The Construct Validity of Teaching Behavior Evaluation Methods: A Multitrait-Multimethod Analysis

M. Bruce Lammers and Don F. Kirchner

Journal of Marketing Education 1985 7: 35
DOI: 10.1177/027347538500700206

The Construct Validity of Teaching Behavior Evaluation Methods: A Multitrait-Multimethod Analysis

M. Bruce Lammers and Don F. Kirchner

The construct validity of five different methods of assessing five separate teaching behavior traits was examined in the framework of a 5 x 5 multitrait-multimethod matrix. The results indicated that videotape rating methods produced convergent and discriminant validity coefficients which were greater than or equal to those derived from the more traditional peer and student rating methods.

The evaluation of teaching ability is typically done by having peers and/or students evaluate the teaching behavior of the professor under review. The peer evaluation procedure currently in use at California State University, Northridge (CSUN) calls for a classroom visitation to be made by at least one tenured faculty member who acts then as the *in vivo*, nonparticipant observer of the professor's teaching behavior. Among the known disadvantages of the *in vivo* method are:

(1) It assumes that the presence of an *in vivo* observer does not significantly alter the behavior of the professor or of the students in the classroom (Samph 1976);

(2) The maximum exposure of the *in vivo* observer to any single behavioral sequence during the class is one; i.e., there is no chance for an instant replay or stopaction of any segments the observer may have missed or miscomprehended; and,

(3) It is an unwieldly and inefficient method to schedule and execute when the number of available observers is small and the available observation times are limited.

Given the drawbacks of the *in vivo* method, an alternative peer evaluation method suggested by the present article is the videotape method. With this method, a professor's classroom behavior is videotaped and submitted for subsequent evaluation to the department's personnel committee. Unfortunately, both the *in vivo* observer method and the videotape method suffer from the possibility that the presence of an observer, whether *in vivo* or implicit, may alter the behaviors of interest (Campbell and Stanley 1963). According to Mercatoris and Craighead (1974), "There is no consensus concerning reactivity of the observational process, particularly within pedagogical settings" (p. 513). Their own study found that the quality of teacher-pupil interaction was unaffected by direct observation, but the quantity of the interaction was greater under observation than under nonobservation.

Most researchers, though, agree that the perception of being observed somehow affects one's behavior. In this vein, Harrop (1979) noted that several problems are associated with observing human behavior in the classroom, some of which are (1) observer expectation, (b) the effect of the observer's presence on the students, (c) the effect of the observer's presence on the teachers, and (d) the phenomenon of observer drift, the dramatic variation over time in the behavioral ratings given by individual observers or groups of observers. It is important to keep in mind that Harrop was specifically referring to the problems associated with being observed by a live observer. A mechanical observer, such as the videotape camera, contains advantages which promise to alleviate some of the problems associated with *in vivo* observation. For example, among the advantages of the

H. Bruce Lammers and Don F. Kirchner are Professors of Marketing at California State University, Northridge, California.

videotape method are (Dowrick and Biggs 1983) the following:

(1) It allows for multiple exposures (instant replays) of behavioral sequences, thereby increasing interrater reliabilities and attenuating the risk of miscomprehension of what occurred (see Meadowcroft and Moxley 1980, p. 27);

(2) It provides a permanent record which can be submitted in the case of appeals;

(3) It allows for greater flexibility in the scheduling of rating times;

(4) It eliminates the influence of varying observer characteristics; e.g., clothing, mannerisms, age, sex, race, friend, foe;

(5) It eliminates interpersonal contact between the observer and the observed;

(6) By collecting all data on videotapes and then presenting the tapes to observers in random or counterbalanced order, the problem of observer drift can be attenuated.

