An Empirical Examination of

Alternative Measures of Job Performance

by

Diana L. Deadrick

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Ph.D.

in

Management

APPROVED:

K. Dow Scott, Chairman	T. W. Bonham, Ph.D.
Frederick S. Hills, Ph.D.	Robert M. Madigan, Ph.D.
Raymond H. Myers, Ph.D.	Dianna L. Stone, Ph.D.

March 26, 1987

Blacksburg, Virginia

HGW 10-5-87

An Empirical Examination of

Alternative Measures of Job Performance

by

Diana L. Deadrick

K. Dow Scott, Chairman

Management

(ABSTRACT)

This research addresses the dual aims of selection research: the understanding and prediction of job performance. Two areas of research regarding criterion construct validity are examined and a research model is developed in an attempt to integrate this literature. This research model formalizes suggestions made by James (1973) and sets forth different levels, referents, and methods for criterion validation. A series of hypotheses regarding the interrelationships among alternative job performance measures and the relationships between criteria and predictors are presented.

A longitudinal study was conducted to test this Job Performance Model in a field setting. Five measures of job performance and six ability tests for performance prediction were examined for sewing machine operators in a garment manufacturing plant. Data analyses indicated: High convergent validity among multiple methods of job performance measurement when the level of specificity was matched; Low to insignificant predictability of the alternative job performance criteria; and Differential prediction of job performance, depending on the method and referent for performance evaluation.

It was concluded that measurement characteristics of job performance criteria represent boundary conditions for subsequent prediction. The model presented here has merit for addressing the interrelationships among multiple performance criteria as well as the relationships between criteria and performance predictors.

Acknowledgements

The completion of this research exercise finally affords me the time to stop, sit back, and give thanks to the numerous "support systems" that have helped me see my way through this past year. First, I want to thank my committee members for their confidence in me when I had none and their support when I was floundering. These are the people that I have aspired to call my colleagues. Second, I am indebted to the project team. Just when I thought there were no more hours in the day, these people gave their time and devotion to the "cause." Third, I want to recognize the project participants for making my dream a reality. The V.E.C. provided the financial resources and the B.V.U.S.A. provided the data resources for this research. Special thanks go to

and my family. These are the people who listened to my apprehensions, helped me deal with the setbacks, and believed in my dreams as much as I do.

My dissertation, and my career, are dedicated to the memory of my father:
This man instilled in me the perseverance and self-confidence that got the words on paper. In his memory, I am reminded that "There is a joy, too, in loneliness."

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

It is important on both theoretical and practical grounds that selection research incorporate the pursuit of performance construct validity along with performance prediction. The performance construct is admittedly multidimensional (Thorndike, 1949; Dunnette, 1963), dynamic (Prien, 1966; Ghiselli and Haire, 1960), and hence problematic (Smith, 1976). In fact, several authors have addressed the issue of "what are we measuring" and have come to no firm conclusions (i.e., Weitz, 1961; Kane, 1982). Given the lack of consensus about the definition of "performance," a detailed examination of different forms/concepts of job performance criteria is necessary. Guion (1976) points out that if selection research is to follow the scientific method, then rational hypotheses must be generated based on the explication of the criterion constructs of interest--not on the predictor variable at hand. The logic of hypothesis development would therefore dictate that the construct validity of performance be addressed prior to the empirical validation of predictor-criterion relationships.

Many authors (Guion, 1976; Dunnette, 1963; Smith, 1976) have stressed that the overriding aim of selection research is the prediction of performance. After all, the practical application of selection research findings has a great impact on the ability composition of a workforce. Thus the errors of prediction are to be minimized on an institutional, if not individual, basis. Yet the traditional methodology for this research is correlational, thus assuming a meaningful/parsimonious criterion variable. Dunnette (1963) has criticized this "classic validation model" as being too simplistic and has suggested that future research follow a different methodology whereby the nature, and shape, of multiple predictor-criteria relationships can be examined. A more informative method would assess not only the degree of association between the variables but also the magnitude of prediction errors.

