Research in Higher Education, Vol. 36, No. 2, 1995

THE EFFECT OF IMPLICIT THEORIES ON RATERS' INFERENCE IN PERFORMANCE JUDGMENT: CONSEQUENCES FOR THE VALIDITY OF STUDENT RATINGS OF INSTRUCTION

Nand Kishor

Performance judgment is a situation of incomplete information where raters' inference would play an important role. Consequently, the schematic nature of human cognition may introduce implicit personality theory bias in performance judgment. To demonstrate this, a causal model of performance rating judgment was framed from the theories of person perception and social cognition. The model yielded a good fit to the data obtained from a performance rating task where the availability of performance information was manipulated. The results supported the hypotheses that student raters' inferences are partly contaminated by their implicit theories of a good instructor. Student raters inferred traits and behaviors and provided ratings for corresponding items even when the instructor behavior was limited to a subset of performance data only. The findings imply that one aspect of invalidity in student ratings of instructors is the bias in human inference due to the implicit theories of effective instructional behavior.

Student ratings of instructors' teaching performance in colleges and universities are quite common. In addition to providing feedback to the instructor, these ratings are widely used in making tenure, promotion, and merit pay decisions (McCallum, 1984). Given these purposes, it is necessary that student ratings be reliable, valid, and accurate. Increasing the reliability and validity of student ratings has been the focus of many psychometrically driven investigations. Researchers have manipulated the content and format of the rating scales to find the best method of instrumentation. As a result, there are instruments that yield high reliability. However, findings on the validity of student ratings are perplexing (Doyle, 1981; Feldman, 1989; Gaski, 1987; Howard and Maxwell, 1982; Marsh, 1987). The present study sought to examine the issue of

Nand Kishor, Ph.D., Dept. of Educational Psychology & Special Education, The University of British Columbia, 2125 Main Mall, Vancouver V6T 1Z4.

0361-0365/95/0400-0177$07.50/0

validity within a cognitive information processing perspective. The focus was on the students rather than the instrumentation.

Previous research on the validity of student ratings has mainly been conducted within the psychometric tradition where the focus has been on rating outcomes. Construct validation through factor analysis has been used more commonly than other techniques, and the factorial validity of some rating instruments is quite impressive. For example, the same nine factors in Student Evaluation of Educational Quality (SEEQ) are often obtained in many contexts (Marsh, 1984). However, the factorial validity of rating instruments as evidence for the validity of students' rating is not without problems. The critics have argued that the factor structures also reflect students' implicit theories (Whitely and Doyle, 1976), or the semantic similarity of the items in the scale (Cadwell and Jenkins, 1985). Comparatively, there are fewer studies of the predictive validity of student ratings, and the ones that exist have several methodological limitations (see Abrami, d'Apollonia, and Cohen, 1990). Researchers have also investigated the validity of student ratings by comparing student ratings with that of the instructor, instructor's peers, former students, and administrators. Establishing validity of student ratings against the ratings of peers and former students is also problematic (see Feldman, 1989, pp. 163–166).

In most of the well-developed rating scales (e.g., SEEQ), Likert-type behavioral items with adjectival and numerical anchors are quite common. Behavioral items are popular on the assumption that these provide a more concrete behavior to be rated, resulting in performance ratings least biased by personality and other factors. However, psychometrically oriented studies have largely ignored the fact that performance judgment is a situation of incomplete information. In situations of incomplete data, part of the judgment depends on the raters' ability either to infer or to recall performance behavior. Consequently, performance ratings may not be free from the pervasive shortcomings of human inference (see Nisbett and Ross, 1980; Holland et al., 1986).

Cognitive processes of attention, storage, retrieval, and integration are involved in producing a numerical rating (Anderson, 1981). The rating judgments may only be valid to the extent a rater follows rational procedures in forming that judgment (cf. Simon, 1978). Thus, the validity of students' ratings will also depend on their ability to make inferences when they complete the performance rating instruments. Their inferences may be biased. Bias in judgment has been a major threat to the validity of student ratings (e.g., the Dr. Fox effect), but previous research in this area has been described as a "witch hunt" because researchers have not usually provided any operationalization of bias (Marsh, 1987, chapter 5). In the present study, a theoretical framework of cognitive processing was posited to investigate inferential bias in student ratings.

INFORMATION PROCESSING PERSPECTIVE

In any performance judgment, the raters themselves are one source of bias or error. This conception is common in the current information processing models of performance evaluation (Cooper, 1981; DeNisi, Cafferty, and Meglino, 1984; Ilgen and Feldman, 1983). In these models, performance rating is considered a specific instance of person perception, which may be biased in the direction of the raters' implicit personality theory of the occupation being rated. It is suggested that biases in performance rating result from the operation of implicit personality theories that influence the cognitive processing of performance information. Consequently, psychometrically sound rating instruments alone are unlikely to provide valid evaluations. Attention to raters' cognition or how they mentally arrive at their ratings may help us identify some causes of bias in performance ratings (Landy and Farr, 1980).

Although the influence of raters' implicit theories of an occupation has been acknowledged, little empirical evidence has accumulated regarding the extent to which students' implicit personality theories of teaching (IPT) bias their ratings of instructors. In an earlier study, Whitely and Doyle (1976) claimed that the factorial validity of the rating instruments was due to the commonly held IPT by college students. In their study, one sample of students evaluated teaching performance using a rating instrument, and another sample of students categorized the 26 teaching characteristics used in the instrument. Both rating and categorization data were separately factor analyzed, and the resulting factors were found to be congruent, implying the existence of shared implicit theories of teaching. Larson (1979) also provided evidence showing that students have a shared implicit theory of instructor behaviors. Marsh (1984) suggested that students' IPT contains behavioral covariation of instructional effectiveness dimensions. These findings were further elaborated by Cadwell and Jenkins (1985) who argued that the congruence between factors resulted from the semantic similarity between the items in the scale. Cadwell and Jenkins claimed that the robust factor structure of rating instruments results from synonymous items researchers include in the rating scale. They also implied that factor structures of rating instruments reflect students' IPT of teaching and not necessarily the actual covariation in an instructor's teaching behaviors.

Marsh and Grove (1987) discounted the semantic similarity argument mainly on the grounds that the instructors' and students' ratings are usually quite similar with respect to the factor structure of SEEQ. However, factorial similarity based on correlations implies only relative similarity, not absolute similarity, and comparing students' ratings with the instructor's own ratings assumes independence of the two sets of ratings, which hardly seems to be the case (Feldman, 1989, pp. 163–166). Moreover, recent research on the role of semantics

and language appears to support the semantic similarity argument. For example, Carlson and Mulaik (1993) suggested that language permeates the rating process, and is the basis for the stability and regularity of the rating process.

Although Marsh (1987, chapter 5) advocated the construct validation approach through factor analysis, he did not consider the importance of raters' cognition. The importance of cognitive processes underlying performance rating judgments has been suggested by several theorists and researchers (Cooper, 1981; DeNisi, Cafferty, and Meglino, 1984; Ilgen and Feldman, 1983; Jenkins, 1987). If ratings are based on students' IPT to any degree, then how the behavioral covariation in students' IPT is mapped on their ratings needs to be understood. There is a lack of evidence elucidating the nature of any causal relationships between the behavioral covariation in students' IPT and their ratings. An alternative to what Marsh (1987) labeled as "witch hunt" would be to investigate the specific psychological mechanisms involved in the production of the ratings. Literature on person perception can be drawn upon to postulate a model of how students' IPT influences their ratings of an instructor's teaching.

Schematic Processing

A person's implicit theories operate through schematic processing. As Taylor and Crocker (1981) have suggested, schemas are inherent in human cognition. Schemas control attention, encoding, storage, recall, and evaluation of information. Schemas are of three types: person schema, event schema, and role schema. Person and role schemata are quite applicable in performance evaluation. The organization of schemas is hierarchical: behavioral information being more concrete is subordinate to traits being more abstract. Traits being superordinate tend to influence the perception of behavioral information (Hastie, 1981) and permit extensive inferences about persons (Wyer and Martin, 1986). Therefore, in situations of limited information, such as in performance evaluation, schemas or the implicit personality theories of the target occupation would enable the raters to infer the unavailable information.

Recent studies suggest that the IPT have an action-oriented mental representation, where specific actors have specific goals (Trzebinski et al., 1985). Person schemas represent actors; role schemas represent goals. For example, in a person's schema of teachers, *principal* would be an actor category and carrying out supervisory activities would be a category of goals associated with the actor. The trait *leadership ability* would be an important condition for realizing the goal of supervision. So, the implicit personality schemas not only contain stereotypical traits but also the corresponding role behaviors. And because behavioral and trait information are encoded together (Cantor and Mischel, 1979; Trzebinski, 1985; Trzebinski et al., 1985), either trait or behavioral data can activate or prime the rater's IPT.

In the evaluation of teaching performance, students' attention to an instructor's traits or personality factors would be automatic and uncontrollable. Once the IPT is activated, it would automatically influence the evaluation of both behavioral and trait data, leading students to compare the instructor for a "goodness-of-fit" against their own subjective IPT of a good instructor. Where performance information is obscure or unavailable, IPT would facilitate the inferences. Consequently, the ratings on behavioral dimensions would not be free of inferential bias resulting from raters' perception of the instructor's personality. Even if personality information is unavailable, it would be implied by the available behavioral data, because behaviors and traits are naturally covarying dimensions in the IPT. Moreover, good teaching depends both on effective role behaviors (teaching methods) and desirable traits (Medley, 1979). The traits being superordinate will tend to affect the ratings of behavior. It seems that the manner in which the IPT functions would bias all ratings of instructor behavior obtained even on the most reliable instrument.

Within the schematic processing framework, Krzystofiak, Cardy, and Newman (1988) examined the influence of IPT on ratings of performance behavior. These researchers found that traits attributed to a ratee were a function of ratee behaviors, and the inferred traits influenced the perception of those behaviors, and vice versa. Their results showed a mutual association between traits and behaviors; they did not explore causal relations. Raters in their study inferred traits only, although in performance appraisals behaviors may be inferred as well.

PRESENT STUDY

The schematic processing framework provided hypotheses to examine the influence of IPT on performance ratings. It suggested an operative model of cognitive strategies by which performance ratings appear to be actually formed in a rater's mind. The general conjecture was of causal effects at the level of constructs underlying ratings for items corresponding to available and unavailable performance information. This hypothesized causal model is presented in Figure 1. Each construct was measured through multiple rating items. The unidirectional path γ represents the causal influence of available behavioral information on implied traits. It represents the hypothesis (1) that the available behavioral information about an instructor activates raters' person schemata in terms of traits. Path β is also unidirectional, showing the causal influence of traits or person schemata on inferred behavior. It represents the hypothesis (2) that person schema or traits of an effective instructor influences inferred behaviors when data regarding such behaviors are required but unavailable to the rater.

In construct validation research where validity of a rating instrument is estab-

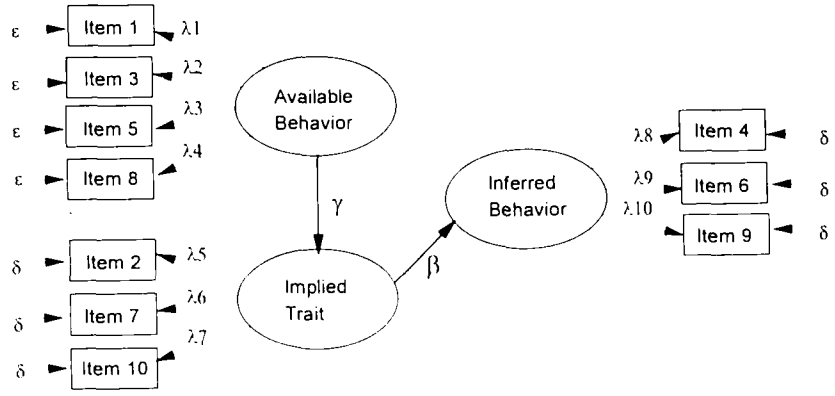

FIG. 1. Hypothesized operating model.

lished through factor analysis, it is assumed that students have implicit behavioral covariation of teaching effectiveness (Marsh, 1984). Carefully developed rating instruments usually omit items related to personality traits, yet reflect an underlying factor structure. Consequently, a competing causal model can also be formulated. It is quite plausible that the basis for inferred behavior could be the behaviors that are available. Ratings of inferred behavior may not be based on inferred traits but on the available behaviors. Thus, a plausible competing model is shown in Figure 2. This model was also estimated with the same measurements as for the theorized model in Figure 1.

Research Strategy

There are three ways to test hypotheses in performance evaluation research. An actual instructor evaluation could have been used, but this procedure would

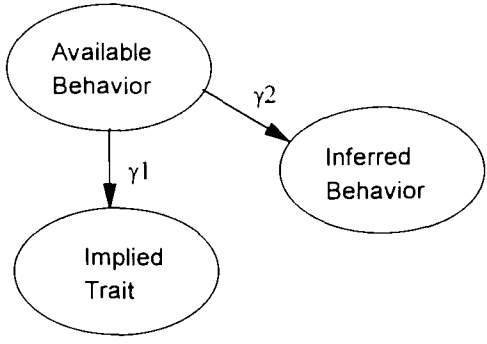

FIG. 2. Competing model.

have gained external validity at the expense of internal validity of the study. Such procedures have been used commonly in the psychometric tradition of construct validation, where the interest is in the final ratings and not in how the ratings are arrived at mentally. Use of simulated videotapes was another possibility, but this methodology would also have had more external validity at the expense of internal validity. In an actual performance evaluation, it is difficult to control for the "leniency effect," "Dr. Fox effect" (i.e., effect of gestures, appearance, etc), and the "grading satisfaction effect." The first two of these effects would also prevail in the evaluation of short simulated teaching presented on videotapes. All of these effects were potential threats to the internal validity of the present study.

The question of internal and external validity of the study should be raised in relation to the purpose of the study (Mook, 1983). For the purpose of testing hypotheses and inferring causality in raters' information processing in the present study, it was necessary to manipulate the nature and amount of information available to the student raters. Consequently, a written vignette of an instructor was most suitable as the stimulus. The vignette enabled manipulation of information and provided constancy of stimulus across all raters. As well, a vignette can provide a summary of instructional episodes very much like what the students may actually retrieve from their memories about an instructor during an actual evaluation. It should be noted that an actual evaluation by students is not done when the instructor is "in action." At some institutions, it is required that the instructor be absent from the classroom when students complete the rating instrument. Consequently, the students usually have to retrieve a lot of the instructor's behavior from memory. Summary descriptions similar to the vignette may result from students' memories if students are asked to provide a qualitative evaluation of an instructor. Nevertheless, the results should be interpreted in relation to the purpose and the constraints on the design of the study.

The causal models were tested on performance rating data via linear structural equation modeling procedures. In testing causal models, elimination of competing models is just as important as confirming the hypothesized model (Breckler, 1990). A competing model based on research literature was proposed (Figure 2). Other competing models relating the three constructs were eliminated by features of the design. The performance rating task was such that certain symmetrical relationships (bidirectional paths) could be ruled out. Performance data were manipulated in terms of availability, just as some performance data are usually unavailable to students in actual appraisal situations. All trait and part of behavioral data were withheld from the student raters. Consequently, it was logically impossible for implied traits or inferred behavior to have a causal effect on ratings of available behavior. Nor was it possible for inferred behavior to have a causal effect on implied traits. Therefore, the basis of causal interpretation is not only the LISREL analysis, but also the design of the rating task where some data were purposefully withheld. Causal interpreta-

tion ultimately depends on manipulation and control of independent variables (Cliff, 1983). In the absence of some needed data, the raters' IPT must operate causally, enabling them to impute the missing information. Unless the student raters' IPT was operative, they could not have rated the items indicating implied trait and inferred behavior.

METHOD

Subjects

The subjects were undergraduates at a university where students' evaluation of instructors' teaching is mandatory, and forms part of the data used in personnel decisions. Therefore, these students had some experience in rating instructors' teaching performance. On a voluntary basis, 222 students participated. These students completed a performance rating task specially designed for this study. Three cases were not included in the final analysis because of multiple ratings on some items, leaving the final sample to 219. This sample size was sufficiently large for LISREL analysis (cf. Bollen, 1989; Hayduk, 1987).

Performance Rating Task

The performance rating task consisted of a vignette followed by a 10-item rating scale. The vignette allowed control of information that was necessary to test the hypotheses predicting causal relations. Only behavioral information regarding the teaching performance of a hypothetical college instructor was provided in the vignette. No statements were made about the instructor's personality dispositions or traits, and some behavioral information was withheld as well. The information available was incomplete in relation to the 10 rating criteria (items) on which the instructor was to be evaluated. The subjects were neither informed of the information withheld nor forced to rate all of the items. This meant that if ratings were provided for the items for which there was no relevant information in the vignette, those ratings would be imputed by the raters' IPT of an effective instructor. Kishor (1990) used a similar approach to investigate the halo error in performance ratings.

The vignette was evaluated for content. Two individuals independently reviewed the original vignette. They identified any personality descriptions in the vignette and the rating items for which there was lack of relevant information. Discrepancies were found in vignette description. The initial vignette was revised and reviewed again by two different individuals. The final form of the vignette was as follows:

> Professor P teaches courses in research methodology in the faculty of education. P requires students to complete weekly assignments and to take a test at the end of the course. This requirement is included in the course outline given in the first class. The

weekly papers are returned with several comments. In each of Professor P's courses, a few students fail the final test. In the class, P always emphasizes the key concepts by providing many examples. Chalkboard, overhead, and other instructional aids are used regularly as appropriate. The lesson objectives are not presented at the beginning of a class, but at the end of every class key ideas are always summarized. All class time is used up covering new ideas. Students try not to miss a class, because Professor P follows the course outline quite rigidly and always completes the syllabus.

In the 10-item rating scale, items 2, 7, and 10 measured personality dispositions: enthusiasm, leadership ability, and dependability, respectively. These items served as indicators of the implied trait construct because there was no information about these traits in the vignette. Items 4, 6, and 9 measured implied behavior: interaction with students, command of subject matter, and attention to individual differences, respectively. These items served as indicators of the inferred behavior construct because there was no information relevant to these behaviors in the vignette. Items 1, 3, 5, and 8 measured clarity of presentation, feedback on student learning, use of class time, and assessment technique, respectively. These items were the indicators of available behavior construct, and there was information relevant to these items in the vignette. All rating items consisted of 6-point Likert scales with 1 labeled as poor and 6 as excellent.

Procedure

The performance rating task was administered to seven classes during the last 10 minutes of regular class time. The researcher told the students that the purpose of the study was to understand certain characteristics of performance ratings, explained what the task involved, and invited those who were willing to complete the rating task to do so. The students did not show any anxiety in completing the rating task, although they did not have all the information necessary to rate all the items. The average time to complete the task was about 6 minutes. Average participation per class was about 90 percent, and all responses were anonymous. Except the hypothetical vignette, the procedure followed in this study is similar to what happens when students evaluate their actual course instructors: some information is recalled or inferred, ratings are made on Likert scales, typically ratings are completed in 6 to 8 minutes, only one instructor is evaluated at a time, and the ratings are completed when the instructor is not in action.

Data Analysis

The data were analyzed via linear structural equation modeling procedures using PRELIS and LISREL VII (Joreskog and Sorbom, 1989). The responses

on Likert scales are strictly ordinal measures and usually produce skewed distributions in performance ratings. The PRELIS program offers the underlying normally distributed interval-level measurement from ordinal data. Therefore, the polychoric covariances obtained from PRELIS were analyzed using the weighted least squares estimation. This approach overcomes the problem of attenuation in correlations and skewness of distributions. It enables inferences permissible of interval measurement from ordinal scales (see Joreskog and Sorbom, 1989, chapter 7). For identification, the variance of the latent constructs were fixed to unity.

Analysis was in two stages as recommended (Anderson and Gerbing, 1988; Bollen, 1989). First, a measurement model was tested to determine whether the rating items measured three distinct latent constructs as purported. This was necessary because the rating scale was specifically constructed for this study and its factor structure was unknown. Then, the structural equations relating the constructs were included in the analysis. Identification at each step is a sufficient condition for the identification of the whole model (Bollen, 1989, p. 328).

Given the literature on the existence of factor structures in instructor rating scales, the more general measurement model was the baseline model for model comparison instead of a null model that implies no correlation among the constructs (cf. Sobel and Bohrnstedt, 1985). Comparing a theoretical model with a measurement model not only provides "an assessment of the fit of the substantive model of interest to the estimated construct covariances, but it also requires the researcher to consider the strength of the explanation of this theoretical model over that of the confirmatory measurement model" (Anderson and Gerbing, 1988, p. 419). Model comparison was accomplished following the decision tree outlined by Anderson and Gerbing (1988).

Multiple criteria were used to assess the fit of the models. The χ^2 test, the LISREL goodness-of-fit index (GFI), the adjusted goodness of fit index (AGFI), and the root mean square residual (RMS) showed the adequacy of the models. A model yields a good fit to the extent that χ^2 is not significant ($p > 0.05$), the GFI and AGFI approach one, and RMS approaches zero. These indexes of fit are only descriptive, and there are no absolute standards for judging the fit of a model. However, for model comparison, hypotheses can be tested by comparing the relative fit of more restricted and less restricted models with sequential χ^2 difference tests (cf. Bollen, 1989; Hayduk, 1987). This comparison is achieved by the statistical significance of a χ^2 test relative to the degrees of freedom (*df*) between the models. If a more restricted model yields a fit that is not significantly worse, that is, if the null hypothesis is not rejected, in comparison to a saturated or measurement model, then on grounds of parsimony the more restricted model is acceptable (Anderson and Gerbing, 1988, pp. 419–420). The χ^2 statistics are affected by sample size (Joreskog and Sor-

TABLE 1. Means and Standard Deviations for the Rating Items

Constructs and Items	Mean	SD
Available Behavior		
1. Clarity of presentation	3.65	1.07
3. Feedback on learning	3.23	1.14
5. Use of class time	3.86	0.78
8. Assessment technique	4.37	0.89
Implied Trait		
2. Enthusiasm	3.90	1.02
7. Leadership ability	3.68	1.30
10. Dependability	3.14	1.19
Inferred Behavior		
4. Interaction with students	3.02	1.48
6. Command of subject matter	3.12	1.01
9. Attention to individual differences	3.55	1.17

bom, 1989), but this was not a problem in the present study because the sample size was constant across the models compared. The sign and size of the path coefficients and the conceptual framework of the study were also used in assessing the fit of the models.

RESULTS

Distributional Assumptions

Univariate descriptive statistics and graphical analysis revealed that the assumption of linear interitem relationships was met. Multivariate normality assumption was satisfied as well. Because of the incomplete information in the rating task, missing data were of particular interest in this study. Surprisingly, only 2 percent of the data were missing with no consistent pattern, allowing "pairwise" treatment of missing data. Descriptive statistics are reported in Table 1, and intercorrelations are presented in Table 2.

Measurement Model

To test the existence of three distinct but interrelated latent constructs, a measurement model (MM) was fit to the data via confirmatory factor analysis. As shown in Figure 3, a three-factor model was hypothesized. Consistent with the conceptual framework about how information is encoded in the IPT, factors were allowed to intercorrelate, but each item could load only on the factor it was intended to measure. The data supported the MM quite well, $\chi^2(32, N = 219) = 31.30, p = 0.50$. Thus, there was support for the presence of three

TABLE 2. Item Intercorrelations, Variances, and Covariances

Items	1	2	3	4	5	6	7	8	9	10
1	.91	.17	.24	.30	.15	.10	.14	.02	.12	.10
2	.20	.90	.15	.17	.04	.06	−.04	.04	.13	.10
3	.33	.20	.80	.19	.15	−.03	.08	.06	.05	.06
4	.39	.22	.27	.88	.07	−.01	.08	.06	.05	.06
5	.18	.05	.20	.08	.91	.12	.07	.16	.26	.08
6	.13	.08	−.05	−.02	.15	.87	.01	.07	.06	−.01
7	.17	−.05	.12	.10	.07	.01	.90	.19	.12	.10
8	.03	.05	.09	.07	.20	.09	.23	.93	.12	.16
9	.15	.06	.07	.17	.33	.07	.15	.14	.89	.20
10	.17	.12	.10	.07	.10	−.01	.13	.21	.27	.85

Polychoric correlations are below the diagonal. Asymptotic variances are on the diagonal, and asymptotic covariances are above the diagonal.

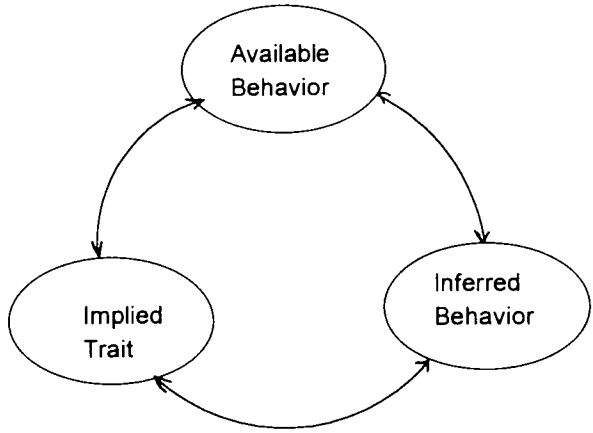

FIG. 3. Measurement model.

latent constructs that comprised the hypothesized operating model. The three latent constructs accounted for 76.4 percent of the variance in the 10 rating items.

Structural Models

First, the theoretical or the operating model (OM) shown in Figure 1 was fit to the data. This model differs from the MM in two ways. The relationships between the latent constructs are asymmetrical, and the path between available and inferred behavior does not exist. The OM model yielded a good fit to the

TABLE 3. Goodness-of-Fit and Model Comparison Information

Model	*df*	χ^2	*p*	GFI	AGFI	RMR
Measurement Model (MM)	32	31.30	=.50	.973	.954	.037
Operating Model (OM)	34	37.11	=.33	.970	.951	.043
Competing Model (CM)	34	48.47	=.05	.960	.937	.052
Model Comparisons						
OM-MM	2	5.81	>.05			
OM-CM	1	11.36	<.05			

data, χ^2 (34, $N=219$) = 37.11 p = 0.33. It is more restricted (fewer paths and $df=34$) than the measurement model ($df=32$). Comparatively, the fit of the OM was not significantly worse, $\chi^2(2) = 5.81$, $p > .05$, than the more general MM in which all latent constructs are intercorrelated. The structural equations accounted for 33 percent of the variance. Thus, the OM with theoretically imposed constraints was the more parsimonious model that adequately described the covariance structure in the data.

Further analysis was done to examine whether the competing model (CM) described the data just as well as the hypothesized OM. Global and comparative fit indexes for all models are reported in Table 3. The data did not quite support the CM, $\chi^2(34, N=219) = 48.48$, $p = 0.05$. Given the marginal probability of fit, CM was also compared with the MM and the OM. In both comparisons, the fit of the CM was significantly worse. Consequently, the theoretical OM best described the covariance structure of the data. Acceptance of OM was also based on the parameter estimates for the structural relations. Standardized estimates of all parameters for the simultaneously fitted measurement and the structural components are presented in Table 4. The two hypothesized causal paths were statistically significant, $\gamma = 0.578$, $t = 2.61$, $p < 0.05$ and $\beta = 0.746$, $t = 2.41$, $p < 0.05$. These coefficients were nonnegative supporting the causal direction as predicted. Moreover, the GFI and AGFI were closest to one and the RMR was closest to zero for the OM.

DISCUSSION

The hypothesized structural model was the model best supported by the data. Not only did the model fit as a whole, but the predicted causal paths in this model were statistically significant and in the direction specified. Thus, the hypothesized model was interpretable within the conceptual framework on which it was posited. Parsimony preferred, the hypothesized model fit the covariance structure better than the confirmatory measurement model. On the other hand, the competing structural model did not fit the data by any criteria.

The hypothesized model predicted that students' IPT causally influences their

TABLE 4. Parameter Estimates of the Hypothesized Model

	Path	Standardized Coefficients
Measurement Component		
Available Behavior		
Clarity of presentation	λ_1	.603
Feedback on learning	λ_2	.312
Use of class time	λ_3	.398
Assessment technique	λ_4	.489
Implied Trait		
Enthusiasm	λ_5	.350
Leadership ability	λ_6	.149
Dependability	λ_7	.307
Inferred Behavior		
Interaction with students	λ_8	.429
Command of subject matter	λ_9	.482
Attention to individual differences	λ_{10}	.381
Structural Component		
γ_1		.578
β_1		.746

*$p < .05$

rating judgments of instructors' teaching behavior. As expected, the path from available behavior to implied traits was statistically reliable. The available behavioral information alone was sufficient for students to rate items indicating the trait construct. As indicated by the schematic processing framework, available behavioral information most likely activated raters' IPT that implied trait dispositions. The IPT enabled the raters to draw inferences to provide ratings on trait items. The path from implied trait to inferred behavior was also large and statistically significant. Again as predicted, the IPT implied behavioral information that was unavailable. As noted in the results, missing data were minimal and patternless. The raters inferred the unavailable behavioral information that enabled them to rate the items for which there was no relevant information available.

The findings that behavioral information can imply traits and traits imply unavailable behaviors are consistent with previous research and theory. In an experimental study, Krzystofiak and associates (1988) found that traits were inferred from performance behaviors and vice versa. Theories of how trait and behavioral information are mentally encoded (Cantor and Mischel, 1979; Trzebinski et al., 1985) also support the findings. Unavailable traits and behaviors were inferred from given behaviors because traits and behaviors are encoded together. Hence, the inferred ratings were most likely a measure of the

covariance of traits and behaviors in the raters' IPT and not in the instructor's teaching behaviors. This conclusion echoes the critique that the factors in performance rating instruments result from the covariance in students' IPT of teaching and not from actual behavioral covariance (cf. Whitely and Doyle, 1976).

The findings also imply that performance rating instruments free of trait items are not immune to inferential bias. The significant causal effect of implied trait ratings on inferred behavioral information suggests that ratings of some behavioral items would be biased by the students' perception of an instructor's personality. Evaluation of teaching performance involves mental comparison where the evaluator compares the teacher with his/her IPT or the exemplar (cf. Shulman, 1986). In this comparison process, personality dispositions will enable extensive inferences about the ratee (cf. Wyer and Martin, 1986). The action-oriented mental representation of the IPT indicates the normative role behaviors (Trzebinski et al., 1985). Thus, the mental comparisons and inferences automatically bias the ratings, even when rating instruments do not contain any trait items. This is why the halo error is so ubiquitous in performance ratings. Cooper (1981) stated that raters' IPT is activated no matter how the rating categories are obtained and defined.

As hypothesized, students' IPT functions through schematic processing and enables them to fill in the missing information. Although schematic processing is an advantage in limited information situations, it introduces inferential bias in judgments (Taylor and Crocker, 1981). Most of the ratings in the present study resulted from raters' inferences based on trait and behavioral covariation in their IPT, or else they could not have rated 6 of the 10 items in the scale. If raters' IPT is an undesirable influence on their ratings, then their ratings contain error. Thus, errors of measurement in performance rating instruments do not seem to be truly random as assumed by the classical test theory, the common basis for the design of performance rating instruments.

Several researchers have investigated the amount of improvement in instruction as a result of feedback from students' evaluation. In a recent meta-analysis, L'Hommedieu, Menges, and Brinko (1990) found a small effect (0.342) of feedback on improvement in instruction. These authors offered a long list of methodological problems in studies they analyzed. Another explanation can be added to their list. If, as found in the present study, the effect of IPT is so pervasive, then student ratings in some of the studies did not vary simply because both pre- and postinstruction ratings shared a common variance from the students' IPT. Indeed, student ratings are fairly stable over time (Marsh, 1987). The stability of student ratings may be due to the stability of their IPT, which could also inhibit them from noticing any changes an instructor might implement during the course (cf. Rotem and Glasman, 1979). Abrami and associates (1990) have suggested that class mean ratings from different sections in a

course are a more reliable and valid measure of an instructor's teaching performance. Means are always more reliable than single observations, but reliability is not a sufficient condition for validity. Mean ratings will not necessarily be valid if the ratings that produce the mean are biased. Likewise, neither within-class nor between-class correlations would produce more valid factor structures if the ratings contain inferential bias.

The findings of the present study have an important implication for performance appraisal practice. The findings suggest that performance rating instruments provide the option "cannot rate" for each item in the rating instrument. This is not a common procedure, but it may minimize IPT induced biased inference. Although this will introduce the problem of missing observations in data analysis, the ratings made and the factors extracted will then be least influenced by students' IPT.

The design of the present study did not capture all aspects of an actual evaluation, but had many features of an actual appraisal: some information had to be recalled or inferred, ratings were made on Likert scales, ratings were completed in 6 to 8 minutes, only one instructor was evaluated, and the instructor was not in action when the students completed the rating instrument. Nevertheless, because the instructor's performance was presented in a vignette, the findings may lack external validity. At this stage of investigation on student raters' cognition and given the purpose of the present study, it is felt that the control available by the use of the vignette outweighs its shortcoming. External validity would no doubt become an important concern when a substantial body of evidence on raters' cognition has accumulated. Besides, consider certain features of the design of the present study. No particular instructor was identified and no personality information was presented. In actual evaluation, a target instructor is identified in the raters' mind. When a target person is identified, there would be a greater chance of raters recalling the target's personality dispositions while processing behavioral data. This would happen because evaluations of teaching do depend on both trait and behavioral information (Medly, 1979), and trait and behavioral information are encoded together (Cantor and Mischel, 1979).

Furthermore, in real settings students will interact with and observe the instructor for a longer period and know the instructor as a person, which most likely will ease the retrieval from memory of personality information during the appraisal. Consequently, depending on how much an evaluation depends on memory, the effects of IPT on performance ratings may be even stronger in an actual appraisal, even if the rating instrument is free of items measuring personality dispositions. Although the instructor evaluated was hypothetical, to discount the findings it has to be shown that in real appraisal processes of recall and inference do not occur. The more an appraisal depends on memory and inference, the more biased the ratings will be in the direction of the raters' IPT (cf. Srull and Wyer, 1989).

Can we ever eliminate the influence of IPT in performance evaluation? The prevalence of behavioral rating scales is founded on the belief that well-designed instruments can reduce inferential bias. However, Landy and Far (1980) concluded that traditional psychometric attempts have had limited success in addressing validity issues. Raters' cognition or actually how the ratings are actually formed in the rater's mind may be quite revealing (Cooper, 1981; DeNisi, Cafferty, and Meglino, 1984). The IPT of good teaching may have a relevant and meaningful place in performance evaluation. If schematic processing is so natural and automatic and the IPT is the basis for comparison, it might be better to capture the raters' IPT. It may be possible to remove the effect of IPT from the ratings by some statistical means. In this regard, asking students to rate an instructor against the ideal may be a methodology to explore. Asking for ideal teacher comparisons may also provide a more authentic evaluation from the students' point of view. The results also suggest that the validity of student ratings may be improved further by focusing on actually how the students form their ratings. Relatively fewer studies have focused on the students' ability to rate without bias. More effort needs to be devoted to students rating strategies rather than on instrument development (cf. Cook, 1989; Kishor, 1992). Training student raters to rate without inferential bias needs to be explored.

It was found that student raters' inference in evaluation of teaching was biased by their implicit personality theories of teaching. The bias resulted from the natural and automatic schematic information processing mechanism of the human mind. Therefore, understanding raters' cognition may at least be as important as improving instrumentation. If one can identify how judgments are caused, those judgments could be improved for their reliability and validity. Understanding how students cognitively form their ratings will most likely provide additional ways of enhancing the validity of students' ratings of instructors.

REFERENCES

Abrami, P. C., d'Apollonia, S., and Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology* 82(2): 219–231.

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin* 103(3): 411–423.

Anderson, N. H. (1981). *Foundations of Information Integration Theory*. New York: Academic Press.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern. *Psychological Bulletin* 107(2): 260–273.

Cadwell, J., and Jenkins, J. (1985). The effects of semantic similarity of items on student ratings of instructors. *Journal of Educational Psychology* 77(4): 383–393.

Cantor, N., and Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, vol. 12 (pp. 3–52). New York: Academic Press.

Carlson, M., and Muliak, S. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research* 28(1): 111–159.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research* 18: 115–126.

Cook, S. S. (1989). Improving the quality of student ratings of instruction: A look at two strategies. *Research in Higher Education* 30(1): 31–45.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin* 90: 218–244.

DiNisi, A. S., Cafferty, T. P., and Meglino, B. M. (1984). A cognitive view of the performance appraisal processes. *Organizational Behavior and Human Performance* 33: 360–396.

Doyle, K. O., Jr. (1981). Validity and perplexity: An incomplete list of disturbing issues in instructional evaluation research. *Instructional Evaluation* 6(1): 23–25.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education* 30(2): 137–194.

Gaski, J. F. (1987). On construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology* 79(3): 326–330.

Hastie, R. (1981). Schematic principles in human memory. In E. Higgins, C. Herman, and M. Zanna (eds.), *Social Cognition: The Ontario Symposium*, vol. 1, (pp. 39–88). Hillsdale, NJ: Erlbaum.

Hayduk, L. A. (1987). *Structural Equational Modelling with LISREL: Essentials and Advances*. Balitmore: The John Hopkins University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1986). *Induction: Processes of Inference Learning and Discovery*. Cambridge, MA: MIT Press.

Howard, G. S., and Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education* 6: 175–188.

Ilgen, D. R., and Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw and L. Cummings (eds.), *Research in Organizational Behavior*, vol. 5 (pp. 141–197). Greenwich, CT: JAI Press.

Jenkins, J. (1987). Implicit theories and semantic similarities: Reply to Marsh and Groves (1987). *Journal of Educational Psychology* 79(4): 490–493.

Joreskog, K. G., and Sorbom, D. (1989). *LISREL 7 User's Reference Guide*. Moresville, IN: Scientific Software Inc.

Kishor, N. (1990). The effect of cognitive complexity on halo in performance judgment. *Journal of Personnel Evaluation in Education* 3: 377–386.

Kishor, N. (1992). Compensatory and noncompensatory information integration and halo error in performance rating judgments. *Journal of Personnel Evaluation in Education* 5: 257–268.

Krzystofiak, F., Cardy, R., and Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology* 73(3): 515–521.

Landy, F. J., and Farr, J. L. (1980). Performance rating. *Psychological Bulletin* 87: 72–107.

Larson, J. R., Jr. (1979). The limited utility of factor analytic techniques for the study of implicit theories in student ratings of teacher behavior. *American Educational Research Journal* 16(2): 201–211.

L'Hommedieu, R., Menges, R. J,. and Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology* 82(2): 232–241.

Marsh, H. W. (1984). Students' evaluation of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology* 76(5): 707–754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11(3): 253–388.

Marsh, H. W., and Groves, M. A. (1987). Students' evaluation of teaching effectiveness and implicit theories: A critique of Cadwell and Jenkins (1985). *Journal of Educational Psychology* 79(4): 483–489.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education* 21: 150–158.

Medley, D. (1979). The effectiveness of teachers. In P. Paterson and H. Walberg (eds.), *Research on Teaching: Concepts, Findings and Implications*. Berkeley, CA: McCutchan.

Mook, D. J. (1983). In defense of external validity. *American Psychologist* 38: 379–387.

Nisbett, R., and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall.

Rotem, A., and Glasman, N. S. (1979). On the effectiveness on students' evaluative feedback to university instructors. *Review of Educational Research* 49: 497–511.

Simon, H. (1978). Rationality as process and as product of thought. *American Economic Review: Papers and Proceedings* 68: 1–16.

Sobel, M. E., and Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (ed.), *Sociological Methodology* (pp. 152–178). San Francisco: Jossey-Bass.

Srull, T. K., and Wyer, R. S. (1989). Person memory and judgment. *Psychological Review* 96: 58–83.

Sulman, L. S. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (ed.), *Handbook of Research on Teaching*, 3rd ed. (pp. 3–37). New York: Macmillan.

Taylor, S. E., and Crocker, J. (1981). Schematic bases of social information processing. In E. Higgins, C. Herman, and M. Zanna (eds.), *Social Cognition: The Ontario Symposium*, vol. 1 (pp. 89–134). Hillsdale, NJ: Erlbaum.

Trzebinski, J. (1985). Action-oriented representations of implicit personality theories. *Journal of Personality and Social Psychology* 48(5): 1266–1278.

Trzebinski, J., McGlynn, R. P., Gray, G., and Tubbs, D. (1985). The role of categories of an actor's goals in organizing inferences about a person. *Journal of Personality and Social Psychology* 48(6): 1387–1397.

Whitely, S. E., and Doyle, K. O. (1976). Implicit theories in student ratings. *American Journal of Educational Research* 13(4): 241–253.

Wyer, R. S., and Martin, L. L. (1986). Person memory: The role of traits, group stereotypes and specific behaviors in the cognitive representation of persons. *Journal of Personality and Social Psychology* 50: 661–675.