In light of the many possible advantages in evaluation methodology that the videotape method offers, it is important to determine the extent to which recordings of teaching behavior from the videotape method are comparable to recordings from the *in vivo* method. The empirical research on this question has come from areas outside of the pedagogical setting (cf. Kessler 1979, 1981). Moreover, the research interest on live versus videotape effects comes largely from social psychologists concerned with juror behavior. Miller (1975) reported no significant differences in jurors' responses to live and videotaped trials, with the exception that jurors who viewed the videotaped testimony retained more trial-related information than did jurors who viewed a live trial. Kassin (1984) found that the presence of a camera did not affect jurors' verdicts, awards, or a series of self-report measures. Even the Supreme Court in its famous 8-0 ruling in *Chandler v. Florida* (1981) has taken the stand that "no one has been able to present empirical data sufficient to establish that the mere presence of the broadcast media has an adverse impact on the [judicial] process." Finally, in an extensive overview of various observational procedures in many different situations beyond the courtroom applications, Kent and Foster (1977) concluded that alternatives to the *in vivo* method (e.g., videotape, one-way mirrors) possess clear "methodological advantages of lessening or eliminating potential biases which . . . seem to *more than justify considering such alternatives to in vivo recordings*" (emphasis added, p. 294).

Nevertheless, given the obvious method variances involved between the *in vivo* versus the videotape procedures, it would be presumptuous to expect no significant differences between the videotape method and the *in vivo* method across all situations. And, as mentioned earlier, the issue of *in vivo* vs. videotape procedures for evaluating teaching performance has received little or no research attention from educational psychologists. Consequently, the purpose of the present study was to assess the construct validity of five methods of evaluating teaching behavior. These five methods are variants of the *in vivo* and the videotape methods and are described in more detail in the Method section.

The assessment of construct validity requires that measures of both convergent and discriminant validity be taken. Campbell and Fisk's (1959) Multitrait-Multimethod Matrix (MTMM) allows for a convenient method of determining these validities. The MTMM technique is primarily concerned with the adequacy of tests as measures of constructs. It provides information on (a) whether the trait or construct can be observed under more than one experimental condition, (b) whether the trait can be meaningfully differentiated or discriminated from other traits, and (c) how much of the variation between traits can be attributed to characteristics of the trait versus the measure of these traits (Goodman, Furcon, and Rose 1969).

METHOD

Procedure

An invitation to participate in the CSUN Marketing Department's project on "evaluating methods used to assess teaching behavior" was sent to all nontenured faculty members of the department (five full-time and 14 part-time

faculty). Six faculty members (four full-time and two part-time) volunteered to have their teaching behavior observed. Two separate undergraduate sections from each of the participating faculty's teaching schedule were randomly selected to be the observation sessions. For those faculty who were teaching more than one undergraduate course, the two observation sessions were drawn from two different randomly selected courses, excluding independent study courses. The specific dates on which the sessions were to be observed were also randomly selected.[1] In the final analysis, three sessions of introduction to business, three sessions of introduction to marketing management, and one session each of marketing research, sales management, retailing, consumer behavior, and marketing management were observed.[2] The sessions were observed within a two-week period near the end of the fall semester.

Multimethods

Five methods of scoring the teaching behavior of the participants were used: the *in vivo* student method, the *in vivo* peer method, the videotape peer prediscussion method, the videotape peer postdiscussion method, and the students' end-of-the-semester evaluations method.

The in vivo peer and in vivo student methods. Each session was videotaped and rated by both the students in the class (*in vivo* student method) and by a peer observer (*in vivo* peer method). The *in vivo* peer observers were randomly selected from the available pool of four tenured marketing department faculty. The assignment of the *in vivo* peer observers was constrained by the limitations that no observer be assigned to the same teacher more than once and that each observer evaluate three sessions.

The videotape peer pre- and postdiscussion methods. Approximately four weeks after the sessions were observed, the four observers met as a group and evaluated the sessions by playing back the videotapes in a randomly determined sequence. Three meetings were needed to review all the tapes over a two-week period. During the playback session, the observers first completed their ratings independently of one another, and then the group arrived at a consensus rating following free discussion. The mean of the independent ratings constituted the videotape peer prediscussion method, and the consensus ratings constituted the videotape peer postdiscussion method.

The students' end-of-the-semester evaluations method. A fifth rating method took place at the end of the semester during the School of Business and Economics' regularly scheduled teacher evaluation. Unlike the previous *in vivo* student ratings, the students were instructed this time to evaluate the teacher's performance over the entire semester, rather than over a single class session. This rating constituted the students' end-of-the-semester evaluations method.

Multitraits

The LEF rating instrument. The rating instrument used by both the student and peer observers was the 24-item lecture evaluation form (LEF) currently in use by the department of marketing (see Appendix A).[3] Students were assured of their anonymity in this project.

