado

Suzanne Young
University of Wyoming

Jay Schaffer
University of Northern Colorado

Daniel Mundfrom
University of Northern Colorado

This paper deals with the validation of an objective teacher effectiveness instrument for rating the classroom effectiveness of college and university teachers. It includes a description of how the instrument was developed and the process by which validity evidence for the instrument was generated and analyzed via regression and factor analyses.

The purpose of this study was to create a research-based teacher evaluation instrument and analyze data gathered with it to obtain validity evidence for its use as a measure of college and university teacher effectiveness. Institutions use such instruments to collect student ratings of teachers for one or more of the following purposes: (1) to provide teachers with feedback for improving their teaching, (2) to provide students with information they may use to select future courses and instructors, and (3) to provide administrators and faculty with a measure of a teacher's effectiveness that may inform their decisions about a faculty member's tenure, promotion, or retention (Marsh & Dunkin, 1992). Typically the validation of such an instrument requires several studies dealing with at least two aspects of validity: (1) to determine the degree to which obtained ratings reflect the true feelings of students, which is important for the first and second purposes above, and (2) to establish the degree to which the items collectively capture, or truly do measure, the construct of teacher effectiveness, which is important to the third purpose. The focus of this paper is on those aspects of validity that surround the instrument's use for the third purpose, that is, to provide a measure of teacher effectiveness.

This paper consists of an explanation of how the instrument was developed and a description of how data were collected and analyzed for validity evidence. First, items were developed that met two or more of the following three criteria: (1) the item is used prevalently in other teacher rating instruments, (2) the item bears a high relationship to the global construct of teacher effectiveness as evidenced in previous research, or (3) it is a key item in a previously developed teacher effectiveness model (i.e., McKay, 1997). In her model of teacher effectiveness, McKay argues that the three most important items to include in a teacher rating instrument are subject matter knowledge, teacher enthusiasm, and communication skills. Second, data were gathered about college and university teachers from former students in an effort to acquire data about teachers from the entire spectrum of teacher effectiveness. Third, these data were submitted to regression and factor analyses. Evidence of the instrument's construct validity could be indicated in several ways, including: (1) high multiple correlation coefficients between a global score and the collection of items or subsets of the items suggested by previous teacher effectiveness models (i.e., McKay, 1997), (2) high factor loadings in the first extracted principle component suggesting that the items provide a common measure of a unitary construct, (3) obtaining a meaningful factor structure consistent with the work of other teacher effectiveness researchers (Marsh, 1991; Marsh & Hocevar, 1984, 1991; Abrami, d'Apollonia, & Rosenfield,1997).

Instrument Development

One hundred twenty-five different items were gleaned from objective teacher effectiveness instruments described in research studies published since 1985. Only items that were demonstrated to be correlates of teacher effectiveness in the studies wherein students provided ratings of teachers were selected for our study. In all we found 44 studies that identified items that were teacher effectiveness correlates. This pool of one hundred twenty-five items was analyzed for duplicates and near-duplicates, and was edited to achieve a uniformity of presentation in style and format. Twenty-five items from this pool were retained for further consideration. We relied heavily on the works of Feldman (1976, 1984, & 1986), Murray (1980), Erdle, Murray, & Rushton (1985) and Marsh (1987) as we sought to assess the adequacy of the twenty-five items to collectively capture the construct of teacher effectiveness. The twenty-five items include all but two of the nineteen instructional rating dimensions that Feldman (1976)

identified in his classic teacher effectiveness review study as well as two additional items recommended by Murray (1980). These items are teacher's interest in the course, enthusiasm, subject matter knowledge, breadth of subject coverage, preparation and organization, presentation skills, speaking skills, sensitivity to student achievement, clarity of objectives, value of the course, value of supplementary materials, classroom management, course difficulty including appropriateness of workload, fairness, value and frequency of feedback, openness, encouragement and challenge, respect and friendliness, availability, clear explanations and encouragement of student participation.

A pilot study of the twenty-five-item instrument led us to conclude that, at twenty-five items, the instrument was much too long to be practical. An eleven-item version was developed from the twenty-five-item version by selecting in large part those items that bore the highest relationships with teacher effectiveness while still covering the spectrum of issues captured in the original item pool. The eleven items are presented in Table 1. In the form for administering the items, a 9-interval rating scale from 1 to 9 with anchors 1 (Very Low), 3 (Low), 5 (Average), 7 (High), and 9 (Very High) followed the presentation of each item.

Data for addressing the validity of the instrument were obtained from students in 22 undergraduate and graduate classes who were asked to rate three professors of their choice from whom they had taken a course in the recent past. The students were given a brief training regarding halo effect and leniency effect in ratings and admonished to not succumb to these rater errors as they filled out the instrument. They were also asked to select professors to rate from a variety of points along the teacher effectiveness continuum to the extent that it was possible for them to do so. In a cover sheet, the students were given written instructions regarding the study and an overall or global rating item to be filled out for each instructor that they planned to rate on subsequent rating forms. The global item, that served as the criterion variable in the regression analyses below, was worded "Everything considered, I would rate the instructor's effectiveness" and was rated on the same 1 to 9 scale as the 11 items. In all, 1082 useable cases were obtained from 384 students. These data was submitted to regression and factor analyses in an effort to acquire evidence of the 11-item instrument's validity to measure college and university teacher effectiveness.

Regression analysis

Table 2 presents information about 4 regression models. The first model is the complete model derived from the data collected in this study by regressing the global score onto all eleven items. An R^2 of 0.8918 was obtained for this model indicating that 89% of the variance in the global scores is accounted

Table 1. Instrument Items

Item Name	Actual Wording on the Instrument
1. Subject matter knowledge	The instructor's subject matter knowledge
2. Communication skills	The effectiveness of the instructor's communication skills
3. Enthusiasm	The instructor's enthusiasm for teaching
4. Comfortable atmosphere	The degree to which the instructor created a comfortable learning atmosphere
5. Respectful of students	The degree to which the instructor was respectful of students
6. Warm and friendly	The instructor's warmth and friendliness
7. Motivate & stimulate	The degree to which the instructor was motivating and stimulating
8. Concern for learning	The instructor's genuine concern for student learning
9. Increased interest	The degree to which the course increased my interest in the subject
10. Increased understanding	The degree to which the course increased my understanding of concepts
11. Course organization	The degree to which the course was well organized
Global Item	Everything considered, I would rate the instructor's effectiveness

Table 2. Regression Models

Item	Complete	Feldman	Young/Shaw	McKay
Subject matter knowledge	X*	X		X
Communication skills	X	X	X	X
Enthusiasm	X	X		X
Comfortable atmosphere	X			
Respectful of students	X	X		
Warmth and friendliness	X	X		
Motivate and stimulate	X	X	X	
Concern for learning	X	X	X	
Increased interest	X			
Increased understanding	X		X	
Course organization	X	X	X	
R^2	0.8918	0.8659	0.8788	0.7877

* An X indicates that the item is included in the model.

for by this eleven-item instrument. The multiple correlation coefficient for the global score and the best linear combination of the 11 items is 0.9444 indicating that the global score and the eleven-item instrument score bear a very high relationship to each other. Considering the criteria used to select items for inclusion in the instrument, this is compelling validity evidence. The 11-item instrument does indeed capture the construct of overall teacher rating extremely well.

Additional validity evidence is provided by the Feldman, Young/Shaw, and McKay models presented in Table 2. Of the 11 items in the instrument, eight were among those that Feldman identified as being used prevalently in teacher evaluation instruments at many colleges and universities. To the extent that an item's prevalence of use in other scales serves as a validity criterion for its inclusion in this teacher effectiveness scale, the subset of eight commonly used items identified by Feldman alone accounts for almost 87% ($R^2 = 0.8659$) of the variance in the global ratings. This provides further substantial evidence of the eleven-item instrument's validity. In a like manner, the Young/Shaw and McKay models offer additional validity evidence. These authors have demonstrated that communication skills, instructor enthusiasm, subject matter knowledge, and ability to motivate and stimulate students are among the most important items to include in a teacher effectiveness instrument (Young & Shaw, 1999 and McKay, 1997). The 5-item subset of Young/Shaw and the 3-item subset of McKay account for 88% and 79% of the variance in global scores, respectively. Regarding the validity of the eleven-item instrument developed in this study, validity is evident in that the instrument contains subsets of items, known to have validity as measures of teacher effectiveness in their own right, that bear high relationships to the global score.