Purpose

The purpose of this study is to directly examine the interrelationships between multiple criteria of job performance and to assess "validity" in terms of the quality of prediction. The model used herein is based largely on the criterion development efforts of James (1973). This model is in fact a multiple criterion model for job performance whereby multiple levels (i.e., global effectiveness criteria and dimensional results criteria) and multiple methods (i.e., judgmental and nonjudgmental) of performance measurement are examined in order to determine the degree of convergence, thus the equivalence of performance criteria. This model has been extended to encompass multiple referents for performance (typical and maximal) in order to incorporate the differences in effort with regard to observed (rated) performance criteria. A "typical" construct of job performance refers to a measure of central tendency, which by its very nature smooths-out the peaks and troughs of short-term performance and assumes an average to be the best estimate of the true value (Kane, 1982). In contrast, a "maximal" performance referent relates to a peak level of performance achieved during a period of time. Conceptually, this construct refers to a feasible (attainable) level of performance that may or may not be maintained on a consistent basis, depending on the motivational and situational constraints that occur on the job. The point here is that while traditional selection models have relied on a typical measure of performance, a maximal referent of performance might be more indicative of the level of performance that "can be achieved" versus the level that "was achieved". Because these different referents are conceptually distinct criteria, an empirical examination of their interrelationship is warranted.

Validation refers to the process of providing evidence that a selection instrument is related to, or predictive of, subsequent job performance (Dreher and Sackett, 1983). One type of validation evidence is termed criterion-related validity, which relies on comparing performance on some selection instrument with performance on the job. Guion (1976) argues that, in practice, a selection procedure (predictor) is initially chosen for examination and then a convenient performance measure is selected as the criterion of interest. This convenient performance measure is almost always

some kind of judgmental performance rating (Guion, 1976), and the validity of judgmental ratings is still in question (Bernardin and Beatty, 1984).

The validity assessment used in this study is two-part. First, the ***interrelationships among multiple criteria*** are examined using measures of fit and prediction. The "quality of fit" between multiple criteria relates to how these performance measures covary; i.e., does an increase in actual output coincide with an increase in the perceived rating of performance output? While a correlation might indicate the strength of the relationship between different criteria, it does not assess the predictive relationship between them. Therefore the validity of criteria is also examined in terms of the "quality of prediction"; e.g., is an increase in the actual quantity of output predictive of the perceived rating of quantity of output? This dual examination of criterion relationships is a means of investigating the construct validity of performance measures--the convergence between multiple methods of the same performance construct (i.e., judgmental v. nonjudgmental measures). In addition, the divergence of different performance measures (i.e., dimensional v. global, and/or typical v. maximal) can be examined in the same manner. The outcome of this criterion validity assessment is a better understanding of the relationships among the underlying criterion constructs, which is a prerequisite to assessing the validity of hypotheses regarding a predictive relationship between performance criteria and selection predictors.

The second validity assessment focuses on the ***relationships between criteria and predictors.*** This investigation follows the traditional strategy of criterion-related validity yet is extended beyond a reliance on the "validity coefficient" (correlation). As noted previously, the validity coefficient is indicative of the quality of fit yet does not provide information about the quality of prediction. Therefore, predictive validity is examined in terms of the correlation between predictors and criteria and also in terms of the errors of prediction in a predictor-criterion model. It is possible that while a given predictor-criterion relationship is marked by a statistically significant correlation, the inherent errors of prediction might be so large as to outweigh the apparent significance of the relationship. Because selection research is primarily concerned with prediction, this validity assessment is

extended to include prediction errors as a means of determining the significance of the predictor-criterion relationships.

Significance

This research is significant for both theoretical and practical reasons. With respect to theory development, it is important to critically examine (and question) the relationships between alternative measures of job performance apart from their relationship to various performance predictors. Judgmental performance ratings are by far the most commonly used method of performance measurement, in spite of research which indicates that these measures are subject to both intentional and inadvertent bias (see Landy and Farr, 1983). In fact, several authors suggest that it is because of this reliance on performance ratings that selection validation findings are weak (Smith, 1976) and generalizable findings are impaired (Guion, 1976). An effort to examine and understand the relationships between alternative methods of performance measurement is a first step toward developing improved measures of performance constructs and strengthening our understanding of predictor-criterion relationships. From a theoretical view, it is also important to explicitly investigate any differential prediction of certain job performance criteria. By extending the concept, and methodology, of validation to encompass errors of prediction, the robustness and stability of our selection models can be examined.