Factor analysis of the LEF. Rather than treating each of the 24 items as separate traits, a principal components factor analysis with varimax rotation was performed on the faculty observers' ratings to identify the traits to be used in the MTMM analysis. The results of the factor analysis showed that the questionnaire was largely unidimensional (see Table 1 for factor loadings matrix). Factor A accounted for 78.7 percent of the variance, while Factors B, C, D, and E accounted for 6.8 percent, 5.9 percent, 4.6 percent, and 4 percent of the variance, respectively.

For the present study, however, all five factors were retained and five trait scores were computed by averaging each rater's responses to those items which loaded most heavily on

[1] As would be expected in a field study, certain events made it impossible for a "pure" random sampling across the board. In specific, a second random assignment was necessary when the original random assignment of an observer to a particular section conflicted with the observer's own schedule. This difficulty only highlights further a disadvantage of using *in vivo* observers.

[2] One observation session was completely cancelled owing to the professor's illness. At the professor's request, there was no rescheduling of the observation session. Thus, a total of 11 rather than 12 observation sessions were completed.

[3] For ease of data entry only, scores on the lecture evaluation form were converted from a 1 to 6 scale from the original -3 to +3 scale.

TABLE 1

VARIMAX ROTATED FACTOR MATRIX

	Questionnaire Item	A	B	C	D	E	h^2
No. 5	Lectures at comfortable speed	.71	.24	.38	.10	.28	.80
No. 14	Uses time efficiently	.66	.31	.18	.16	.09	.59
No. 3	Emphasizes conceptual understanding	.64	.20	.09	.41	.33	.73
No. 22	Explains clearly	.61	.19	.29	.34	.25	.67
No. 23	Overall evaluation of lecture	.59	.48	.33	.30	.30	.87
No. 19	Is well prepared	.53	.41	.37	.45	.11	.80
No. 4	Varies speed and tone of voice	.09	.73	.05	.13	.40	.71
No. 16	Interesting style of presentation	.25	.71	.29	.23	.08	.71
No. 15	Appears enthusiastic about subject	.50	.68	.30	.14	.17	.84
No. 17	Seems to enjoy teaching	.38	.63	.41	.35	.10	.84
No. 24	Overall evaluation as teacher	.52	.58	.34	.23	.14	.79
No. 18	Appears self-confident	.50	.56	.10	.45	-.01	.77
No. 8	Students appear interested	.38	.45	.44	.19	.36	.70
No. 13	Encourages class discussion	.28	.12	.85	.18	.03	.85
No. 10	Invites questions	.29	.34	.82	.11	.15	.91
No. 12	Answers questions well	.31	.20	.59	.41	.34	.77
No. 6	Has good eye contact	.45	.44	.52	.15	.18	.72
No. 1	Gives lecture perspective	.00	.14	.50	.30	.24	.42
No. 11	Summarizes main points	.32	.23	.29	.77	.10	.84
No. 21	Discusses other points of view	.12	.15	.16	.56	.37	.51
No. 20	Identifies what is important	.43	.35	.39	.47	.11	.69
No. 9	Lecture is organized	.37	.30	.19	.44	.32	.56
No. 2	Presents origins of ideas and concepts	.18	.17	.12	.16	.72	.62
No. 7	Writes clearly on blackboard	.45	.16	.42	.22	.60	.81
	EIGEN VALUE	13.79	1.19	1.03	0.80	0.70	
	PERCENTAGE OF VARIANCE	78.70	6.80	5.90	4.60	4.00	

each of the five factors. These heavy loaders are the items with boxed-in loadings in Table 1. For example, Trait A was computed by averaging responses to the items which loaded most heavily on Factor A (items 5, 14, 3, 22, 23, and 19). Trait B consisted of the average response to Factor B items (4, 16, 15, 17, 24, 18, and 8). Trait C consisted of the average response to items 13, 10, 12, 6, and 1. Trait D consisted of the average response to items 2 and 7. Given the distribution of the factor loadings among the items of the LEF, the traits were tentatively labeled as speed, clarity, and conceptual emphasis of lecture (Trait A), enthusiasm and self-confidence (Trait B), interaction with students (Trait C), organization of lecture (Trait D), and use of references (Trait E).

RESULTS

Table 2 presents the Multitrait-Multimethod Matrix (MTMM) for the five traits, each measured by the five different methods.

Convergent Validity

Convergent validity refers to the extent to which different methods agree on their assessment of the same trait. Two criteria can be used to provide evidence for convergence validity.

Criterion 1. From an MTMM approach, convergent validity is demonstrated by showing that the monotrait-heteromethod correlations (the italicized validity diagonals in Table 2) are statistically significant. Table 3 presents pairwise comparisons of these convergence coefficients.

It can be seen in Table 3 that the greatest degree of convergent validity was found among the *in vivo* student, videotape peer prediscussion, and videotape postdiscussion methods. All 15 of these relevant coefficients were statistically significant at $p < .05$ and ranged from .76 to .97 with a median value of .93. It is interesting to note the particularly high degree of convergence between the *in vivo* student and the videotape peer prediscussion methods. *None* of

TABLE 2

MULTITRAIT-MULTIMETHOD MATRIX[a]

METHOD/TRAIT	A_1	B_1	C_1	D_1	E_1	A_2	B_2	C_2	D_2	E_2	A_3	B_3	C_3	D_3	E_3	A_4	B_4	C_4	D_4	E_4	A_5	B_5	C_5	D_5	E_5	Mean	sd
In Vivo Peer																											
A_1																										3.69	1.79
B_1	97																									3.53	1.78
C_1	97	93																								3.99	1.92
D_1	97	93	95																							3.57	1.73
E_1	81	83	84	74																						3.67	2.34
In Vivo Student																											
A_2	*73*	*73*	69	70	59																					4.57	1.49
B_2	73	*73*	67	70	59	99																				4.43	1.51
C_2	68	66	*66*	66	55	98	97																			4.73	1.53
D_2	71	69	69	*67*	59	99	98	99																		4.64	1.50
E_2	70	67	64	66	*55*	99	98	98	99																	4.33	1.41
Videotape Peer (Prediscussion)																											
A_3	*56*	59	55	53	61	*94*	95	95	95	94																4.03	1.42
B_3	62	*64*	57	58	61	93	*96*	94	94	93	98															3.90	1.43
C_3	57	57	*58*	50	63	89	88	*92*	92	88	96	95														4.00	1.47
D_3	62	64	61	*57*	65	96	95	96	*97*	95	99	97	97													4.19	1.45
E_3	76	75	75	72	*75*	96	96	95	97	*95*	94	93	91	96												3.82	1.31
Videotape Peer (Postdiscussion)																											
A_4	*45*	45	36	40	49	*76*	83	75	76	79	*85*	85	77	83	77											3.65	1.83
B_4	56	*57*	46	50	52	88	*93*	87	87	89	93	*95*	86	91	86	95										3.70	1.52
C_4	41	38	*45*	34	47	70	71	*77*	76	70	81	82	*91*	81	72	66	75									3.77	1.72
D_4	58	57	51	*48*	59	86	92	88	*90*	84	93	93	91	*93*	91	91	94	80								3.87	1.53
E_4	75	73	70	70	*67*	96	96	95	97	*96*	93	93	89	94	*92*	80	89	75	94							4.05	1.41
Students' End-of-Semester Evaluations																											
A_5	*09*	20	06	01	54	*06*	13	02	05	04	*28*	29	27	27	20	*43*	34	18	28	11						4.41	1.50
B_5	05	*15*	01	05	44	02	*07*	07	04	04	19	*22*	16	17	10	43	*31*	10	24	04	98					4.33	1.61
C_5	01	13	*01*	05	48	03	09	*01*	03	01	28	28	*26*	25	17	38	29	*18*	22	05	98	95				4.55	1.49
D_5	05	16	05	*03*	54	07	13	05	*07*	05	32	31	32	*30*	22	42	33	23	*29*	11	99	94	99			4.51	1.47
E_5	02	14	02	06	*52*	03	09	00	03	*02*	27	26	26	26	*18*	39	29	16	25	*07*	99	95	99	99		4.28	1.41

[a]Decimal points have been omitted. Correlations greater than .54 are significant at $p < .05$. Heterotrait-Monomethod triangles are enclosed by solid lines. Heterotrait-Heteromethod triangles are enclosed by broken lines. Monotrait-Heteromethod coefficients (validity diagonals) are italicized.

TABLE 3

CONVERGENT VALIDITY COEFFICIENTS

Methods Converged[a]	Lowest	Highest	Median
VPPR with IVS	.92	.97	.95
VPPR with VPPO	.85	.97	.93
VPPO with IVS	.76	.96	.90
IVS with IVP	.55	.73	.67
VPPR with IVP	.56	.75	.58
VPPO with IVP	.45	.67	.48
VPPR with SEOS	.18	.30	.26
VPPO with SEOS	.07	.43	.29
IVP with SEOS	.01	.52	.09
IVS with SEOS	.01	.07	.06

[a]VPPR = Videotape Peer Prediscussion Method.
VPPO = Videotape Peer Postdiscussion Method.
IVS = *In Vivo* Student Method.
IVP = *In Vivo* Peer Method.
SEOS = Students' End-of-Semester Evaluation Method.

of the convergent validity coefficients between these methods fell below .92.

A significant but somewhat lower degree of convergent validity was found among the *in vivo* peer, *in vivo* student, and videotape peer prediscussion methods. Again, all 15 of these convergent validity coefficients were statistically significant and ranged from .55 to .97 with a median value of 0.73.

Although both the *in vivo* peer and the videotape peer postdiscussion methods showed high convergence with the *in vivo* student and the videotape peer prediscussion methods, they showed only marginal convergence with one another. Only two of their convergence coefficients were statistically significant. The range was .45 to .67 with a median value of .48.

It is interesting and noteworthy that the students' end-of-the-semester evaluations method showed essentially no convergence with any of the other four methods (the range of the 20 relevant coefficients was from .01 to .43 with a median value of .15, all $p > .05$).

Criterion 2. Convergent validity can also be expressed by showing that the mean ratings obtained by one method do not significantly differ from the mean scores obtained by another method. Table 4 presents the results of a series of correlated t-tests performed on the trait means. It can be seen from Table 4 that the *in vivo* student ratings were significantly higher than ratings obtained from three peer (faculty) methods in 13 out of 15 comparisons. While this finding does not enhance the convergent validity of the *in vivo* student method with the peer methods, it is important to keep in mind that the relative ratings as expressed by the validity diagonals strongly argue in favor of convergent validity. That is, although the *in vivo* students and the *in vivo* faculty peers differed in their assignment of absolute scale values (as shown by the results in Table 4), they did agree highly on the direction and relative size of scale value shifts (as shown by the results in Table 3).

Discriminant Validity

Discriminant validity refers to the degree to which a trait can be differentiated from other traits. In the MTMM approach, three criteria can provide evidence for discriminant validity.

Criterion 1. First, a trait should correlate higher with itself over two methods than with another trait measured by the same two methods. This involves computing the proportion of times that each of the monotrait-heteromethod coefficients exceeds the eight corresponding heterotrait-heteromethod coefficients (Ostrom 1969). For example, in Table 2 the A_2A_1 coefficient should be greater than A_2B_1, A_2C_1, A_2D_1, A_2E_1, B_2A_1, C_2A_1, D_2A_1, E_2A_1. Table 5 presents the results of this criterion check. The extent to which the methods met this criterion was not altogether impressive. The highest proportions were found for the videotape peer prediscussion and the videotape peer postdiscussion methods (.6000 and .6125, respectively), while the lowest proportion was found for the *in vivo* peer method (.5000). The average of all proportions was .5550.

Criterion 2. Secondly, a variable should correlate more highly with an independent effort to measure the same trait than with measures designed to get at different traits with the same method. This involves computing the proportion of times that a trait's value in its validity diagonal exceeds its eight values in the two corresponding heterotrait-monomethod triangles. For example, in Table 2 the A_2A_1 coefficient should exceed B_1A_1, C_1A_1, D_1A_1, E_1A_1, B_2A_2, C_2A_2, D_2A_2, E_2A_2. Table 5 shows that none of the methods passed this test, although it is encouraging that the rank order of the proportions for this criterion is identical to the rank of the proportions for Criterion 1.

TABLE 4
DIFFERENCES AMONG METHODS ON THE TRAIT MEANS

	Method	Mean[a]	SD	II	III	IV	V
Trait A: "Speed, Clarity, and Conceptual Emphasis"							
I.	*In vivo* peer	3.69	1.79	.881[b]	.344[b]	.038[b]	.724
II.	*In vivo* student	4.57	1.49		.537[b]	.918[b]	.157
III.	Videotape peer prediscussion	4.03	1.42			.382	.379
IV.	Videotape peer postdiscussion	3.65	1.83				.761
V.	Students' end of semester	4.41	1.50				
Trait B: "Enthusiasm and Self-Confidence"							
I.	*In vivo* peer	3.53	1.78	.896[b]	.363[b]	.168[b]	.794
II.	*In vivo* student	4.43	1.51		.533[b]	.729[b]	.102
III.	Videotape peer prediscussion	3.89	1.43			.196	.430
IV.	Videotape peer postdiscussion	3.70	1.52				.626
V.	Students' end of semester	4.33	1.61				
Trait C: "Interaction with Students"							
I.	*In vivo* peer	3.99	1.92	.739	.013[b]	.218[b]	.566
II.	*In vivo* student	4.73	1.53		.726[b]	.958[b]	.173
III.	Videotape peer prediscussion	4.00	1.47			.231	.553
IV.	Videotape peer postdiscussion	3.77	1.72				.784
V.	Students' end of semester	4.55	1.49				
Trait D: "Organization of Lecture"							
I.	*In vivo* peer	3.57	1.73	1.070[b]	.617[b]	.296[b]	.943
II.	*In vivo* student	4.64	1.49		.453[b]	.774[b]	.127
III.	Videotape peer prediscussion	4.19	1.45			.321	.326
IV.	Videotape peer postdiscussion	3.87	1.53				.647
V.	Students' end of semester	4.51	1.47				
Trait E: "Use of References"							
I.	*In vivo* peer	3.67	2.34	.667	.155[b]	.388[b]	.608
II.	*In vivo* student	4.33	1.41		.511[b]	.279[b]	.058
III.	Videotape peer prediscussion	3.82	1.31			.232[b]	.453
IV.	Videotape peer postdiscussion	4.05	1.41				.221
V.	Students' end of semester	4.28	1.41				

[a]Trait scale value range = 1 to 6.
[b]$p < .05$ by correlated t-test, df = 11.

TABLE 5
DISCRIMINANT VALIDITY COEFFICIENTS

Method	Criterion 1[a]	Criterion 2[b]
Videotape peer postdiscussion	.6125	.2580
Videotape peer prediscussion	.6000	.2250
In vivo student	.5563	.1500
Students' end-of-semester evaluations	.5063	.0000
In vivo peer	.5000	.0000
Mean[c]	.5550	.1266

[a]Criterion 1 represents the proportion of times that each of the monotrait-heteromethod coefficients exceeded the eight corresponding heterotrait-heteromethod coefficients in Table 2.
[b]Criterion 2 entries represent the proportion of times that a trait's value in its validity diagonal in Table 2 exceeded its eight values in the two corresponding heterotrait-monomethod triangles.
[c]For each criterion 400 comparisons were necessary (50 monotrait-heteromethod coefficients times eight comparison coefficients).

Criterion 3. A third way to assess discriminant validity is to show the same pattern of trait relationships in all of the heterotrait triangles (those triangles in Table 2 which are enclosed by either broken or solid lines). A noted absence of a consistent pattern of trait relationships was found in Table 2.

DISCUSSION

In general, the results provided evidence for the convergent validity of the *in vivo* student, *in vivo* peer, and both videotape peer methods. Several exceptions, though, must be taken into account. First, student ratings, particularly when obtained *in vivo*, were consistently higher than the three sets of faculty ratings. However, positive coefficients among these ratings indicated that there was significant convergence

on the direction and relative strength of trait ratings by these methods.

Secondly, the *in vivo* peer and the videotape peer postdiscussion methods showed only marginal convergence with one another. However, both methods converged well with all other methods excepting the students' end-of-the-semester evaluations method.

Third, the students' end-of-the-semester evaluations method did not converge with any of the other four methods. One possible explanation is that the students were asked to evaluate the teaching behavior of the professor over the entire semester. By way of contrast, students and faculty in the other four methods were instructed to give ratings relevant to the specific session being observed and taped. Of course, another confounding variable is simply time itself, because the end-of-semester evaluations were collected about three weeks after the *in vivo* ratings. This pesky time-related phenomenon is known in the literature as "observer drift."