Factor analysis

Factor analysis was used to extract the first principal component from the data in an effort to ascertain the degree to which the eleven-item instrument captures a single, unitary construct. The results are presented in the first column in Table 3. With the single exception of subject matter knowledge that had a moderate loading, loadings are high to very high providing substantial evidence that the eleven-item instrument is indeed capturing a unitary construct of teacher effectiveness. The items were also factored to determine whether the unitary dimension would sub-divide into two or more factors. A five-factor solution, with well-identified factors that is easily interpreted, is presented in Table 3. The single dimension of teacher effectiveness in this study subdivides into 5 factors: instructor's subject knowledge; course organization; instructor communication skills, enthusiasm and ability to motivate; increased student interest and understanding; and instructor's general regard for, and treatment of, students. This sub-division of the overall dimension of teacher effectiveness into two or more (in this case, five) factors closely matches factor structures reported by other teacher effectiveness researchers (Marsh ,1991; Marsh and Hocevar, 1984 and 1991; Abrami, d'Apollonia, & Rosenfield,1997).

Table 3. Factor Analysis Results

Item	First Principal Component	Rotated Five Factor Orthogonal Solution				
		F1	F2	F3	F4	F5
Subject matter knowledge	.553					.937*
Course organization	.726				.847	
Communication skills	.861			.606		
Motivate and stimulate	.891			.617		
Enthusiasm	.835			.736		
Increased interest	.820		.831			
Increased understanding	.815		.819			
Comfortable atmosphere	.868	.724				
Respectful of students	.834	.866				
Warmth and friendliness	.796	.860				
Concern for learning	.870	.647				

* Loadings less than .500 are not reported.

Results and Discussion

Our findings consist of the following two statements: 1) the 11 items capture 89% of the variation in overall teacher ratings indicating that the instrument does indeed capture a very large portion of the variation in teacher ability, and 2) the 11 items have high loadings on a single factor indicating the extent to which the instrument is indeed unidimensional, however, the items do subdivide as expected into five, easily interpreted sub-factors, some of which deal more with the instructor and the others more with course-related matters. These findings provide substantial validity evidence for the eleven-item instrument. In general, the evidence is compelling. Our conclusion is that the instrument indeed appears to capture the construct of teacher effectiveness very well.

This work has resulted in the development of a teacher effectiveness instrument to which is attached a substantial body of validity evidence. This instrument may ultimately prove to be a viable teacher-rating instrument for use in a college or university, however, it is important to point out that its intent is to calibrate teacher effectiveness as a global construct. It may or may not be very useful as a device for providing teachers with itemized student feedback or students with information for their future scheduling. However, of possibly greater value than the creation of a single instrument, is the process by which the instrument was developed and validated. This process may be used again with different or modified item bases or underlying dimensions of teacher effectiveness.

References

Abrami, P., d'Apollonia, S., and Rosenfield, S. (1996). *The dimensionality of student ratings of instruction: What we know and what we do not*. In J. C. Smart (ed.) Higher Education: Handbook of Theory and Research (Vol. 11). New York: Agathon Press.

Erdle, S., Murray, H. & Rushton, J. (1985). Personality, classroom behavior, and college teaching effectiveness: A path analysis. *Journal of Educational Psychology*, *77*, 394-407.

Feldman, K. (1976). The superior college teacher from the student's view. *Research in Higher Education*, *5*, 243-288.

Feldman, K. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, *21*, 45-116.

Feldman, K. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, *24*, 139-213.

Marsh, H. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285-296.

Marsh, H. & Dunkin, M (1992). *Students' evaluations of university teaching: A multidimensional approach*. In J. Smart (ed.) Higher education: Handbook of Theory and Research (Vol. 8). New York: Agathon Press.

Marsh, H & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal*, *21*, 341-366.

Marsh, H. & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level and course level. *Teaching and Teacher Education*, *7*, 9-18.

McKay, J. (1997). Examining the elusive essence of the superlative teacher. *Education*, *177*, 3-9.
Murray, H. (1980). *Effective teaching behaviors in the college classroom*. In J. Smart (ed.) Higher education: Handbook of Theory and Research (Vol. 7). New York: Agathon Press.
Young, S. & Shaw, D. (1999). Profiles of effective teaching in higher education. *Journal of Higher Education*, *70*, 670-686.

Send correspondence to: Dale Shaw, Department of Applied Statistics and Reseach Methods
University of Northern Colorado, Greeley, Colorado 80639
Email: dale.shaw@unco.edu