On a practical level, this criterion model for selection research is important for a thorough understanding of the selection process itself. The usefulness of selection research is determined by the quality of information it provides to the decision-maker. It is important to know whether a given predictor is better for predicting a maximal (peak) level of performance or for predicting a typical (average) level of performance. It is also important to examine the degree of equivalence between different performance constructs. While a given predictor might be found to predict future performance, the question remains as to whether future performance measured with a supervisory

rating is substitutable for an objective record of performance. An investigation aimed at examining such equivalence will add to our knowledge of rater accuracy as well as criteria convergence. While an ultimate, bias-free performance criterion is not feasible, the differential power of a selection instrument to predict these alternative performance measures provides insight as to what the predictor is predicting.

In addition, the limiting (boundary) conditions that have an impact on the practical utility of a selection model might take the form of the ***purpose*** of the selection procedure (prediction of one aspect of performance versus overall performance), the performance ***referent*** of importance (a maximal, peak level versus a typical, average level), or the available ***method*** of measuring/evaluating performance (judgmental ratings versus nonjudgmental production records). These conditions might influence the final hire/reject decision. If different criterion variables do in fact influence the predictor-criterion relationships, this information should be used in the matching process for selection decisions. And if these different criterion variables are differentially related to the same predictors, then generalizable validity of predictors across different criteria is open to question.

Summary

To summarize, the purpose of this research is to address the dual aims of selection research: an ***understanding*** of job performance and the ***prediction*** of job performance. With respect to understanding, this research extends the model of selection beyond a simplistic predictor-criterion assessment and examines multiple criteria of job performance. Although several authors have stressed the need for an explication of job performance criteria and a more scientific approach to validation (Guion, 1976; Smith, 1976; James, 1973), most of the existing validation research follows a simplistic model which correlates predictors with criterion measures. As a result, our understanding of the underlying relationships between predictors and criterion constructs is hampered. For instance, are dimensional performance constructs (i.e., specific results) equivalent to the more

global performance constructs (i.e., general outcomes)? If not, a question arises as to whether a strong correlation between a selection procedure and judgmental performance ratings (global performance) signifies that that procedure is also strongly related to a more specific performance index (dimensional performance). These different levels of criteria might also be measured using different methods, and the equivalence (substitutability) of these multiple performance measurements also affects our understanding of the selection process. For example, is an objective count of performance output convergent with a subjective rating of performance output? Not only does this question relate to our understanding of criterion equivalence but also relates to our knowledge of rater accuracy.

With respect to the second aim of selection research, prediction, this study directly examines the quality of prediction by directly examining theoretical boundary conditions. Guion (1976) suggests that the traditional reliance on bivariate correlations and situational performance ratings has resulted in a technology of employee selection rather than a science. Such research would appear to focus on "test evaluation" rather than subsequent performance prediction; and by focusing on "testing," we neglect the multiple facets of job performance (Guion, 1976). An ideal paradigm for selection research is a validation strategy that incorporates both prediction and explanation; a paradigm whereby performance prediction is achieved through an understanding of the linkages between multiple criteria. In this study, it is postulated that job performance prediction might be hampered by these simplistic correlational models. Boundary conditions such as the different levels of criterion measurement (dimensional and global) and/or the different methods of measurement (objective and subjective) might limit the "power" of prediction. A direct examination of these potential boundary conditions vis a vis multiple levels and methods of performance measurement will help to establish the range of conditions under which our selection models hold, thus aiding our prediction as well as understanding of job performance.

This chapter has outlined the purpose and significance of this research study. ***Chapter 2*** reviews the literature that is pertinent to the topic of criterion validity and then presents the research model used for this study. Prior research dealing with the validity of job performance measures and the

statistical techniques for examining the predictive validity of job performance criterion models are summarized. Based on this review, the conceptual model and research hypotheses are presented as the framework for this study.

Chapter 3 presents the methodology used for the empirical test of the proposed model. In this chapter, the research setting is described, the measurement procedures are discussed, and the statistical analyses employed for each research hypothesis are delineated. ***Chapter 4*** contains the empirical results of this study. The first section presents the research findings regarding the relationships among alternative job performance measures. The following section deals with the relative predictability of these performance measures. ***Chapter 5*** provides a discussion of the research findings with implications for future selection research. The limitations of this study are noted as well as recommendations for the design of future selection research.

Chapter 2: Literature Review and Model Development

Overview

This chapter reviews the literature relevant to the issues of the construct validity of job performance criteria and the quality of prediction. The first section, "Construct Validity of Selection Criteria," presents conceptual models for criterion validity and empirical studies dealing with this topic. The second section, "Performance Rating Criteria," presents research on the validity of performance appraisal data. In this section, rating validity is investigated through the process of establishing "criteria for the criterion." The empirical studies selected for this review are investigations of the convergence among alternative methods of measuring job performance.

The third section, "Quality of Prediction," presents a review of statistical techniques available for analyzing the predictive capability of research models. Topics such as residual analysis and prediction errors are applied to the selection aims that guide pursuant research. The last section, "Conceptual Model and Hypotheses," presents the research model used in the present study. The conceptual basis for this model is based on the work of James (1973). The resultant hypotheses address the issues of criterion validity as well as predictive (predictor-criterion) validity.

Construct Validity of Selection Criteria

> The uncritical acceptance of the predictive paradigm (of validation) is too often accompanied by an equally uncritical acceptance of the criterion measure, and many criteria are charitably described as casual (Guion, 1976, p. 789).

An explication of the nature of the performance criterion establishes boundary conditions for selection research models (theories) and addresses the question as to ***what*** are we predicting. The

critical issues here are criteria convergence and equivalence: Can we substitute criterion measures of performance as if these different measures are equivalent to one another? Without first knowing whether one measure of performance, for instance production output records, is equivalent to another, such as judgmental ratings of production performance, there is a danger that the underlying models (theories) we are testing are in fact conceptually different selection models. If this is the case, then the resultant predictor-criterion findings are based on different criterion constructs and, thus, are not necessarily equivalent.

Weitz (1961) highlights the fact that the criterion parameter might in fact represent a boundary condition on our theories. The validity of a particular theory can be examined more thoroughly if we explore the differential impact of alternative criterion measures on the adequacy of the model. Using an example from verbal learning theory, Weitz shows how the significance of the research findings can be altered if the nature of the criterion variable is "manipulated;" i.e., if a different "type" of criterion measure is selected and/or a different level of the criterion is chosen. Applying this direct approach of investigating the criterion parameter to selection research would entail examining different measures, and levels, of job performance in our selection models in order to explore the boundary conditions of these models. Rather than concluding that a particular predictor is not "valid" for predicting job performance, it could be that a particular predictor is predictive of certain types, or levels of job performance but not others. The implications of this concern for the criterion parameter are that:

1. "Laws of criteria" can be generated which specify the bounds (scope) of our selection models,
2. "Cleaner hypotheses" can be stated which specify the impact of an independent variable (predictor) on a range of performance criteria, and
3. "Conceptually relevant" criteria can then be selected from the boundary conditions under which a selection model holds.

The outcome of this research is a step toward the "scientific" pursuit of the validity of not only our selection models of selection but also the criterion parameter itself. This is in line with Guion's (1976) call for "rational hypothesis testing" as a means of moving selection from a technology to a science.

Criterion Models

Weitz (1961) and others (e.g., Smith, 1976; James, 1973; Guion, 1976; Dunnette, 1963) have presented eloquent arguments for a paradigmatic shift in selection research. Rather than pursuing research aimed at test validation, more explanatory, thus scientific, research should follow a construct validation approach. Such research is aimed at understanding the criterion parameter of our selection models before attempting to explain the relationship between the observed variables. Noting that selection research is atheoretical and empirically-driven, James (1973) stressed the need for selection models that focus on the performance criteria and address such critical issues as ***what to measure*** and ***what has been measured.*** This author identified, and analyzed, two existing criterion models which traditionally guide selection research. The Ultimate Criterion Model (Blum and Naylor, 1968; