Student Evaluations of Courses and Faculty Based on a Perceived Learning Criterion: Scale Construction, Validation, and Comparison of Results

Richard D. Freedman and Stephen A. Stumpf
New York University

Validation studies of the Course-Faculty Evaluation Instrument (CFI) are described. Seven dimensions were constructed which characterize each class and predict student rating of the instructor, course, and text. Different measurement scales and methods were analyzed, using a multitrait-multimethod (MTMM) strategy. The MTMM matrix for the CFI and a similar MTMM matrix for the Course-Evaluation Instrument (CEI) reported by Schwab (1974) were analyzed and compared. The same method of scaling was found to be superior in both studies. Using an analysis of variance framework to summarize MTMM matrices, the CFI demonstrated greater discriminant validity using more dimensions (traits) and had a lower error component than the CEI. The benefits of comparing instruments and implications for future course-faculty evaluation research are discussed.

Student evaluations of courses and instructors are often being used in academic decision-making (Rodin & Rodin, 1972). Because of their growing use, the validity of student evaluations is of critical importance. Yet Wallace and Schwab (1973) state that there have been "only a few empirical studies . . . reported on the validity of student evaluations" (p. 229). After evaluating 500 such instruments, Sockloff (1973) concluded that few would have been used if they were properly researched.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 2 Spring 1978 pp. 189–202

A Course-Faculty Evaluation Instrument (CFI) was developed to provide (1) information to business students for use in selecting courses and instructors; (2) feedback to instructors for improving courses and teaching; and (3) information to administrators for planning and decision-making purposes. The primary goal of the research was to develop a valid and reliable instrument for the Graduate Business School (GBA) at New York University. However, it was also intended to construct an instrument that could be used, with or without modifications, in other graduate and undergraduate universities.

The importance of student evaluations in higher education is evidenced by the development, validation, and use of a number of instruments (e.g., Aleamoni & Spencer, 1973; Barnoski & Sockloff, 1976; Holmes, 1971; Wallace & Schwab, 1973). These instruments measure students' perceptions of instructors and/or classes along various dimensions. The criteria used to validate such instruments include objective student examinations, student beliefs that the course achieved the stated objective, "consumer satisfaction," student decisions to take further courses in the same area, perceived learning, and long-term usefulness (Costin, Greenough, & Menges, 1971). Since the validation processes and instruments are seldom analytically compared, it is difficult to determine their relative advantages. This retards the de-

velopment of evaluation systems or approaches that could be useful in a variety of institutions.

In addressing this problem, Freedman and Stumpf (1976) developed the Course-Faculty Evaluation Instrument (CFI) following the general framework used by Schwab (1974; Schwab & Wallace, 1975). However, methodological changes were made to improve the instrument's validities. This paper describes validation studies of the CFI, summarizes its measurement characteristics, and compares it to the earlier instrument developed by Schwab.

Validity Studies of the CFI

Instrument Development

The CFI was developed during the spring and summer of 1976. Items were generated by asking over 100 graduate business students for critical incidents, short descriptive phrases, or adjectives about effective or ineffective performance on five dimensions used by Schwab (1974, 1976). The dimensions were Instructor in Class, Subject Matter, Text and/or other Required Reading, Graded Assignments and Examinations, and Instructor in General. Sixteen or more nonredundant items were generated for each dimension for a total of 96 items. Sample items for the Instructor in Class dimension are: clear, thought provoking, stimulating; sample items for the other dimensions are presented in Table 1 (for additional items see Freedman & Stumpf, 1976).

Construct validity evidence for the CFI was defined as the instrument's ability to differentiate courses and faculty in relation to a criterion of teaching effectiveness. Since perceived student learning is a criterion of effective teaching that has considerable acceptance (McKeachie, Lin, & Mann, 1971; Schwab, 1976), it was chosen as the criterion for developing the CFI. Student perceptions of the extent to which a course and instructor met their learning goals was used to assess item validity (Harari & Zedeck, 1973; Schwab, 1976; Sharon, 1970).

The learning criterion was developed by analyzing student perceptions of the extent to which each item described a "high" or "low" learning experience. The response scale used was "Y" for Yes, "?" for Undecided, and "N" for No. Items which could distinguish between "High" and "Low" learning experiences were retained for further analysis. This was determined by calculating a point-biserial correlation between two independent samples. One sample was instructed to complete the questionnaire while considering a high learning class ($N = 187$, coded as "1"), the other a low learning class ($N = 184$, coded "0"). Of the 96 items the median point-biserial correlation was .43; the range was from .02 to .73, with negative items reverse scored (Freedman & Stumpf, 1976). Items with $r_{pb} < .35$ were subsequently excluded from the final CFI due to their low validity.

In order to estimate the stability of the point-biserial correlations, the High and Low learning samples were randomly divided in half. Separate point-biserial correlations were computed. These were then rank-ordered for each dimension and a Spearman rank-order correlation coefficient was computed. The results indicated that the point-biserial correlations were stable for each dimension (median $r_s = .87$). In addition, each half sample point-biserial correlation was converted to a Fisher z-score which was then compared to the same item's Fisher z from the other half sample. Using a .05 level of significance, only 4 out of 96 pairs of correlations were significantly different, which would be expected by chance.

Each dimension was factor analyzed using principal factoring with Varimax rotation ($N = 371$). The simplicity and strength of factor loadings, the strength of the point-biserial correlations, and the internal consistency of the dimensions (as indexed by coefficient alpha) were criteria for instrument construction. The criteria were used sequentially to exclude items within a dimension which did not contribute to the psychometeric properties of the CFI. (For a description of this iterative process see Freedman & Stumpf, 1976, pp. 7–8.) The use of these criteria resulted in seven factorially simple dimensions

which were relatively independent of each other: Subject Functionality, Subject Affect, Subject Difficulty, Instructor in Class, Instructor in General, Graded Assignments/Examinations, and Text/Required Readings. All items within each dimension had correlations of .39 or higher with the perceived learning criterion and contributed substantially to coefficient alpha. Items which met the other criteria, but did not contribute to coefficient alpha, were deleted for reasons of parsimony.

The factor analyses were replicated (N = 1,332) and the dimensions examined for factor stability. The median coefficient of congruence was .98, indicating that the factor structure was virtually identical between the two samples (Harman, 1967). All dimensions were reliable in the replication group, where the median coefficient alpha was .82 with a range of .72 to .85. Table 1 summarizes these properties, provides sample items for each dimension, and presents the dimension intercorrelations.

Inspection of the intercorrelation matrix in Table 1 indicates that the seven dimensions were not independent. However, the dimensions were developed to correlate highly with perceived learning; the intercorrelations are to some extent the result of the common relationship of each dimension to the student's perception of learning. Since each dimension correlates .35 or below with most other dimensions, each contributes meaningful new information. When the relationship of each variable to perceived learning was controlled (i.e., learning is partialled from the intercorrelations), the largest intercorrelation (.55) was reduced to a partial correlation of .24. The lower intercorrelations (.35 and below) were reduced to near zero.

Correlates of CFI Dimensions

Data on potential correlates were collected during the development of the CFI for their descriptive value and/or because they had been previously found to covary with evaluations by others. Correlates not believed to be related to the perceived learning criterion could be indicative of an evaluation bias. Therefore, such variables were investigated as part of the validation process. The variables studied included: learning style as measured by the Learning Style Inventory (Kolb, Rubin, & McIntyre, 1974); expected grade; years of full-time work experience; grade point average; reason for taking the course (i.e., requirement or elective); present major; hours spent studying per week; perceived course difficulty; perceived degree of course structure; sex; age; program of study (e.g., MBA, MS); full-time or part-time student status; number of courses completed; and undergraduate major. Although many of these variables did show significant relationships with some CFI dimensions, the relationships were generally weak (r's $< .15$). Furthermore, since a large number of correlations were tested for statistical significance, some of the significant results may have occurred by chance. Table 2 presents representative CFI dimension correlates.

Learning styles. Since the CFI was developed using a perceived learning criterion, it is possible that different learning styles might correlate with evaluations on various dimensions. The Learning Style Inventory (LSI) was administered to the replication sample (N = 1332) in order to investigate this possibility. If students with identifiable styles of learning provided different evaluations on course and instructor dimensions, this should be considered when using CFI results in course development or decision-making.

The LSI measures learning style along two bipolar factors, referred to as Active Experimentation-Reflective Observation (AE-RO) and Abstract Conceptualization-Concrete Experience (AC-CE). The AE-RO factor was the student's preference for learning by either experimentation or reflection, measured along a bipolar continuum. The AC-CE factor was the student's preference for learning by conceptualization or experience, again measured along a bipolar continuum. The low correlations of the two LSI factors with each CFI dimension (Table 2) indicate

Table 1
CFI Scales, Measurement Characteristics, and Intercorrelations

Scale	Number of Items	Median Item Point-Biserial[a]	Coefficient Alpha[b]	Scale Intercorrelations[c] SF	SA	SD	IC	IG	ASG
Subject Functionality (SF) (valuable, useful, relevant, and significant)	4	.53	.84						
Subject Affect (SA) (interesting, stimulating, enjoyable, and exciting)	4	.70	.81	.55					
Subject Difficulty (SD) (challenging, demanding, and difficult)	3	.49	.77	.21	.16				
Instructor in Class (IC) (clear, thought-provoking, stimulating)	8	.63	.83	.33	.49	.16			
Instructor in General (IG) (helpful, patient, bad, interested in students, and tolerant)	5	.47	.72	.15	.22	.07	.50		
Graded Assignments/Exams (ASG) (clear, valuable, vague, ambiguous)	7	.52	.85	.21	.23	.07	.35	.30	
Text/Required Readings (TEXT) (worthwhile, interesting, good, practical)	7	.55	.82	.52	.51	.29	.33	.16	.18

[a] N = 371

[b] The sample N=1332 but the N's for each scale varied an average of 5% due to missing data. N = 1104 for the ASG dimension because some classes had been given no graded assignments or examinations at the time the questionnaire was administered.

[c] All correlations are significant at $p \leq .05$ (two tailed), $N \leq 1332$.

that students with different learning styles did not evaluate courses or faculty differently. However, these correlations should be interpreted cautiously since the validity of the LSI has subsequently been questioned (Freedman & Stumpf, in press).

Expected grade. As Table 2 shows, expected grade correlated significantly with each dimension, but was able to account for only a small amount of student variance in dimensions. Multiple regression analysis was used to determine if other correlates (particularly years of work, grade point average, and reason for taking course) could explain significant dimension variance beyond that of expected grade. No substantial increments in the correlations were obtained.

The correlation of expected grade with evaluations has frequently been considered an indication of student bias (e.g., Bausell & Magoon, 1972; Schuh & Crivelli, 1973). Stumpf and Freedman (1977) provided an alternative explanation of the expected grade correlate as covariance introduced by the professor, through his/her projection of being a difficult or easy grader. However, for either interpretation it may be desirable to partial out the expected grade covariance with CFI dimensions. A method for removing this variance, should it become substantial, is available (Freedman & Stumpf, 1976; Schwab & Forrest, 1975).

Dependent Variables and Coursewise Cross-Validation

Since the CFI was to be used for decision-making, three overall rating variables were developed: an instructor rating (three questions using 5-point scales, coefficient alpha = .90), course rating (three questions using 5-point scales, coefficient alpha = .83), and text rating (one question).

The rating variables were conceptually distinct from the CFI dimensions. The dimensions were designed to relate to perceived learning, whereas the ratings were overall evaluations. Rating questions included: (1) Would you recommend this instructor to friends? (2) Give this instructor (course) a grade? (3) All things considered, how would you rate [the amount learned, instructor, subject matter, or text] compared with all *other* courses and instructors you have had?

By treating the overall ratings as dependent variables, the seven CFI dimensions could be used as predictors. Based on the theory that instructor dimensions should predict instructor ratings and that subject dimensions should predict course ratings, relevant dimensions were selected and hierarchically regressed on each rating, using the class as the unit of analysis (N = 47, where data from the 1,332 students were averaged by class). The class was believed to be the appropriate unit of analysis when using CFI

Table 2
Correlates of Student Variables with CFI Scores

CFI Scale	Learning Style AE-RO	Learning Style AC-CE	Expected Grade	Years of Work	Grade Point Average	Reason for Taking Course (Requirement or Elective)
SF	.01	.05	.14*	.03	.04	.08*
SA	.01	.03	.17*	.13*	.04	.08*
SD	-.01	-.07*	-.14*	.01	.08*	.01
IC	.00	-.01	.14*	.10*	.10*	-.03
IG	.05	-.01	.12*	.05*	.05*	.07*
ASG	.06*	.07*	.15*	.12*	.12*	.05
TEXT	.02	.06*	.13*	.09*	.09*	.00

*$p \leq .05$ (two-tailed), $N \leq 1332$.

results to make decisions comparing courses or instructors. Eighty-three percent of the variance in instructor ratings was predicted by IC, IG, and ASG; 72% of the variance in course ratings by SF, SA, SD, and ASG; and 66% of the variance in text ratings by TEXT.

In a classwise cross-validation study of the CFI, the initial weights were used to predict the new sample ($N = 171$ classes involving 4,592 student ratings). These regression equations could have accounted for 83% of instructor rating and 77% of course rating variance. The optimal weights accounted for 84% and 80% of the variance, respectively. Thus, the CFI dimensions could meaningfully predict student ratings.

Alternative Scaling Procedures and Convergent-Discriminant Validity

There has been considerable debate over the extent to which different rating scales and formats are susceptible to bias (e.g., Fogli, Hulin, & Blood, 1971; Norton, Gustafson, & Foster, 1977; Zedeck, Kafry, & Jacobs, 1976). To facilitate selecting a method less susceptible to bias, three scaling methods were investigated using a multitrait-multimethod (MTMM) approach. An additional purpose of this analysis was to obtain reliability and validity data on the CFI dimensions that were comparable to data previously collected by Schwab (1974). In so doing, two instruments developed against the same criterion could be directly compared.

New York University Sample and Procedure

The sample consisted of 201 graduate business students (three classes) enrolled during the summer semester of 1976. Each student was asked to evaluate his/her class using a MTMM version of the CFI. Sufficient data was obtained from 176 students. The MTMM instrument developed contained the seven dimensions with three different question format and scaling methods. Method I asked the student if an adjective or descriptive phrase described the present class and used a three-point scale labeled "Y" for Yes, "?" for Undecided, and "N" for No. For example, in the subject difficulty dimension, "demanding" was presented as follows:

Demanding Y ? N

[See Dubois & Burns (1975) for analysis of the question mark anchor, and Smith, Kendall, & Hulin (1969, pp. 31–35) for use of similar scales]. Method II used short descriptive statements which were responded to on a 5-point Likert scale, ranging from "strongly agree" through "undecided" to "strongly disagree." Method III was a 7-space sematic differential type of scale with bipolar adjectives, such as "Demanding—Undemanding". All negative items were reverse scored. The three method scores for each dimension were intercorrelated and cast into a MTMM matrix (Campbell & Fiske, 1959).

University of Wisconsin-Madison Sample and Procedure

Schwab (1974) performed a similar MTMM analysis using data from 149 graduate business students (four classes) taking courses at the School of Business, University of Wisconsin-Madison in 1973. The Course Evaluation Instrument (CEI) developed by Schwab (1974) consisted of five dimensions with nine items each. The MTMM version of the CEI used three scaling methods. Method I was the same three-point scale (Y ? N) as that used at NYU. Method II was a triadic scoring system which compared student evaluations of the present course to his/her ideal of a high and low learning course (Schwab, 1974). Triadic scoring, developed by Smith et al. (1969, pp. 32–34), determined the scoring direction for each question based on individual preferences rather than a group norm. Each student was given three questionnaires using the Method I scale and the following separate directions: (1) describe your present course; (2) describe a high learning course; and (3) describe a low learning course. The scoring scale derived for the triadic method was "+1," "0,"

and "−1," based on the degree to which the three scores were congruent. "For example, if a student indicated that an item characterized a High learning course, the present course but not the Low learning course, it would receive a positive score" (Schwab, 1974, p. 7). Method III used the same adjectives as Method I but was responded to on a 7-point Likert type scale (1—strongly agree to 7—strongly disagree). All negative items were reverse scored. The three method scores for each dimension were intercorrelated and presented in a MTMM matrix by Schwab (1974).

Results

MTMM matrices. Tables 3 and 4 provide the MTMM matrices for the CFI and CEI, respectively. Interpretation of each matrix is as follows (Campbell & Fiske, 1959):

1. Convergent validity exists when different methods of measuring the same characteristics significantly intercorrelate. Thus convergent validity is demonstrated in the matrix by the significant (and large) monotrait-heteromethod correlations, i.e., the validity diagonal entries.
2. "Evidence for discriminant validity is threefold. First, the correlations in the validity diagonal should be higher than those in the same column and row in which neither [dimensions] nor [method] are in common. Second, the values in the validity diagonal should be higher than those correlations between the [dimension] and other [dimensions] with common [method]. Third, the pattern of [dimension] interrelationships should be the same within and between [methods]" (Kavanagh, MacKinney, & Wolins, 1971, p. 35). Both matrices showed general support for each of the above comparisons, with the exception of the CEI method II/III correlations presented in Table 4.

The convergent validity diagonals were similar for the three methods used with the CFI (see Table 3). CEI correlations (see Table 4) for methods I/II validity diagonals were substantially larger than other CEI validity diagonals. This suggests that method I may be superior. Inspection of the heterotrait-monomethod triangles in both studies provided additional support for Method I. The intercorrelations of dimensions were generally lower, indicating less method bias.

There are two effects on MTMM correlations due to differences in scaling methods that need explanation:

1. Methods which use fewer scale points are less reliable (Jenkins & Taber, 1977; Lissitz & Green, 1975); therefore, the correlations within methods using fewer scale points were attenuated more than methods using a greater number of scale points.
2. Methods which use fewer scale points *may* have a restricted range, curtailing the correlations with other methods (Cohen & Cohen, 1975, p. 65).

Since the methods selected involve either 3, 5, or 7 scale points, substantial differences in reliability might be expected. Using a monte carlo approach, Lissitz and Green (1975) simulated the effect on coefficient alpha (based on ten items) of various interitem covariances and number of scale points. Their data relevant to the CFI indicated that for a covariance of .5 among items within a dimension and 3, 5, and 7 scale points, coefficient alpha would be .71, .74 and .75, respectively. This suggests that the correlations involving method I in the CFI study were more severely attenuated than methods II or III. Similarly, methods I and II were more attenuated in the CEI study. However, the effect was small. Since the validity diagonal correlations involving method I were generally equal to or larger than the other validity diagonals even though they were more severely attenuated, method I would still be preferred.

Table 3

Convergent-Discriminant Validity Correlations for Three Scaling Methods of the CFI

(New York University, Graduate School of Business, N=176)

Scaling Method and CFI Scale	Method I ("Y," "?," "N")							Method II (Likert Scale)							Method III (Semantic Differential)						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Method I																					
1. Subject Functionality	--																				
2. Subject Affect	58	--																			
3. Subject Difficulty	32	14	--																		
4. Instructor in Class	55	47	24	--																	
5. Graded Assignments	40	24	08	39	--																
6. Text and Readings	56	58	30	36	27	--															
7. Instructor in General	16	13	04	33	54	14	--														
Method II																					
1. Subject Functionality	61	44	40	39	42	50	22	--													
2. Subject Affect	62	70	34	51	37	60	23	76	--												
3. Subject Difficulty	31	14	79	18	13	28	10	56	40	--											
4. Instructor in Class	44	37	24	55	48	31	35	61	64	34	--										
5. Graded Assignments	33	34	23	20	87	45	53	52	43	29	46	--									
6. Text and Readings	51	46	29	31	36	82	18	67	66	36	42	57	--								
7. Instructor in General	22	19	07	38	61	22	79	34	34	11	61	62	32	--							
Method III																					
1. Subject Functionality	69	55	32	38	34	62	17	65	68	40	41	47	62	27	--						
2. Subject Affect	56	62	27	44	33	57	24	50	80	30	45	33	56	32	80	--					
3. Subject Difficulty	31	17	75	14	09	32	12	51	42	83	27	28	37	12	51	42	--				
4. Instructor in Class	42	48	14	57	35	38	38	42	52	19	62	33	35	49	61	68	27	--			
5. Graded Assignments	38	17	24	18	85	44	59	51	37	29	47	89	55	60	58	48	32	43	--		
6. Text and Readings	47	52	27	30	29	79	15	48	64	27	37	45	80	28	72	74	37	51	53	--	
7. Instructor in General	25	23	04	30	50	30	74	25	27	05	41	51	29	78	37	44	13	65	61	40	--

Note. Decimals are omitted. Due to some classes not having had a graded assignment at the time the data were collected, N=37 for the ASG dimension.

Table 4

Convergent-Discriminant Validity Matrix for Three Scaling Methods of the CEI

(University of Wisconsin-Madison Graduate Students, N=149)

Scaling Method and CEI Scale	Method I ("Y," "?," "N")					Method II (Triadic)					Method III (7-point Likert Scale)				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Method I															
1. Subject Matter	--														
2. Instructor in Class	54	--													
3. Assignments/Exams	32	46	--												
4. Text/Required Reading	45	56	23	--											
5. Instructor in General	38	68	53	42	--										
Method II															
1. Subject Matter	60	34	19	40	24	--									
2. Instructor in Class	31	67	34	50	48	48	--								
3. Assignments/Exams	24	27	56	16	17	48	45	--							
4. Text/Required Reading	43	50	26	54	36	73	62	48	--						
5. Instructor in General	30	42	25	36	46	51	66	41	57	--					
Method III															
1. Subject Matter	57	45	51	31	36	35	33	46	33	25	--				
2. Instructor in Class	43	61	69	22	55	26	41	48	26	30	67	--			
3. Assignments/Exams	16	17	39	08	20	19	09	37	13	12	46	50	--		
4. Text/Required Readings	52	48	46	54	36	42	43	44	36	33	66	54	29	--	
5. Instructor in General	40	54	67	21	62	24	41	42	32	31	56	77	37	46	--

Note. Decimals are omitted. The matrix is a reconstructed version of the matrix reported by Schwab (1974, p. 16).

The question of restriction in range due to the number of scale points can be addressed by considering the standard deviations for each CFI dimension and method. Dimension scales were formed by taking the simple average of the items within the dimension. Since dimensions were intercorrelated (not individual items), the number of scale points varied with methods *and* the number of items in the dimension. For example, the 3-point scale (method I) for the Subject Affect dimension (4 items) resulted in an 8-point Subject Affect scale; the 5-point scale (method II) for the Instructor in Class dimension (8 items) yielded a 32-point scale.

Since all dimension scales involved a moderate to large number of scale points, substantial restriction in range was unlikely. Comparison of the standard deviations for each CFI dimension across methods confirmed this expectation; the standard deviations were homogeneous. The correlations were comparable within the CFI MTMM matrix. Although the standard deviations for the methods used by Schwab (1974) were not reported, the similarity of methods across studies suggests similar results. Furthermore, each CEI dimension contained 9 items; the number of scale points for methods I and II was actually 18.

Comparison of Results Between Studies

Comparison of two multitrait-multimethod matrices is by no means straightforward. Kavanagh et al. (1971) indicated that the usual procedure suggested by Campbell and Fiske (1959) "is inferential, implicit, and, in the case of large numbers, awkward. In addition, the comparison of effects within and between studies is difficult, if not impossible" (p. 37). To overcome these difficulties, both matrices were further analyzed in ANOVA terms; the results are presented in Table 5. The ANOVA uses the correlations in the matrix as raw data (Boruch, Larkin, Wolins, & MacKinney, 1970, p. 839; Kavanagh et al., 1971). This permits a clear and direct comparison of the two instruments. The advantage of the method is that researchers may perform comparisons based upon published MTMM tables rather than following the time-consuming, and possibly fruitless, procedure of gaining access to the raw data of other researchers required by other comparison approaches (e.g., Jöreskog, 1974).

All F ratios in both studies were significant at the $p \leqslant .001$ level (see Table 5). The main effect for the course represented convergent validity, where the variance component indicated the strength of agreement on the course characteristics by the students. The course main effect summarized the data across all dimensions and methods and thus provided a single index of convergence. Both studies showed moderate convergence (variance component for CFI = .44, CEI = .43).

The course by dimension interaction indicated the degree of discrimination among dimensions. The CFI results showed moderate discriminant validity (.39); however, the CEI had low discriminant validity (.15). Students were not able to discriminate among the five CEI dimensions, but could discriminate among the seven dimensions of the CFI.

Two sources of error were also identified by the ANOVA. The course by method interaction is a measure of method bias. Both studies showed little method bias (CFI = .08, CEI = .17). The three-way interaction of course, dimensions, and methods is a measurement error term. The CFI had relatively low error (.18), while the CEI had moderate error (.35). (Main effects for dimension, method, and the interaction of dimensions and methods were not included in the Kavanagh et al. method of analysis.)

The practical significance of the error terms is clarified, and a more direct comparison of ANOVA results is obtained, by computing comparison indices (Kavanagh et al., 1971). Each comparison index was computed by dividing the respective variance component by the sum of the variance component and the error term. For example, the CFI course main effect comparison index equalled .71 (.44 divided by the sum of .44

Table 5
Analysis of Variance Results for CFI and CEI Convergent-Discriminant Validity Matrices

Instrument and Source	Degrees of Freedom	MS	F	Variance Component	Comparison Index
CFI					
[1]Course (C)	175	9.39	51.8*	.44	.71
C X Dimension (C X D)	1050	1.34	7.4*	.39	.68
C X Method (C X M)	350	.75	4.1*	.08	.31
Error (C X D X M)	2100	.18		.18	
CEI					
[1]Course (C)	148	6.80	19.4*	.43	.55
C X Dimension (C X D)	592	.79	2.3*	.15	.30
C X Method (C X M)	296	1.18	3.4*	.17	.33
Error (C X D X M)	1184	.35		.35	

Note. The unit of analysis is the student. However, the ANOVA calculations are based on the correlations within the MTMM matrix (Kavanagh et al., 1971).

[1]The course main effect is an average of within class variance.

*$p \leq .001$

and .18). Examination of the comparison indices for the two instruments highlighted their differences.

Special Considerations for Interpreting Results

The data for each study were collected from the student populations used to develop the respective instruments. Thus, the data presented in Tables 3 through 5 may overstate the strengths of each instrument; but the comparisons made are, nonetheless, appropriate.

A second consideration is that the methods within each study used similar question content. In the NYU study each method used different formats and response scales; the University of Wisconsin–Madison study used a different response scale or scoring method. Thus, convergent validity herein should be interpreted along a validity-reliability continuum. Since both studies employed analogous methods, their comparison is valid.

In performing the MTMM analysis, it has been assumed that the class provides a distinct stimulus to each student, i.e., the student is the appropriate unit of analysis. For instrument development purposes, the authors believe this to be a valid assumption since there was substantial variance in student evaluations of the same course. Yet, the class may be a more relevant unit of analysis when instructor or course evaluations are compared for decision-making. Since research by Linn, Centra, and Tucker (1975) and Barnoski and Sockloff (1976) indicated that different units of analysis had little effect on the factor structure of student evaluations, the CFI is believed to be valid for both units of analysis.

Discussion

Validities

Both the CFI and CEI demonstrated acceptable validities. Thus, student learning perceptions are a meaningful and relevant, albeit limited, criterion of educational effectiveness. The high point-biserial correlations, with modified dimensions and in different business school environments, support the criterion's generalization as an evaluation construct.

To further investigate this criterion, additional research has been undertaken. For ex-

ample, the CEI has demonstrated validity generalization to three undergraduate business schools and their associated graduate business schools (University of Wisconsin–Whitewater, University of Missouri—Columbia, and Cornell University; Schwab & Wallace, 1975). The CFI has demonstrated validity generalization to NYU's College of Business and Public Administration (undergraduate) and to Harvard Business School. The CFI has also demonstrated validity extention to undergraduate social science students (Rutgers University) and to military graduate engineering students (Air Force Institute of Technology; Stumpf, Freedman, & Krieger, 1977).

Differences in instrument construction are believed to have contributed to the higher discriminant validity of the CFI. This can be attributed to the factor analytic procedures employed by Freedman and Stumpf (1976). By insuring that each dimension was comprised of items with simple factor loadings, the relative degree of scatter around each dimension was reduced. The lower bias and error terms observed in the analysis of the CFI may to due to methods selected in the multimethod analysis or the dimensions used within a method. Further research on each instrument would be necessary to explain these differences.

Scaling Method

Although only a few scaling methods and formats were investigated, the consistent results across two independent studies are encouraging. The "Y, ?, N" format was superior in both studies. This format has other desirable qualities, such as its simplicity, the small amount of space required on the questionnaire, and that no additional computations are necessary (in contrast to the triadic method). Since course-faculty instruments are generally administered to hundreds (or thousands) of students each semester, the above attributes of method I are meaningful. For example, 48 items are asked on one side of one page of the CFI; and the instrument can be completed in about 5 minutes.

Dimensions

Students can discriminate between course and instructor dimensions within a class. The degree of correlation for the CFI was low (the median r was .19 for the three subject dimensions with the two instructor dimensions). Considering the fact that all dimensions were developed to correlate highly with perceived learning (as indexed by the point-biserial correlations in Table 1), the low intercorrelations among dimensions are encouraging. A factor analysis of the seven dimensions confirmed the existence of two general factors: instructor and course.

Implications

Since evaluations are often used in decision-making (Rodin & Rodin, 1972), it is imperative that the instruments be reliable and valid. It appears that students can provide valid evaluations on the perceived learning criterion. In addition to the level of reliability and validity demonstrated, a comparative strategy permits the determination of an instrument's relative advantages. Future research should compare instruments validated against *different* criteria using the *same* student population. In this way different validation criteria can be assessed. By incorporating different scaling methods and formats into the above studies, other measurement characteristics of the instruments can be compared. If such studies can subsequently be cross-validated or shown to generalize and/or extend to other populations, a general instrument and/or process of validation would emerge.

Additional research should also investigate the extent to which instruments, validated against different criteria which have similar dimensions, converge and discriminate. A MTMM strategy investigating this aspect of course-faculty evaluation instruments exemplifies the initial logic suggested by Campbell and Fiske (1959). Once an instrument is developed and shown to have similar measurement properties for various student populations, the

courses, instructors, and student perceptions across universities can be compared. Such research is currently underway using the CFI.

References

Aleamoni, L., & Spencer, R. The Illinois course evaluation questionnaire: A description of its development and a report of some of its results. *Educational and Psychological Measurement*, 1973, *33*, 669–684.

Barnoski, R. P., & Sockloff, A. L. A validation study of the faculty and course evaluation (FACE) instrument. *Educational and Psychological Measurement*, 1976, *36*, 391–400.

Bausell, R., & Magoon, J. Expected grades in a course, gradepoint average, and student rating of instructor. *Educational and Psychological Measurement*, 1972, *32*, 1013–1023.

Boruch, R. F., Larkin, J. D., Wolins, L., & MacKinney, A. C. Alternative methods of analysis: Multitrait-multimethod data. *Educational and Psychological Measurement*, 1970, *30*, 833–853.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multi-trait multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81–105.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1975.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, *41*, 511–535.

DuBois, B., & Burns, J. A. An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, 1975, *35*, 869–884.

Fogli, L., Hulin, C. L., & Blood, M. R. Development of first-level behavioral criteria. *Journal of Applied Psychology*, 1971, *55*, 3–8.

Freedman, R. D., & Stumpf, S. A. *Course-Faculty Evaluation Instrument: Development and applications* (Working Paper #76–59). New York: New York University, Graduate School of Business Administration, November 1976.

Freedman, R. D., & Stumpf, S. A. What can one learn from the Learning Style Inventory. *Academy of Management Journal*, in press.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. *Journal of Applied Psychology*, 1973, *58*, 261–265.

Harman, H. *Modern Factor Analysis*. Chicago: University of Chicago Press, 1967.

Holmes, D. The teaching assessment blank (TAB): A form for the assessment of college instructors. *Journal of Experimental Education*, 1971, *39*, 34–38.

Jenkins, G. D., & Taber, T. D. A monte carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, 1977, *62*, 392–398.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Measurement, Psychophysics, and Neural Information Processing*. (Vol. 2). San Francisco: W. H. Freeman & Co., 1974, pp. 1–56.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, 1971, *75*, 34–49.

Kolb, D., Rubin, I., & McIntyre, J. *Organizational psychology: An experiential approach* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1974.

Linn, R. L., Centra, J. A., & Tucker, L. Between, within, and total group factor analyses of student ratings of instruction. *Multivariate Behavioral Research*, 1975, *10*, 277–288.

Lissitz, R. W., & Green, S. B. Effect of the number of scale points on reliability: A monte carlo approach. *Journal of Applied Psychology*, 1975, *60*, 10–13.

McKeachie, W. J., Lin, Y., & Mann, W. Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal*, 1971, *8*, 435–455.

Norton, S. D., Gustafson, D. P., & Foster, C. E. Assessment for management potential: Scale design and development, training effects, and rater/ratee sex effects. *Academy of Management Journal*, 1977, *20*, 117–131.

Rodin, M., & Rodin, B. Student evaluation of teachers. *Science*, 1972, *177*, 1164–1166.

Schuh, A. J., & Crivelli, M. A., Aminadversion error in student evaluations of faculty teaching effectiveness, *Journal of Applied Psychology*, 1973, *58*, 259–260.

Schwab, D. *Development of the CEI: Scale construction and initial testing* (CEI Report No. 1, University of Wisconsin—Madison: Graduate School of Business and Industrial Relations Research Institute, February 1974.

Schwab, D. *Manual for the Course Evaluation Instrument*. University of Wisconsin—Madison: Graduate School of Business and Industrial Relations Research Institute, 1976.

Schwab, D., & Forrest, R. Accounting for the relationship between students' evaluation of courses and expected grades. *Proceedings of the 35th Annual Meeting of the Academy of Management*, 1975, 10–12.

Schwab, D., & Wallace, M. The validity extension of a measure to assess teaching effectiveness in business schools. *Proceedings of the Midwest Academy of Management Meeting*, 1975, *18*, 128–135.

Sharon, A. T. Eliminating bias from student ratings of college instructors. *Journal of Applied Psychology*, 1970, *54*, 278–281.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes.* Chicago: Rand McNally, 1969

Sockloff, A. L. Instruments for student evaluation of faculty: Ideal and actual. In A. L. Sockloff (Ed.), *Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students.* Philadelphia: Temple University, Measurement and Research Center, 1973, 132–151.

Stumpf, S. A., & Freedman, R. D. The nature of expected grade bias in course-faculty evaluations (Working Paper #77–08). New York: New York University, Graduate School of Business Administration, March 1977.

Stumpf, S. A. Freedman, R. D., & Krieger, K. *Validity generalization and extension of the Course-Faculty Evaluation Instrument (CFI).* Unpublished Manuscript, New York University, Graduate School of Business Administration, September, 1977.

Wallace, M., & Schwab, D. The validation of a teaching-effectiveness measure in two business schools. *Proceedings of the 33rd Annual Meeting of the Academy of Management*, 1973, *33*, 229–235.

Zedeck, S., Kafry, D., & Jacobs, R. Format and scoring variations in behavioral expectation scales. *Organizational Behavior and Human Performance*, 1976, *17*, 171–184.

Author's Address

Richard D. Freedman, Management/Organizational Behavior Department, Graduate School of Business Administration, New York University, 100 Trinity Place, New York, NY 10006