DOCUMENT RESUME

ED 295 966 TM 011 752

AUTHOR Mehrens, William A.; And Others
TITLE Fiscal Viability, Conjunctive and Compensatory Models, and Career-Ladder Decisions: An Empirical Investigation.
PUB DATE Apr 88
NOTE 52p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Career Ladders; *Cost Effectiveness; Decision Making; Elementary School Teachers; Elementary Secondary Education; Models; Primary Education; Secondary School Teachers; *Teacher Evaluation; *Weighted Scores
IDENTIFIERS *Tennessee Career Ladder Program

ABSTRACT

A study was undertaken to explore cost-effective ways of making career ladder teacher evaluation system decisions based on fewer measures, assessing the relationship of observational variables to other data and final decisions, and comparison of compensatory and conjunctive decision models. Data included multiple scores from eight data sources in the 1985-86 Tennessee Career Ladder Teacher Evaluation System. The data sources include: (1) classroom observation; (2) teacher/evaluator dialogues in planning, teaching strategies, and evaluation; (3) a peer questionnaire; (4) a principal questionnaire; (5) student questionnaires (primary, elementary, and secondary levels); (6) the Tennessee Career Ladder Professional Skills Test; (7) Professional Development and Leadership Activities Summary; and (8) evaluator consensus judgments. Findings indicate that less data can be gathered without having any major impact on the decisions reached if one uses optimal weighting. Several fairly accurate models using various reduced sets of data were proposed. Although classroom observation data were not highly related to the other variables, there was evidence indicating that the decisions reached without such expensive-to-gather data would be highly similar to the decisions actually reached. It was not possible to compare a purely compensatory model with a purely conjunctive model using these data. However, the comparison of the actual decisions reached in Tennessee with those made using a conjunctive data combination model gave no support for preferring a compensatory model. (TJH)

FISCAL VIABILITY, CONJUNCTIVE AND COMPENSATORY MODELS, AND CAREER-LADDER DECISIONS: AN EMPIRICAL INVESTIGATION

William A. Mehrens
Michigan State University

Joyce R. McLarty
American College Testing Program

Ernest A. Rakow
Memphis State University

S.E. Phillips
Michigan State University

Running Head: FISCAL VIABILITY

Paper prepared for the National Council on Measurement in Education annual meeting, New Orleans, April, 1988.

FISCAL VIABILITY, CONJUNCTIVE AND COMPENSATORY MODELS, AND CAREER-LADDER

FISCAL VIABILITY, CONJUNCTIVE AND COMPENSATORY MODELS, AND CAREER-LADDER DECISIONS: AN EMPIRICAL INVESTIGATION

Abstract

The purpose(s) of the study included: (1) exploration of cost effective ways of making career ladder teacher evaluation system decisions based on fewer measures; (2) assessment of the relationship of observational variables to other data gathered and the final decisions; and (3) comparison of compensatory and conjunctive decision models. Data included multiple scores from eight data sources in the 1985-86 Tennessee Career Ladder Teacher Evaluation System.

The findings from this study suggested that less data can be gathered without having any major impact on the decisions reached if one uses optimal weighting. Several fairly accurate models using various reduced sets of data were proposed. Although classroom observation data were not highly related to the other variables there was evidence indicating the decisions reached without such expensive-to-gather data would be highly similar to the decisions actually reached.

It was not possible to compare a purely compensatory model with a purely conjunctive model using these data. However, the comparison of the actual decisions reached in Tennessee with those made using a conjunctive data combination model gave no support for preferring a compensatory model.

FISCAL VIABILITY, CONJUNCTIVE AND COMPENSATORY MODELS, AND CAREER LADDER DECISIONS: AN EMPIRICAL INVESTIGATION

High stakes decisions should be based on high quality data. In general, the more data that are gathered, the better the decision-making process. However, in gathering data for any decision one must consider the cost of gathering the data relative to the improvement in the decision made. One of the major feasibility standards listed in the Standards for Evaluation of Educational Personnel is fiscal viability (Stufflebeam & Brethower, 1987). Gathering more data will not always be cost effective.

Moreover, when more than one piece of data is gathered one must decide how to weight the various pieces of data. The psychometric literature discusses two prominent models for combining data: the compensatory model and the conjunctive model. Also, a combination of the two can be used. In the compensatory model, high scores on one variable can compensate for low scores on another variable. Multiple regression is an example of this approach. In the conjunctive model an individual must score above the cutoff score on each measure used. An example of a combination model would be to allow compensation above the individual cutoff scores where a combined score higher than the sum of the individual cutoff scores is required (see Mehrens, Phillips, Anderson, 1987).

When criterion data exist for a comparable group, the weights can be established such that they result in the lowest number of incorrect decisions for that group. Then, through validity generalization those weights can be used for subsequent decisions for future comparable groups. However, for many decisions, there is no completely adequate criterion measure against which to determine optimal, mathematically derived, predictor weights.

The Tennessee Career Ladder Teacher Evaluation System is a case in which both the issues of the cost/benefit ratio of gathering data and methods of combining data are relevant. To satisfy the legislative mandate for the career ladder program, the state needed to determine which teachers should be placed at which rungs of the career ladder. One of the basic principles of the system was that "multiple sources of data are essential to the development of a complete picture of teaching performance" (Tennessee Department of Education, 1986a, p. 2). In both 1985-86 and 1986-87 multiple data sources were combined using a combination compensatory/conjunctive approach.

A variety of questions can be raised regarding the specific approaches used by Tennessee. For example, would their method be transportable to other states, or would it be considered too expensive? Could a judicious selection of a subset of the variables used result in a reasonably comparable set of candidates selected at a lower cost? Would different weighting procedures have resulted in much the same set of decisions? If not, what differences characterize the candidates selected under different systems?

Many educators believe observation is the most valid way to evaluate teachers. In Tennessee, observational data included both low inference (frequency counts) and medium inference data (ratings). Are there any appropriate combinations of the observational variables that would result in essentially the same decisions about teachers as would be made using the non-observational data?

OBJECTIVES

The specific objectives of this paper were as follows:

1. To determine for the 1985-1986 Tennessee data how well the less expensive-to-gather data could, when optimally weighted, predict the actual decisions made.
2. To determine how well the decisions made without the classroom observation data would match the actual decisions made.
3. To determine the relationship between the observation variables and all other pieces of data.
4. To determine how similar the decisions would be under the compensatory and conjunctive weighting models.

METHODOLOGY

Instruments Used

The 1985-86 Tennessee Career Ladder Evaluation System was based on the data from the following eight data sources:

1) Classroom Observation
2) Teacher/Evaluator Dialogues in Planning, Teaching Strategies, and Evaluation
3) Peer Questionnaire
4) Principal Questionnaire
5) Student Questionnaires (secondary, elementary, and primary levels)
6) Tennessee Career Ladder Professional Skills Test
7) Professional Development and Leadership Activities Summary
8) Evaluator Consensus Judgment

The Classroom Observations were based on the observations of a trained three member evaluation team. Each evaluator visited once. There were one announced and two unannounced visits. Twenty-five scores were obtained: two in the planning domain, 13 in the teaching strategies domain, one in

evaluation and nine in classroom management (27 scores were originally obtained but one was dropped and two were combined.)

The Teacher/Evaluator Dialogues were conducted by the evaluators during their visits at the school. Each evaluator conducted one dialogue. There were ten Dialogue Scores: three from the first evaluator focusing on planning issues; four from the second evaluator on teaching strategies; and three from the third evaluator, two on evaluation and one based on classroom management.

The Peer Questionnaire was completed anonymously by three colleagues randomly selected from a list of six supplied by the candidate. It provided a single score in the leadership domain.

Fifteen scores were obtained from the Principal's Questionnaire: 3 for planning, 4 for teaching strategies, 3 for evaluation, 3 for classroom management and 2 for leadership. These were subsequently combined to produce 5 means--one in each domain.

The Student Questionnaires were administered to the candidate's students by an evaluator with the candidate out of the room. They differed across the three levels. There were 9 different scores obtained for both the elementary and secondary levels (4 on teaching strategies, 2 on evaluation and 3 on classroom management) and three for the primary level (teaching strategies, evaluation, and classroom management). (For our analyses combining levels we obtained 3 scores for each level.)

The Tennessee Career Ladder Professional Skills Test was a multiple-choice test and produced four scores: Planning, Teaching Strategies, Evaluation, and Classroom Management.

The Professional Levelopment and Leadership Summary (PDL) was based on information the candidate provided about five development and five leadership

activities. These were jointly judged by two evaluators and two scores were produced: (1) enhances instruction with new techniques, and (2) exhibits leadership to improve schooling.

The Evaluator Consensus Judgment Scores were obtained after the data collection phase. The evaluators met to discuss the candidates and provide consensus judgment scores on the first four domains: planning, teaching strategies, evaluation, and classroom management. Although the observation and dialogue scores were not available to the evaluators during consensus meetings, their personal notes from the school visits were. For this reason, the dialogue, observation and consensus data were not independent measures.

All the data were combined in a set of complicated weighting and scaling procedures to provide five domain scores: Planning, Teaching Strategies, Evaluation, Classroom Management and Leadership. This was done by first combining all data except the principal and consensus scores by domain. These we call the data scores. The data scores were placed on a 200-800 scale (the principal and consensus scores were already on this scale), and the three sets were weighted and combined to form domain scores. While there were some differences across the domains, for domains 1-4 the data sources were weighted about 65%, the principal ratings about 10% and the consensus scores about 25%. For domain 5, the data sources were rated 80% and the principal ratings 20%. Domain scores 1-5 were then weighted .15, .35, .15, .25, and .10, respectively, to obtain a total composite score.

To be qualified for a Career Level above Level 1 the teachers had to meet or exceed a minimum scaled score of 450 in each domain and, in addition, have a total weighted composite score that met or exceeded the qualifying score for a particular level (600 for Level 2, and 700 for Level 3).[1] Details of the

instruments can be found in the Teacher Orientation Manual and details regarding the scoring can be found in the teacher's edition of the Career Ladder Technical Manual (Tennessee State Department of Education, 1986a & 1986b).

As can be discerned from the above abbreviated description, the Tennessee Career Ladder Evaluation System uses a multiple data source approach that can be both expensive to maintain and difficult to explain.

> "The processes of aggregating scores, rescaling them, weighting and combining them, ... was confusing to almost everyone. Once the system was implemented...explaining the system to the candidates became a major concern." (McLarty, Furtwengler and Malo, 1985, p. 16).

Nevertheless, while

> "the multiple data source system is difficult and expensive to build, relatively inflexible, and complicated to explain it also provides a thorough and equitable evaluation, is relatively stable, and is based on concepts which are logically appealing" (p. 19).

Although the Tennessee Career Ladder System remains somewhat controversial, the evaluation component has been perceived as producing reasonably valid results. Thus, for our purposes, we used the current classification as the criterion against which to compare other methods.

Population/Sample

Teachers completing evaluations for the 1985-86 cycle of the Career Ladder in Tennessee were used in the analyses. The 1985-86 cohort was chosen because there was more variability in the data than for the 86-87 cohort. All teachers who were in the General Education category were included. Adaptations of the evaluation system were made for special education teachers, Chapter 1 resource teachers, and vocational education teachers. Data for these individuals were not included in our analyses.

The population of respondents was divided into two stratified random samples where the stratification was based on whether the teacher was a primary, elementary, or secondary teacher. One of the two groups was used for the initial analyses and the other group was used for cross validation.

Analyses

The general analytic procedures were based on the four objectives listed earlier. First some preliminary descriptive analyses were completed to assist in the identification of meaningful ways to aggregate the data. Discriminant functions were run to provide data relevant to objectives 1 and 2. Canonical correlation analyses were used for objective 3 and hit rates comparing a conjunctive (by domain) approach to the original combined conjunctive/compensatory approach were obtained to address objective 4.

RESULTS

Preliminary Results

An initial set of descriptive analyses was completed on the first half sample. This sample consisted of 535 teachers: 97 primary teachers, 128 elementary teachers, and 310 secondary teachers. The breakdown by career ladder level was as follows: 68 Level I, 42 Level II, and 325 Level III Teachers.

The means, standard deviations, minimum scores and maximum scores were computed for each variable. Although these results are too voluminous to present in this paper, they are useful in explaining why some variables counted for little in subsequent analyses (high means and little variance). For example, several observations (observations 6, 14, 19, 22, 23, and 27) had

means above 99 (maximum possible was 100) and the principal mean ratings were all above 768 (on a 200-800 scale). Some of the more interesting descriptive data are shown in Table 1.

Insert Table 1 About Here

The seven observation variables which had extremely high means were as follows: Score 6: whether or not the teacher provided the students with correct information; Score 7: whether the teacher provided an appropriate language model for students; Score 14: the proportion of students in the class who were "on task"; Score 19: an index of whether student behavior was under control; Score 22: whether the teacher treated students of different races and both genders equitably; Score 23: the amount of time required to resume class after an interruption (if one occurred) and; Score 27: whether the teacher was "on task."

These scores, except for #19 and #23, focused on identifying serious but infrequent teacher errors. The score distribution for #19 resulted from the complex pattern scoring procedure. That for #23 was due to the small number of situations in which a naturally-occuring interruption was observed.

All five Principal subdomain scores also had very high means (all above 770) and their variances were quite low. The final combined domain score means were all well above the 600 minimum required for Career Ladder Level II. Three of these means approached and two exceeded the 700 required for Level III. The mean composite score of 690 was close to that required for Level III. Out of 535 individuals, 325 (61%) were identified as Level III teachers. (The distribution of composite scores was negatively skewed.)

Because of the large volume and considerable complexity of the data, as part of our initial exploration we obtained correlation matrices and a series of regression equations. Because the Tennessee Career Ladder data are weighted and aggregated at multiple stages, it was necessary to determine at what level of aggregation to examine the data. Three levels were selected: (1) the domain score level (five scores, each incorporating data from most of the instruments); (2) the subtotal score level (fourteen scores: five based on all the data except the principal and consensus scores [hereafter called the data scores], five from the principal, and four from the evaluator consensus); and (3) the indicator level (61 scores--23 from the observations, 10 from the dialogues, 1 from the peer questionnaire, 3 from the student questionnaires, 4 from the professional skills test, 2 from the professional leadership and development survey, 14 from the principal, and 4 from the evaluator consensus judgment).

The domain score correlation matrix for the total sample of 535 cases is given in Table 2A. As can be seen, the correlations among the first four domains were all above .64. Domain 5 (Leadership) appeared to measure something a bit less related to the other domains. The correlations between the domains and the composite were, of course, related to the weights (.15, .35, .15, .25, and .10) of the domains in computing the composite.

The subtotal domain score correlations are presented in Table 2B. Recall that the data scores included the observations and dialogues (as well as other variables). The observations and dialogues were conducted by the evaluators--who also provided the consensus scores. Thus, we would expect the data and consensus subdomain scores to be correlated. It is noteworthy that the five data scores correlate somewhat less with each other than the first four

data subscores correlate with the consensus subscores. In general the data in Table 2B suggest that the homo method--hetro trait triangles for consensus and principal scores present higher correlations than the hetro method--homo trait correlations. These data strongly suggest the lack of convergent-discriminant validation of the traits (domains) (Campbell & Fiske, 1959).

The correlation matrix for the 61 indicators is presented in Table 2C. Observation data were aggregated across occasions so occasion variance was treated as error. Therefore, many of the individual indicators based on observations, as expected, had low correlations with other variables. Indeed, the correlations among the observation indicators are quite low. However, the ten dialogue scores have fairly high correlations with each other as do the three student questionnaire scores, the four test scores, the two professional development and leadership scores, the fourteen principal scores, and the four consensus scores. Only the four consensus scores tended to correlate very much with the other scores. (Recall, however, the possible interaction of the observations, dialogue, and consensus scores.) Note that the consensus scores have low correlations with the four test scores (.02 to .23). They correlate moderately with the principal scores (.22 to .38).

Insert Tables 2A through 2C About Here

Regression equations using the backward stepwise deletion procedure were constructed to identify variables which might be deleted without significant loss of information. In each case, the composite score was treated as the dependent variable. When the five domain scores were treated as the independent variables, none was deleted. When the composite was regressed on

the 14 subtotal scores again no variables were dropped although the beta weights for the 5 subdomain scores from principals were all quite small. However, when the composite was regressed onto the 61 indicators only 24 scores remained in the equation. The multiple R for those 24 was 0.974. Of the 37 variables deleted, 17 were from the observations, 5 were dialogue scores, 1 was a student questionnaire score, 1 a test score, 1 a Professional Development and Leadership (PDL) score and 12 were principal questionnaire scores. In a reduced analysis using 43 independent variables (all the above 61 except the 14 principal questionnaire scores and the 4 consensus judgment scores), only 23 remained in the equation after stepwise deletion. The multiple R was 0.95 and the 20 variables deleted were: 14 observation variables, 3 dialogue, 1 student, 1 test and 1 PDL.

Separate analyses by grade level (primary, elementary, and secondary) showed much the same results on both the descriptive statistics and the various regression analyses. Thus we decided to complete all the remaining analyses using the combined data.

Variables for inclusion in the discriminant function and canonical correlation analyses were selected based on the purposes of the study, the initial findings, and the costs of obtaining certain types of data. Further, hit rates were computed and a set of contingency tables were constructed to compare the actual outcome by levels with what would have been obtained using strictly a multiple cutoff (conjunctive) approach on the domain scores.

Discriminant Analyses

One purpose of the study (objective 1) was to determine how well a subset of data that was less expensive-to-gather could, when optimally weighted, predict the actual decisions made. A specific sub-purpose (objective 2) was to determine how well decisions made without expensive-to-gather observation data would match actual decisions.

A series of discriminant function analyses were run based on 512 cases (the original 535 minus 23 cases which had at least one missing variable). Some variables were also dropped due to missing data.

To provide "baseline" data the first discriminant analysis was run using all the data. Had we used the combined conjunctive/compensatory model (450 minimum on each domain; 600 and 700 composites for Level II and III) we would have had no misses. However, the discriminant function did not use the conjunctive component of the model. Further, the discriminant function assumes multivariate normality of the distributions and equal dispersion and covariance structures for the groups. Although the analyses is not very sensitive to violations of these assumptions we knew they were violated to a considerable degree. Therefore obtaining the baseline fit using all the data seemed useful. The prior probabilities for the three levels were .13, .26, and .61. Thus, the maximum chance criterion hit rate was .61 and the proportional chance criterion was .46 ($.13^2 + .26^2 + .61^2$).

The discriminant analysis classification using all the scores is presented in Table 3A. Note that 93.2% of the cases were correctly classified. There were 4.7% false positives and 2.1% false negatives.

Insert Tables 3A through 3F About Here

The data in Tables 3B through 3F indicate the hit rate for various deletions of independent variables: 3B omits the observation scores (90.8% hit rate); 3C omits the consensus scores (91.8%); 3D omits the dialogue scores (90.5%); 3E omits the observation and consensus scores (87.5%); and 3F omits the observation, dialogue and consensus scores (71.4%).

Keeping in mind the baseline of 93.2% hits (cross validation evidence will come later), it would seem that omitting all observations and consensus scores and obtaining a hit rate of 87.5% (Table 3E) may be a reasonably cost-effective approach. Training the evaluators to observe and conducting the actual observations was very expensive. If judicious weighting of the remaining variables produces a hit rate close to the baseline hit rate using all the variables it may be cost effective to forgo the observation and consensus data obtained from the evaluators.

If one wished to save even more resources (time and money) the dialogues also could be omitted. However, the drop to a hit rate of 71.4% (Table 3F) seems like too large a loss. The outside evaluators are probably needed to conduct the dialogues to keep the system accurate and acceptable to the public. Note particularly the 3 to 1 ratio of false positives to false negatives if all the data from the outside evaluators are dropped.

In an effort to make some substantive interpretations of the findings for the data in Table 3E, we looked at the discriminant structure correlations (loadings). These are simply the linear correlations between the independent variables and the discriminant function and reflect the variance shared between the independent variables and the discriminant function. Although the loadings are subject to instability, they are considered more valid than the

discriminant weights as a means of interpreting the discriminating power of the independent variables (see Hair, Anderson, & Tatham, 1987, p. 91).

The structure correlations (loadings) for the data in Table 3E are presented in Table 4. These data indicate that the 10 dialogue scores counted the most in differentiating between Levels II and III. The principals' scores weighted heavily in differentiating between Levels I and II.

Insert Table 4 About Here

The data in Table 3F indicate clearly the loss of the dialogue scores in differentiating between Levels II and III. Note that there are 16 individuals actually in Level I who were classified in Level II, and 13 in Level I who were classified in Level III, for 29 errors. For Level II, 13 were classified in Level I and 81 in Level III. For Level III, 3 were classified in Level I and 22 in Level II. The major difference between the hit rates in 3E and 3F were the false positives in 3F due to placing many more actual Level II individuals into Level III (as mentioned earlier, this results in a 3 to 1 false positive to false negative ratio).

The discriminant structure correlations (not shown here) for the data in Table 3F (which omits observation, dialogue and consensus scores) indicated that the three student questionnaires counted the most in differentiating between Levels I and II. These are followed by the two Professional Development and Leadership Summary (PDL) scores. In addition, the 14 principal scores, one peer score, and four test scores were all significant in

differentiating between the first two levels. None of the indicators was significant in differentiating between Levels II and III.

Cross Validated Discriminant Functions

The cross validated discriminant function hit rates are shown in Tables 5A through 5F. The original and cross validated hit rates are summarized in Table 6. The baseline cross validated hit rate using all the variables was 84.5%.

The cross validated hit rate for two of the other analyses (omitting observations and omitting dialogue scores) exceeded the base line value! Of course one expects less shrinkage in a cross validation if fewer variables are used. Nevertheless, obtaining these values suggests that the original function using the observations (and perhaps the dialogues) capitalized on random error.

The data presented in Table 6 suggest several possible options for making career ladder decisions with less data: (1) omitting only observations, (2) omitting only consensus scores, (3) omitting only dialogue scores, and (4) omitting both observations and consensus scores. The hit rate of 63% (25.5% false positives and 11.5% false negatives) suggests it may be unwise to eliminate all data from the external evaluators. More will be said about this in the discussion section

---

Insert Tables 5A through 5F and 6 About Here

Canonical Correlations

Some educators believe that the "bottom line" of teacher effectiveness is what actually goes on in the classroom and that external observations are needed to obtain such information. Thus, for these educators, the observation data represent the best measures for Career Ladder decisions. But, again, such data are costly to obtain. Perhaps other data, appropriately weighted, could serve as a surrogate set (objective 3).

Canonical correlation "measures the strength of the overall relationships between the linear composites of the predictor and criterion sets of variables" (Hair, Anderson, & Tatham, 1987, p. 187). Canonical correlations between the observations (dependent variables) and all the other data (independent variables) were obtained for 512 cases. (Twenty three of the original 535 were dropped due to missing data.) The results are summarized in Table 7. The first four canonical covariates produced Rs of .90, .64, .61, and .45 respectively (significant $\leq$ .015). Note that these values reflect the variance explained in the linear composites, not the original variables. The percent of variance (redundancy index) of the dependent variables (observational data) explained by the non observation data was 23.8% by the first canonical variate and cumulated to 29.3% by the first four variates. For the independent variables the corresponding values were 20.3% for the first root and 24.1% cumulative variance explained by the dependent (observation) variables.

Insert Table 7 About Here

While there are no generally accepted standards for the minimum acceptable redundancy index we believe the data explain too small a percent of the variance to advocate using one set of scores as a surrogate for the other set. Consequently, we did not cross-validate the canonical correlation equations.

Conjunctive vs. Compensatory Models

The final purpose of the study (objective 4) was to explore whether different decisions would have been made under a purely conjunctive model. To estimate an optimal conjunctive solution an iterative procedure was used to vary the cut scores in various domains until the resulting contingency table provided a maximum number of hits.

However, computer iterations are expensive and time consuming. Further, we know from data presented earlier, that one can eliminate some of the data sources and still make decisions highly comparable to the actual decisions made in Tennessee. Because fiscal viability in data gathering is essential regardless of the data combination model employed, we obtained hit rates separately for two components of the domain scores: data scores and principal scores. We also obtained hit rates for the total domain scores. This type of analysis is not reported for consensus scores because these scores are not procedurally independent from the dialogue and observation scores. For this analysis it was not feasible to eliminate any of the instruments which went into the data scores (such as observation or dialogues) because separate scores were not readily available for these as instruments.

For each set, prior probabilities of the hit rate were obtained separately for each domain by setting the cut score between Career Ladder

Levels 2 and 3 equal to the score obtained by the person who ranked 325th out of the 535 cases (because 325 individuals actually obtained Level 3 status) and by setting the cut score between Levels 1 and 2 the score of the 469th person (tied scores changed both ranks considerably in some instances). Next maximum hit rates were obtained for each domain by iteratively varying the cutoff values. Then, the best combinations of cut scores for two domains and for three domains were obtained. We could have continued the iterations for the best combination of five domains. But given the intercorrelations of the domains (Table 2), it seemed unlikely that the cross-validiated hit rate would have been much higher. The more relevant data from the conjunctive analyses are presented in Tables 8A to 8C.[2]

Insert Tables 8A to 8C About Here

As can be seen from the data in Tables 8A to 8C, domains 1, 2, and 4 seemed to provide the best hit rates (using the final decision as the criterion). Using only one domain score, domain 2 gave the best results for the data scores with an 81.2% hit rate. Domain 1 worked best for the principal scores and total domain scores providing hit rates of 63.0% and 86.8%, respectively. The best combination of two domains for the data scores was domains 1 and 2 with a hit rate of 86.1%. For principal scores it was a tie between 1 & 4 and 2 & 4 with a hit rate of 63.3%. For domain scores it was 90.0%. The best combination of three domains was domains 1, 2, and 4 for all types of scores. For the data scores, the hit rate was 88.3%. For the principal scores the hit rate was 63.7%, and for the domain scores it was 91.9%.

Given the weightings of the domain scores to obtain the composites (.15, .35, .15, .25, and .10 respectively) it was not surprising that domains 2 and 4 should give some of the better hit rate data. Given the intercorrelations of the domain subscores to each other and the composite (Table 2B) one can understand why adding the second and third subdomain scores did not serve to increase the hit rate very much. As pointed out before, there seems to be convergence of the data across domain within method of data gathering. That is to say, the convergent/discriminant validity data support the importance of method and are dissonant with the hypothesis that the domains measure different constructs. This finding is consistent with results reported by Furtwengler and others (1985, 1986).

We cross-validated the total domain score classifications with the conjunctive model using the iteratively derived optimum cut scores from the first sample. A contingency table showing the cross-validated hit rates is given in Table 9. Note that using only three domains the cross-validated hit rate was 86.3% (down from 91.9% in the first sample). That was higher than the cross-validated discriminant function using all the data (Table 5A).

Insert Table 9 About Here

DISCUSSION

Objectives 1 & 2

Because external evaluators comprised the major expense of the evaluations in the career ladder program, we analyzed the data with various portions of their tasks omitted to determine how the decisions might differ.

The evaluators gathered the observation and dialogue data on the same school visit. Training was needed for both these tasks. However it is difficult to separate out the costs of the two pieces of data. Further, conducting the dialogues could conceivably have had some impact on the observation scores and vice versa. The consensus data were gathered by bringing the evaluators together at the close of the overall evaluation process. This was a time consuming and expensive process. It required extensive training time and travel money. Because the consensus ratings were completed <u>after</u> the observations and dialogues the consensus discussions could <u>not</u> have impacted the other two sets of data. The evaluators' perceptions from the observations and dialogues were used in reaching the consensus.

The results of the original and cross validated discriminant analyses suggested several possibilities regarding making career ladder decisions with greater fiscal viability.

(a) Dropping observations results in a cross validated hit rate of 85.7% which exceeded the baseline cross validated hit rate (84.5%) with all scores included. Thus, dropping observations would be wise if such actions would not result in less valid dialogue or consensus scores. This would result in considerable savings.

Any decision regarding the elimination of observations must consider the existing philosophy about the value of observations in career ladder decisions (or any other personnel decision) and the psychometric characteristics of the observation scores. Fourteen states now mandate observations for certification (Sandefur, 1986) and individuals supporting such mandates would surely hesitate to give them up. Others believe observations are invalid and/or ethically and legally improper (Macmillan & Pendlebury, 1985; Scriven;

1987). Current research suggests that extensive observations are necessary to provide observation scores of sufficient reliability for use in personnel decisions (see, for example, Capie and Ellett, 1987).

(b) Dropping consensus scores provides a cross validated hit rate of 81.2% compared with the base rate cross validated hit rate of 84.5%. This would result in considerable savings, but less savings than dropping the observations. However, it would not impact any of the other data.

(c) Dropping dialogue scores results in a cross validated hit rate of 84.6% which exceeds the cross validated hit rate using all variables. Again, if such an action did not reduce the validity of observation or consensus scores it would be a cost effective move.

(d) Omitting observation and consensus scores results in a cross validated hit rate of 76.6%. There are 3.7% more (of the total) false positives and 1.2% more false negatives using this approach than are obtained in the cross validation using all the variables. Observations and consensus scores are the two most expensive pieces of data to gather. State agencies should consider whether the benefit (assuming all variables as weighted give us something closer to the "Truth") outweighs the cost.

(e) With omission of all the data provided by the external evaluators (observations, dialogues, and consensus) the cross validated hit rate was 63%, there were 25.5% false positives and 11.5% false negatives. While it is possible the principals would have been more rigorous had there not been outside evaluators, the opposite seems more likely. The finding here certainly is in keeping with the results from other states. When an outside evaluator is not used the resulting distribution of scores is so negatively

skewed as to provide almost useless data. It would probably not be reasonable to choose this option.

Objective 3

Professionals differ strongly with respect to the value, and importance, of observations. Although dropping the observations from the discriminant function does not negatively impact the cross validated hit rate, the reasonably low canonical correlations (and redundancy indices) suggest that the observations can not be used as a surrogate for the other data--or vice versa.

Initially it may seem the discriminant functions and the canonical correlations produced contradictory results. How can dropping a reasonably independent variable (as the canonical correlations suggest) not lower the hit rate in a discriminant function? At least one explanation is that the variation in observation scores was largely occasion variance. Whatever the variances in the observation variables indicate it is obvious that using only the observation data would result in basing the career ladder decision on something quite different than that indicated by the total set of data used in Tennessee.

Objective 4

Because of the way Tennessee obtained, scored, and aggregated data it was not feasible to compare the multiple cutoff (conjunctive) and compensatory approaches by instrument. However, using the conjunctive approach at the domain level indicates that a reasonably high hit rate can be obtained by using only data scores from three domains (88.3%). The total domain scores,

again using only three domains, produced a hit rate of 91.9%. Cross-validated this became 86.3%.. This was higher than the cross-validated discriminant function hit rate. Although w do not have a totally congruent comparison between conjunctive and compensatory models, the data we do have do not support the use of a compensatory over a conjunctive model.

SUMMARY AND CONCLUSIONS

The Tennessee Career Ladder Evaluation System was an extremely high stakes evaluation. Considerable controversy existed within the state over whether there should be a career ladder system and whether adequate evaluations could be performed. It was essential that any developed and implemented system be based on concepts which were logically appealing using data collection methods which allowed for objective scoring processes. It was necessary given the political/educational climate to build a multiple data source system.

However, after the fact, it is useful to examine the system to determine whether a less costly and/or more easily explainable approach could serve in reaching essentially the same conclusions. In addition, a close look at the data furthers our understandings about the educational constructs (domains) measured and our ability to measure them. The results of our analyses should be useful in any revisions of the Tennessee system, to other states or districts planning a system, and indeed to theorists in educational personnel evaluation.

The descriptive analyses indicated that several indicators were not useful for making differential career ladder decisions. For example, six observation variables had means greater than 99 (100 point maximum), and the

principal mean ratings were very high (all domain means above 768 on a 200-800 scale). This descriptive information alone would be useful for any evaluation system revision.

The correlation matrices of the domain scores (Table 2A) and the domain sub scores (Table 2B) indicate that the data are more impacted by method than trait, suggesting a lack of construct validity. Both practioneers and theorists need to consider the implications of this for future personnel evaluation.

The discriminant function analyses and the related structure correlation matrices indicated that omitting both the expensive-to-gather observation and consensus scores still allowed a cross-validated hit rate of 76.6% when the actual career ladder placement served as the criterion measure. With those scores omitted, the principals' scores were useful in differentiating between Career Ladder Levels 1 and 2, but not between Levels 2 and 3. The dialogue scores were important for differentiation at the higher level.

Although observation data could be omitted from the discriminant function without reducing the cross-validated hit rate, canonical analysis indicated that only 29.3% of the observation data variance was explained by the other data and that only 24.1% of the variance of the other data was explained by the observation data. A reasonable conclusion is that the variance for the observation data contained a large portion of occasion variance treated as error.

The conjunctive model data indicated that with an optimal cut score, one can reproduce the actual decisions made in Tennessee with fair accuracy (91.9% hit rate, cross-validated to 86.3%) using the domain scores for only three domains. Alternatively, the data scores alone for these three domains,

eliminating pricipal and consensus scores, could be used. This gave an original hit rate of 88.3%, which is somewhat lower, but would result in considerable savings.

All of these conclusions lead us to the following suggestions to consider in building future career-ladder evaluation systems:

1. Aggregate and report data by instrument rather than by domain.
2. Carefully evaluate the cost-effectiveness of any observation system.
3. Eliminate principal evaluations.
4. Use a conjunctive model to combine the data.

The above suggestions are given in decreasing order regarding their empirical support from our study. Certainly following one suggestion may impact the value of following another. For example, aggregating and reporting data by instrument rather than domain may greatly impact the decision regarding the choice of a data combination model.

Finally, we recognize all of these suggestions must be considered in light of the political and educational milieus and the fiscal viability of any evaluation system.

Footnotes

1 In addition, in order to make Level I, all applicants were required to pass a reading/writing test. Only two individuals in the sample failed on this criterion. Those who failed to pass generally dropped out of the evaluation process.

2 Note also that there were some optimum cut scores that included a range of values. We arbitrarily chose to use the lowest value that resulted in maximum accuracy. Using a different value in the range would change the ratio of false positives and false negatives.

References

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Capie, W. & Ellett, C. (April, 1987). The generalizability of two classroom observation systems for assessing merit teacher performance in Florida A paper presented at the annual meeting of the National Council on Measurement in Education. Washington, D.C.

Furtwengler, C., Malo, G., McLarty, J., & Strouss, S. (April, 1986). Multiple data sources in teacher evaluation. A paper presented at the annual meeting of the Nationl Council on Measurement in Education. San Francisco.

Furtwengler, C., McLarty, J., & Malo, G. (April, 1985). The career ladder program in Tennessee. A paper presented at the annual meeting of the National Council on Measurement in Education. Chicago.

Hair, J.F. Jr., Anderson, R.E. & Tatham, R. L. (1987). Multivariate Data Analysis (2nd ed.). New York: Macmillan Publishing Company.

McLarty, J., Furtwengler, C., & Malo, G. (April, 1985). Using multiple data sources in teacher evaluation. A paper presented at the annual meeting of the National Council on Measurement in Education. Chicago.

Mehrens, W. A., Phillips, S.E., & Anderson, A.E. (April, 1987). Conjunctive versus compensatory models for teacher licensure decisions: Monte Carlo and logical investigations. A paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Macmillan, C.J.B. & Pendlebury, S. (1985). The Florida Performance Measurement System: A consideration. Teachers College Record, 87, 67-78.

Sandefur, J.T. (1986). State assessment trends. AACTE Briefs, 7, (6), 12-14.

Scriven, M. (1987). Validity in personnel evaluation. Journal of Personnel Evaluation in Education, 1, 9-23.

Stufflebeam, D.L. & Brethower, D.M. (1987). Improving personnel evaluations through professional standards. Journal of Personnel Evaluation in Education, 1, 125-155.

Tennessee State Department of Education (1986a). Teacher Orientation Manual. Nashville, Tennessee, Author.

Tennessee State Department of Education. (1986b). Career Ladder Technical Manual: A Guide to Interpreting Your Scores, Teacher Edition. Nashville, Tennessee: Author.

Table 1
Descriptive Data for Some of the Scores from the Tennessee Evaluation Instrument (N=535)

Observation Score Number		Mean	St. Dev.	Min.	Max.	Possible Range
6		99.82	1.04	83.7	100	0-100
7		94.95	7.60	50.0	100	0-100
14		99.98	.04	99.7	100	0-100
19		99.64	.67	94.1	100	0-100
22		99.97	.72	83.3	100	0-100
23		99.90	.13	99.4	100	0-100[1]
27		99.56	1.44	83.7	100	0-100
Domain #	Principal					
1	Planning	778	35.4	525	800	200-800
2	Teaching Strategy	777	[illegible].7	460	800	200-800
3	Evaluation	772	38.7	550	800	200-800
4	Classroom Management	775	46.0	350	800	200-800
5	Leadership	772	38.9	550	800	200-800
	Consensus					
1	Planning	681	104	300	800	200-800
2	Teaching Strategy	667	112	200	800	200-800
3	Evaluation	662	112	200	800	200-800
4	Classroom Mangaement	670	122	200	800	200-800
Final Domain Score						
1		668	105	358	793	200-800
2		679	85	350	782	200-800
3		692	75	373	788	200-800
4		701	78	340	796	200-800
5		732	52	451	795	200-800
Composite Score		690	71	376	780	200-800

[1] N = 37 for this case.

Table 2A

Correlation Matrix of the Domain Scores (N=535)

Domain	1	2	3	4	5	Composite
1 Planning		.738	.646	.669	.443	.847
2 Teaching Strategy			.727	.806	.431	.947
3 Eval.				.769	.377	.842
4 Classroom Management					.369	.906
5 Leadership						.510

Table 2B

Correlation Matrix of Domain Subscores

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Data - P	-	.59	.45	.51	.39	.71	.66	.64	.57	.27	.27	.21	.25	.26	.78
2. Data - TS		-	.58	.71	.38	.68	.67	.63	.63	.33	.35	.26	.30	.39	.89
3. Data - E			-	.68	.29	.56	.55	.65	.53	.26	.24	.19	.23	.24	.73
4. Data - CM				-	.34	.64	.62	.61	.64	.27	.30	.21	.29	.22	.83
5. Data - L					-	.38	.34	.35	.25	.32	.31	.28	.24	.29	.48
6. Consensus - P						-	.87	.86	.78	.35	.36	.27	.33	.31	.87
7. Consensus - TS							-	.85	.78	.34	.34	.25	.32	.28	.86
8. Consensus - E								-	.78	.36	.38	.28	.35	.32	.84
9. Consensus - CM									-	.31	.37	.24	.37	.26	.81
10. Prin - P										-	.87	.83	.75	.78	.41
11. Prin - TS											-	.82	.89	.76	.43
12. Prin - E												-	.73	.75	.33
13. Prin - CM													-	.67	.41
14. Prin - L														-	.37
15. Composite															-

P = Planning
TS = Teaching Strategy
E = Evaluation
CM = Classroom Management
L = Leadership

Table 2C-1
Correlation Matrix of the 61 Indicators

Observation	1	2	3	4	5	6	7	8	10	11	12	13	14	15	16	17	18	19	21	24	25	26	27
1		.59	.34	.25	.17	.04	.31	.44	.63	.06	.28	.22	.35	.18	.54	.24	.28	.33	.34	.61	.44	.09	.21
2			.40	.21	.22	.10	.47	.42	.83	-.27	.40	.12	.43	.26	.77	.32	.36	.40	.43	.83	.62	.06	.31
3				.19	.30	-.02	.24	.13	.36	-.11	-.28	.19	-.17	-.25	.36	.15	.14	.11	.31	.36	.22	-.05	.12
4					.00	.04	.05	.14	.29	.02	-.03	.87	-.29	-.00	.26	.05	.04	.24	.08	.23	.07	.31	.04
5						-.05	.19	.13	.16	.03	.04	-.14	-.12	-.16	.17	-.07	.17	.14	.12	.17	.16	-.09	.05
6							.09	.07	.11	.11	.06	-.03	-.01	-.03	.05	.08	-.00	-.01	.07	.13	.09	.02	.10
7								.16	.45	.14	.13	-.01	.27	.10	.33	.09	.24	.22	.19	.41	.35	.09	.20
8									.42	.40	.64	.07	.22	.29	.39	.52	.20	.22	.36	.41	.29	.01	.20
10										.29	.39	.22	.51	.27	.75	.39	.38	.49	.48	.83	.58	.11	.33
11											.46	-.06	.13	.21	.26	.40	.16	.14	.32	.33	.29	-.07	.14
12												-.05	.22	.38	.39	.58	.19	.20	.36	.37	.25	-.06	.16
13													.21	-.07	.16	.06	-.03	.16	.05	.15	-.02	.32	.04
14														.20	.48	.22	.50	.88	.16	.38	.23	.21	.29
15															.24	.35	.24	.20	.35	.24	.20	.11	.16
16																.29	.35	.46	.40	.72	.53	.10	.24
17																	.21	.25	.50	.30	.21	.04	.18
18																		.54	.22	.31	.19	.09	.36
19																			.17	.37	.23	.19	.24
21																				.47	.40	.12	.12
24																					.70	.02	.30
25																						.15	.20
26																							.08
27																							

Table 2C-2
Correlation Matrix Of The 61 Indicators

	Dialogue										Peer	Study			PST				PDL	
	1	2	3	4	5	6	7	8	9	11		1	2	3	1	2	3	4	1	2
Observation																				
1	.49	.45	.43	.48	.46	.49	.48	.40	.41	.39	.23	.23	.19	.23	.21	.13	.14	.17	.25	.28
2	.58	.59	.50	.59	.60	.58	.58	.52	.51	.49	.25	.29	.29	.30	.27	.20	.12	.18	.31	.33
3	.28	.31	.25	.24	.22	.20	.25	.21	.25	.22	.10	.10	.09	.09	.12	.07	.00	.08	.09	.13
4	.08	.07	.01	.13	.09	.14	.08	.11	.05	.12	.02	.03	-.06	-.07	.08	.10	.09	.03	.02	.04
5	.07	.15	.15	.17	.16	.16	.18	.12	.13	.14	.03	.13	.11	.17	.00	-.04	-.11	-.05	.11	.05
6	.13	.10	.10	.13	.08	.06	.07	.07	-.00	.05	.07	.03	.03	.05	.15	.14	.08	.16	.06	.10
7	.37	.35	.33	.40	.35	.36	.32	.27	.29	.25	.21	.17	.20	.20	.31	.23	.15	.32	.26	.29
8	.32	.31	.28	.33	.32	.32	.32	.32	.29	.31	.11	.18	.16	.19	.05	.06	.05	.05	.18	.23
10	.35	.54	.50	.58	.59	.57	.56	.52	.49	.50	.22	.29	.28	.29	.24	.15	.09	.16	.25	.26
11	.18	.19	.19	.19	.22	.17	.16	.23	.24	.24	.02	.13	.21	.19	.00	.02	-.05	-.00	.05	.08
12	.23	.29	-.22	.26	.26	.26	.24	.33	.33	.31	.11	.17	.18	.19	.04	.05	-.03	.05	.11	.17
13	.07	.03	-.06	.11	.04	.11	.05	.08	-.01	.03	.06	-.05	-.14	-.16	.10	.14	.16	.07	.02	.03
14	.25	.25	.24	.25	.25	.29	.27	.33	.30	.33	.14	.21	.20	.19	.06	.07	.05	.06	.16	.14
15	.15	.19	.19	.17	.21	.16	.20	.26	.28	.27	.02	.21	.24	.24	.04	-.02	-.12	-.00	.06	.07
16	.47	.51	.43	.48	.49	.37	.46	.47	.48	.48	.14	.28	.30	.29	.15	.10	-.03	.09	.20	.24
17	.20	.20	.20	.21	.26	.21	.24	.26	.22	.24	.03	.15	.15	.14	.01	.02	-.05	.02	.05	.08
18	.19	.19	.21	.29	.29	.32	.30	.26	.27	.26	.12	.24	.25	.28	.10	.06	.01	.11	.14	.17
19	.23	.23	.23	.27	.29	.31	.28	.30	.27	.31	.13	.22	.22	.21	.03	.04	.02	.05	.14	.11
21	.29	.34	.31	.30	.34	.29	.31	.31	.29	.33	.08	.32	.34	.33	.05	-.00	-.09	.01	.12	.15
24	.55	.56	.50	.58	.58	.57	.56	.51	.49	.51	.25	.28	.32	.32	.22	.16	.06	.16	.26	.30
25	.44	.48	.47	.47	.50	.46	.48	.41	.41	.45	.16	.32	.37	.36	.13	.08	.01	.06	.22	.24
26	.07	.03	-.06	.04	-.05	.04	-.03	-.06	-.09	-.06	.07	-.06	-.15	-.13	.07	.06	.13	.05	.05	.03
27	.23	.17	.19	.21	.22	.16	.20	.23	.18	.17	.08	.13	.14	.14	.19	.13	.09	.16	.10	.17
Dialogue																				
1		.84	.70	.32	.33	.35	.35	.34	.34	.31	.28	.20	.22	.21	.25	.17	.17	.19	.31	.32
2			.75	.33	.35	.36	.37	.34	.35	.34	.28	.21	.22	.21	.17	.10	.09	.14	.30	.31
3				.34	.40	.35	.40	.34	.38	.35	.24	.26	.31	.28	.19	.04	.06	.14	.27	.27
4					.80	.83	.82	.36	.35	.33	.21	.25	.25	.27	.28	.18	.18	.22	.28	.31
5						.79	.83	.38	.39	.39	.19	.26	.31	.30	.23	.15	.08	.15	.24	.27
6							.82	.38	.37	.35	.24	.26	.27	.27	.26	.17	.14	.19	.28	.30
7								.38	.37	.36	.22	.25	.28	.29	.22	.15	.09	.13	.29	.32
8									.86	.77	.18	.24	.29	.26	.16	.07	.07	.10	.30	.30
9										.76	.16	.24	.31	.29	.18	.08	.09	.12	.29	.29
11											.12	.24	.28	.27	.05	.01	-.00	.06	.22	.23
Z Peer												.16	.16	.17	.05	.08	.04	.11	.19	.20
Z Study																				
1													.80	.83	.03	-.04	-.04	.02	.19	.22
2														.86	.06	-.01	-.07	.02	.15	.19
3															.05	-.02	-.08	.01	.16	.19
ZPST																				
1																.47	.47	.54	.27	.27
2																	.45	.51	.18	.16
3																		.45	.26	.21
4																			.25	.23
PDL																				
1																				.72
2																				

TABLE 2C-3
Correlation Matrix of the 61 Indicators

	PI														CONS				
Observation	1	2	3	4	5	6	7	8	9	11	12	13	14	15	1	2	3	4	Composite
1	.27	.25	.25	.26	.24	.23	.25	.19	.16	.22	.25	.23	.24	.24	.60	.61	.57	.53	.64
2	.28	.26	.26	.28	.25	.24	.26	.20	.20	.23	.25	.25	.20	.26	.76	.73	.72	.66	.78
3	.12	.13	.10	.11	.10	.09	.10	.08	.10	.10	.09	.12	.10	.14	.32	.32	.28	.30	.34
4	.17	.14	.09	.16	.13	.19	.16	.10	.04	.18	.14	.15	.10	.12	.20	.25	.20	.28	.23
5	.02	.02	.05	.03	.07	.02	.07	.05	.06	.08	.07	.03	.07	.05	.16	.17	.14	.15	.17
6	-.00	-.01	-.03	.02	-.01	-.02	-.01	-.02	-.03	-.03	-.02	-.01	-.02	.01	.12	.20	.11	.06	.14
7	.20	.18	.21	.26	.18	.16	.14	.16	.19	.13	.15	.21	.21	.22	.42	.39	.36	.31	.47
8	.14	.13	.13	.17	.14	.14	.14	.11	.10	.13	.16	.13	.14	.16	.38	.37	.36	.35	.44
10	.28	.27	.25	.31	.25	.26	.27	.20	.20	.24	.27	.23	.23	.27	.75	.75	.71	.73	.79
11	.02	.01	.06	.04	.03	.03	.05	.04	.01	.02	.01	.02	.01	.04	.26	.24	.23	.21	.29
12	.14	.12	.10	.13	.06	.12	.11	.06	.08	.10	.10	.09	.12	.16	.35	.33	.34	.31	.40
13	.18	.17	.10	.18	.14	.18	.15	.10	.05	.17	.14	.17	.11	.11	.13	.18	.13	.20	.15
14	.18	.16	.19	.23	.21	.20	.22	.13	.16	.23	.25	.15	.14	.12	.47	.48	.48	.67	.57
15	.03	.01	-.03	.01	-.01	.00	-.01	-.01	.03	-.02	.02	-.03	.04	.06	.27	.26	.28	.31	.32
16	.22	.20	.21	.22	.20	.22	.23	.16	.14	.21	.22	.18	.14	.21	.63	.61	.59	.63	.67
17	.09	.08	.06	.08	.03	.05	.05	.03	.02	.03	.05	.01	.06	.06	.31	.31	.30	.31	.39
18	.18	.16	.19	.21	.25	.22	.22	.15	.13	.24	.28	.10	.16	.14	.43	.45	.44	.50	.51
19	.18	.17	.21	.23	.23	.21	.24	.14	.16	.25	.25	.12	.13	.11	.43	.45	.47	.66	.56
21	.10	.11	.13	.10	.11	.08	.09	.08	.09	.09	.11	.10	.06	.11	.40	.42	.37	.35	.46
24	.26	.26	.26	.26	.25	.24	.25	.19	.18	.23	.23	.24	.20	.25	.72	.71	.65	.62	.74
25	.16	.13	.17	.14	.13	.09	.14	.12	.13	.09	.12	.13	.09	.17	.54	.54	.50	.46	.57
26	.09	.11	.05	.11	.13	.14	.13	.05	.02	.11	.11	.09	.08	.09	.04	.04	.03	.10	.06
27	.06	.04	.08	.12	.07	.07	.07	.02	.05	.09	.13	.04	.05	.06	.33	.28	.30	.35	.37
Dialogue																			
1	.23	.21	.21	.25	.21	.23	.21	.18	.19	.18	.20	.22	.19	.26	.58	.51	.49	.45	.61
2	.19	.19	.18	.19	.18	.19	.16	.14	.12	.15	.16	.17	.15	.20	.57	.52	.49	.45	.61
3	.15	.12	.17	.17	.16	.16	.12	.12	.16	.10	.13	.15	.13	.17	.54	.51	.47	.43	.58
4	.26	.25	.20	.28	.23	.21	.23	.20	.18	.24	.26	.22	.22	.23	.59	.59	.54	.48	.66
5	.25	.24	.21	.26	.23	.21	.23	.20	.20	.17	.22	.21	.17	.21	.59	.57	.57	.50	.66
6	.27	.25	.21	.28	.25	.24	.26	.21	.21	.24	.27	.25	.23	.24	.57	.57	.55	.48	.66
7	.24	.23	.18	.25	.22	.19	.21	.19	.17	.19	.23	.18	.19	.21	.57	.58	.54	.49	.66
8	.24	.22	.26	.27	.21	.23	.18	.19	.19	.21	.21	.21	.18	.24	.55	.54	.67	.50	.59
9	.21	.17	.23	.23	.17	.19	.14	.19	.20	.15	.17	.16	.16	.20	.53	.49	.62	.45	.57
11	.20	.20	.27	.24	.22	.22	.18	.29	.19	.20	.21	.18	.16	.20	.55	.53	.62	.51	.57
Z Peer	.29	.29	.29	.30	.34	.28	.23	.29	.25	.23	.25	.28	.22	.30	.28	.28	.27	.19	.34
Z Study																			
1	.15	.15	.13	.17	.17	.18	.21	.13	.15	.18	.25	.28	.22	.30	.28	.28	.27	.19	.34
2	.12	.13	.14	.14	.14	.13	.15	.10	.16	.12	.20	.13	.15	.18	.33	.35	.31	.35	.50
3	.15	.15	.15	.15	.18	.16	.19	.14	.17	.17	.15	.12	.09	.16	.36	.34	.35	.34	.49
ZPST																			
1	.12	.12	.06	.12	.09	.10	.10	.11	-.10	.09	.20	.12	.12	.19	.38	.35	.35	.34	.51
2	.11	.07	.03	.10	.07	.06	.07	.07	.04	.04	.08	.13	.13	.16	.23	.21	.20	.13	.30
3	.12	.14	.05	.10	.08	.10	.07	.07	.06	.07	.02	.07	.10	.08	.11	.12	.09	.07	.23
4	.15	.14	.14	.17	.15	.15	.14	.10	.09	.10	.05	.10	.13	.12	.10	.11	.09	.02	.16
PDL 1	.20	.22	.19	.22	.18	.20	.16	.16	.17	.14	.11	.13	.16	.11	.17	.18	.16	.06	.26
2	.24	.22	.23	.22	.20	.22	.20	.17	.16	.15	.13	.20	.20	.24	.30	.29	.31	.24	.40
PI 1		.84	.77	.79	.75	.74	.68	.76	.61	.58	.60	.73	.70	.67	.34	.32	.35	.29	.40
2			.77	.78	.78	.78	.77	.76	.64	.61	.62	.79	.69	.70	.31	.30	.33	.29	.38
3				.74	.76	.68	.66	.75	.60	.53	.56	.75	.59	.63	.31	.33	.33	.30	.38
4					.80	.80	.73	.77	.67	.70	.74	.74	.67	.70	.37	.35	.38	.34	.43
5						.81	.78	.77	.59	.72	.75	.70	.66	.65	.31	.30	.32	.31	.39
6							.82	.72	.62	.72	.74	.74	.62	.63	.31	.30	.34	.35	.38
7								.68	.56	.80	.81	.69	.63	.59	.32	.31	.34	.35	.39
8									.63	.60	.62	.72	.64	.66	.25	.23	.26	.22	.31
9										.49	.51	.66	.59	.60	.24	.23	.25	.22	.29
11											.91	.55	.54	.47	.29	.28	.30	.33	.35
12												.59	.56	.51	.32	.30	.34	.38	.40
13													.64	.71	.27	.28	.30	.27	.34
14														.67	.28	.24	.28	.25	.32
15															.30	.27	.30	.23	.35
CONS 1																.87	.86	.78	.87
2																	.85	.78	.86
3																		.78	.84
4																			.81

Table 3A

Predicted Career Ladder Level Using Discriminant Function Analyses With All Data

Actual Group	Predicted Group Membership 1	2	3	Total	Canonical Correlation	p
1	55	10	0	65		
2	3	119	14	136	.902	.000
3	0	8	303	311	.446	.000
Total	58	137	317	512		

Percent of cases correctly classified	93.2%
Percent of false positives	4.7
Percent of false negatives	2.1

Table 3B

Predicted Career Ladder Level Using Discriminant Function Analysis & Omitting Observation Scores

Actual Group	Predicted Group Membership 1	2	3	Total	Canonical Correlation	p
1	55	10	0	65		
2	6	114	16	136	.892	.000
3	0	15	296	311		
Total	61	139	312	512	.351	.003

Percent of cases correctly classified	90.8%
Percent of false positives	5.1
Percent of false negatives	4.1

Table 3C

Predicted Career Ladder Level Using Discriminant Function Analyses & Omitting Consensus Scores

Actual Group	Predicted Group Membership 1	2	3	Total	Canonical Correlation	p
1	56	9	0	65		
2	4	119	13	136	.883	.000
3	0	16	295	311		
Total	60	144	308	512	.410	.002

Percent of cases correctly classified	91.8%
Percent of false positives	4.3
Percent of false negatives	3.9

Table 3D

Predicted Career Ladder Level Using Discriminant Function Analyses & Omitting Dialogue Scores

Actual Group	Predicted Group Membership 1	2	3	Total	Canonical Correlation	p
1	54	11	0	65		
2	7	109	21	137	.886	.000
3	0	10	305	315		
Total	61	130	326	517	.414	.000

Percent of cases correctly classified	90.5%
Percent of false positives	6.2
Percent of false negatives	3.3

Table 3E

Predicted Career Ladder Level Using Discriminant Function Analyses & Omitting Observation & Consensus Scores

Actual Group	Predicted Group Membership 1	2	3	Total	Canonical Correlation	p
1	54	10	1	65		
2	9	105	22	136		
3	0	22	289	311	.844	.000
Total	63	137	312	512	.283	.158

Percent of cases correctly classified	87.5%
Percent of false positives	6.4
Percent of false negatives	6.1

Table 3F

Predicted Career ladder Level Using Discriminant Function Analyses & Omitting Observation, Dialogue & Consensus Scores

Actual Group	Predicted Group Membership 1	2	3	Total	Canonical Correlation	p
1	36	16	13	65		
2	13	43	81	137		
3	3	22	290	315	.663	.000
Total	52	81	384	517	.207	.532

Percent of cases correctly classified	71.4%
Percent of false positives	21.3
Percent of false negatives	7.4

Table 4

Structure Correlation Matrix for Analysis 3E

Dialogue	Function 1**	Function 2
5	.51 *	-.09
7	.50 *	-.20
6	.49 *	-.18
4	.48 *	-.10
2	.48 *	-.22
1	.48 *	-.17
3	.40 *	-.13
8	.40 *	-.18
9	.38 *	-.29
11	.37 *	-.07
Student		
2	.34 *	.15
1	.32 *	.29
PDC		
2	.26 *	.07
1	.26 *	.03
Peer		
	.20 *	.06
Prin		
15	.19 *	.04
13	.17 *	.16
9	.16 *	.10
Pre Skills Test		
1	.16 *	.09
Prin		
14	.15 *	.11
Pro Skills Test		
3	.08 *	.00

	Function 1**	Function 2
Prin		
12	.20	.44 *
5	.20	.41 *
11	.18	.40 *
7	.20	.37 *
Student		
3	.34	.37 *
Prin		
1	.21	.33 *
3	.20	.32 *
4	.22	.32 *
2	.20	.31 *
6	.20	.25 *
Pro Skills Test		
4	.12	.21 *
Prin		
18	.15	.16 *
Pro Skills Test		
2	.12	.13 *

** Function 1 primarily differentiates Levels II and III whereas function 2 primarily differentiates Levels I and II.

* Indicates the loading adds significantly to the discrimant function at the .05 level of statistical significance.

Table 5A

Cross Validated Discriminant Function Hit Rate Using All Data

Actual Group	Predicted Group Membership 1	2	3	Total
1	41	20	3	64
2	7	118	25	150
3	0	24	272	296
Total	48	162	300	510

Percent of cases correctly classified	84.5%
Percent of false positives	9.4%
Percent of false negatives	6.1%
Max. Chance Hit Rate	58.0%

Table 5B

Cross Validated Discriminant Function Hit Rate Omitting Observation Scores

Actual Group	Predicted Group Membership 1	2	3	Total
1	45	16	3	64
2	6	122	22	150
3	0	26	270	296
Total	51	164	295	510

Percent of cases correctly classified	85.7%
Percent of false positives	8.0%
Percent of false negatives	6.3%
Max Chance Hit Rate	58.0%

Table 5C

Cross Validated Discriminant Function Hit Rate Omitting Consensus Scores

Actual Group	Predicted Group Membership 1	2	3	Total
1	34	27	3	64
2	8	113	29	150
3	0	29	267	296
Total	42	169	299	510

Percent of cases correctly classifed	81.2%
Percent of false positives	11.6%
Percent of false negatives	7.3%
Max. Chance Hit Rate	57.8%

Table 5D

Cross Validated Discriminant Function Hit Rate Omitting Dialogue Scores

Actual Group	Predicted Group Membership 1	2	3	Total
1	46	19	1	66
2	9	111	31	151
3	0	19	278	297
Total	55	149	310	514

Percent of cases correctly classified	84.6%
Percent of false positives	9.9%
Percent of false negatives	5.5%
Max. Chance Hit Rate	57.8%

Table 5E
Cross Validated Discriminant Function Hit Rate
Omitting Observations and Consensus Scores

Actual Group	Predicted Group Membership 1	2	3	Total
1	36	25	3	64
2	6	105	39	150
3	0	31	265	296
Total	42	161	307	510

Percent of cases correctly classified	76.6%
Percent of false positives	13.1%
Percent of false negatives	7.3%
Max. Chance Hit Rate	58.0%

Table 5F
Cross Validated Discriminant Function Hit Rate
Omitting Observations, Dialogue & Consensus Scores

Actual Group	Predicted Group Membership 1	2	3	Total
1	26	18	22	66
2	22	38	91	151
3	5	32	260	297
Total	53	88	373	514

Percent of cases correctly classified	63.0%
Percent of false positives	25.5%
Percent of false negatives	11.5%
Max. Chance Hit Rate	57.8%

Table 6

Original and Cross Validated Discriminant Functions

	Hit Rate	Percent False Pos.	Percent False Neg.
All Variables			
Original	93.2	4.7	2.1
Cross-Validated	84.5	9.4	6.1
Omitting Observations			
Original	90.8	5.1	4.1
Cross-Validated	85.7	8.0	6.3
Omitting Consensus Scores			
Original	91.8	4.3	3.9
Cross-Validated	81.2	11.6	7.3
Omitting Dialogue Scores			
Original	90.5	6.2	3.3
Cross-Validated	84.6	9.9	5.5
Omitting Observ. & Consensus			
Original	87.5	6.4	6.1
Cross-Validated	76.6	13.1	7.3
Omitting Obs, Dialogue, & Consensus			
Original	71.4	21.3	7.4
Cross-Validated	63.0	25.5	11.5

Table 7

Summary Results of the Canonical Correlation Analysis (N=512)

Canonical Variates	Canonical R	p	Percent Ys Explained by Xs (Redundancy Index)	Percent Xs Explained by Ys (Redundancy Index)
1	.90	.000	23.8	20.3
2	.64	.000	1.7	1.4
3	.61	.000	2.9	2.0
4	.45	.015	.9	.4
			29.3	24.1

Table 8A

Classification Using the Conjunctive Model: Data Scores

Priors	Domain	Cut Score 1 vs 2	Cut Score 2 vs 3	Percent Classified F. Neg.	Hit	F. Pos.
Planning	1	441	702	13.3	73.7	13.0
Teaching Strategy	2	566	707	10.3	79.3	10.3
Evaluation	3	614	710	16.0	67.7	16.4
Classroom Management	4	615	717	12.8	72.9	13.7
Leadership	5	707	728	19.0	62.0	19.0
Cutoff-Best Single Domain	2	573	702	5.6	81.2	13.2
Cutoff-Best Combination of Two Domains	1&2	363,573	583,690	7.3	86.1	6.6
Cutoff-Best Combination of Three Domains	1,2,4	360,444,600	570,690,702	6.8	88.3	4.9

Table 8B

Classification Using The Conjunctive Model: Principal Scores

Priors	Domain	Cut Score 1 vs. 2	Cut Score 2 vs. 3	Percent Classified F. Neg.	Hit	F. Pos.
Planning	1	735	781	19.9	57.3	22.7
Teaching Strategy	2	742	781	20.3	58.3	21.4
	3	715	750	15.4	59.0	25.6
Evaluation	4	732	778	21.1	55.6	23.3
Classroom Management	5	719	750	18.1	57.1	24.8
Cutoff-Best Single Domain	1	704	720	2.3	63.0	34.8
Cutoff-Best Combination of Two Domains	1&4	704,660	716,720	3.8	63.3	32.9
Tied With	1&2	704,700	716,710	3.8	63.3	33.8
Cutoff-Best Combination of Three Domains	1,2,4	704,700,660	728,706,720	4.3	63.7	32.0
Tied With	1,4,5	704,660,600	728,725,710	4.9	63.7	31.4

Table 8C

Classification Using the Conjunctive Model: Total Domain Scores

Priors	Domain	Cut Score 1 vs 2	Cut Score 2 vs 3	Percent Classified F. Neg.	Hit	F. Pos.
Planning	1	511	700	11.7	77.1	11.3
Teaching Strategy	2	573	695	7.1	85.7	7.1
Evaluation	3	615	705	10.9	78.2	10.9
Classroom Management	4	611	716	8.3	83.5	8.3
Leadership	5	712	737	19.5	61.3	19.2
Cutoff-Best Single Domain	2	550	693	4.5	86.8	8.6
Cutoff-Best Combination of Two Domains	1&2	465,550	598,686	4.7	90.0	5.3
Cutoff-Best Combination of Three Domains	1,2,4	465,500,591	620,678,700	6.2	91.9	1.9

TABLE 9

Cross Validated Hit Rate Using the Conjunctive Model With Three Domain Scores*

Actual Decision (Level)	Conjunctive Decision (Level) I	II	III	Row Total
I	50	15	1	66
	9.5	2.9	.2	12.5
II	22	126	10	158
	4.2	24.0	1.9	30.0
III		24	278	302
		4.6	52.9	57.4
Column Total	72	165	289	526
	13.7	31.4	54.9	100.0

* The three domains were Planning, Teaching Strategies, and Classroom Management. The first number is the number of individuals in the cell. The second number is the percent of individual in the cell.

Percent of cases correctly classified	86.3%
Percent of false positives	4.9%
Percent of false negatives	8.8%