Finally, the discriminant validity was terribly low for all methods involved, although the videotape methods fared better than did the other methods. While obvious method variances may have contributed to the low discriminant validity coefficients, a more palatable explanation is that traits B, C, D, and E were only weakly captured by the LEF questionnaire (refer to their eigen values in Table 1). Perhaps none of the methods had sufficient discriminatory power to pull out these weaker traits.

Owing to resource constraints, several interesting and important questions were unaddressed by the present study. One of these concerns the effects of being obtrusively observed whether it be *in vivo* or by video camera. Obtrusive observation was held constant in the present study. That is, all subjects, regardless of the observation method, were aware that they were being observed. Some researchers feel that the observer exerts little influence on teaching behavior because a teacher cannot do what (s)he cannot do regardless of the level of observation method (cf. Samph 1976). But as Samph (1976) pointed out, it is equally plausible to hypothesize that direct observation may inhibit what one can do.

Without question, a number of social psychological theories thrive on the notion that people behavior differently under direct observation. One of the forerunning theories in this area is objective self-awareness theory (Duval and Wicklund 1972) which contends, among other things, that self-focused attention (as induced by the awareness of being observed) can increase the salience of the norms for correct, nondeviant social behavior. Thus, putting one's "best foot forward" may be more probable under enhanced self-focused attention. Unfortunately, there seems to be little or no empirical research on whether self-focused attention effects are greater when *in vivo* observers are used or when video cameras are used as inductors of self-focused attention. Furthermore, since the present study involved the simultaneous use of both video cameras and live observers in each session (in order to minimize and control for differences due to shifts in classrooms, time, and student-teacher moods), the question of whether either method alone produces differential behavior change is not answerable.

Another factor which was held constant in the present study was that of forewarning. All participants were forewarned at least one week in advance that their teaching behavior would be observed. Research on the effects of forewarning appears to be limited to persuasion theory (Petty, Ostrom, and Brock 1981). Little or no empirical research exists on the effects of forewarning on teaching behavior evaluations. Nor was it a purpose of the present study to examine such effects. A reasonable hypothesis, however, is that the professors prepared more diligently for the sessions in which they were to be observed. However, since forewarning was held constant across all conditions, it is probable that the behavioral and cognitive changes which may have resulted from being forewarned were equivalent across all conditions. In essence, it is possible that each professor was better prepared than usual, but that this preparedness did not vary across conditions. This would serve to enhance the internal validity of the present findings. Furthermore, recall Samph's (1976) earlier comment that one cannot do what one cannot do.

CONCLUSION

Although many other intriguing questions about methods of evaluating teaching behavior remain, the primary purpose of the present study was achieved. The construct validity of

five different methods of measuring teaching behavior was assessed. The videotape rating methods of the present study produced convergent and discriminant validity coefficients which were greater than or equal to those derived from the traditional *in vivo* peer and *in vivo* student methods. At the very least, these data argue for the serious consideration of using the videotape method in lieu of a live observer to assess teaching behavior.[4]

[4]The results of this study have, in fact, already been instrumental in bringing about approval from the CSUN administration for substituting a camera for a live observer, *at the discretion and option of the professor being observed.*

REFERENCES

Campbell, Donald T. and D. W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56, 81-105.

Campbell, Donald T. and J. C. Stanley (1963), *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally.

Chandler v. Florida (1981), 101 S.Ct. 802, January.

Dowrick, Peter W. and Simon J. Biggs (1983), *Using Video*, New York: Wiley.

Duval, Shelley and Robert A. Wicklund (1972), *A Theory of Objective Self-Awareness*, New York: Academic Press.

Goodman, P., J. Furcon, and J. Rose (1969), "Examination of Some Measures of Creative Ability by the Multitrait-Multimethod Matrix," *Journal of Applied Psychology*, 53, 240-243.

Harrop, L. A. (1979), "Unreliability of Classroom Observation," *Educational Research*, 21, 207-211.

Kassin, Saul M. (1984), "TV Cameras, Public Self-Consciousness, and Mock Juror Performance," *Journal of Experimental Social Psychology*, 20, 336-349.

Kent, Ronald N. and Sharon L. Foster (1977), "Direct Observational Procedures: Methodological Issues in Naturalistic Settings," in *Handbook of Behavioral Assessment*, A. R. Ciminero, K. S. Calhoun, and H. E. Adams, eds., New York: Wiley.

Kessler, Joan B. (1979), "An Overview of Research on the Use of Videotape in the Courtroom and in Legal Education," paper presented to the Western Speech Communication Association, Los Angeles.

———— (1981), "An Overview of Research on the Use of Videotape in the Courtroom for Expert Witnesses," paper presented to the American Academy of Forensic Sciences.

Meadowcroft, Pamela and Roy Moxley (1980), "Naturalistic Observation in the Classroom: A Radical Behavioral View," *Educational Psychologist*, 15, 23-34.

Marcatoris, Michael and W. Edward Craighead (1974), "Effects of Nonparticipant Observation on Teacher and Pupil Classroom Behavior," *Journal of Educational Psychology*, 66, 512-519.

Miller, Gerald R. (1975), "Jurors' Responses to Videotaped Trial Materials: Some Recent Findings," *Personality and Social Psychology Bulletin*, 1, 561-569.

Ostrom, Thomas M. (1969), "The Relationship Between the Affective, Behavioral and Cognitive Components of Attitude," *Journal of Experimental Social Psychology*, 5, 12-30.

Petty, Richard E., Thomas M. Ostrom, and T. C. Brock (1981), *Cognitive Responses in Persuasion*, Hillsdale, NJ: Lawrence Erlbaum.

Samph, Thomas (1976), "Observer Effects on Teacher Verbal Classroom Behavior," *Journal of Educational Psychology*, 68,736-741.

APPENDIX A

LECTURE EVALUATION FORM

Name of Professor Being Evaluated _____
Class _____
Time _____
Date _____
Evaluator _____

Instructions to Evaluator:

Select a plus number for phrases that you think accurately describe the professor, lecture, or class. The more accurately you think the phrase describes it, the larger the plus number you would use. Select a minus number for phrases you think do not describe it accurately. The less accurately you think a phrase describes it, the larger the minus number you would choose. Therefore, you can select any number from +3, for phrases that you think are very accurate, to -3, for phrases you think are very inaccurate descriptors.

-3	-2	-1	Gives lecture *perspective* by indicating what has been discussed and what will be discussed	+1	+2	+3
-3	-2	-1	Presents origins of ideas and concepts	+1	+2	+3
-3	-2	-1	Emphasizes conceptual understanding	+1	+2	+3
-3	-2	-1	Varies the speed and tone of voice	+1	+2	+3
-3	-2	-1	Lectures at comfortable speed	+1	+2	3+
-3	-2	-1	Has good eye contact with students	+1	+2	+3
-3	-2	-1	Writes clearly on blackboard (if applicable)	+1	+2	+3
-3	-2	-1	Students appear interested in lecture	+1	+2	+3
-3	-2	-1	Lecture is organized	+1	+2	+3
-3	-2	-1	Invites questions	+1	+2	+3
-3	-2	-1	Summarizes main points	+1	+2	+3
-3	-2	-1	Answers questions well (if applicable)	+1	+2	+3
-3	-2	-1	Encourages class discussion	+1	+2	+3
-3	-2	-1	Uses time efficiently	+1	+2	+3
-3	-2	-1	Appears enthusiastic about his/her subject	+1	+2	+3
-3	-2	-1	Has interesting style of presentation	+1	+2	+3
-3	-2	-1	Seems to enjoy teaching	+1	+2	+3
-3	-2	-1	Appears self-confident	+1	+2	+3
-3	-2	-1	Is well-prepared	+1	+2	+3
-3	-2	-1	Identifies what he/she considers important	+1	+2	+3
-3	-2	-1	Discusses points of view other than his/her own	+1	+2	+3
-3	-2	-1	Explains clearly	+1	+2	+3
-3	-2	-1	OVERALL EVALUATION OF LECTURE	+1	+2	+3
-3	-2	-1	OVERALL EVALUATION OF PROFESSOR AS TEACHER	+1	+2	+3

Comments

On the back side of this page, please feel free to comment on the lecture you have witnessed. You are particularly encouraged to expound upon or clarify the ratings you have given to the professor/lecture. In addition, it would be helpful if you would offer constructive suggestions on how the professor might improve his/her teaching. Thank you for your cooperation.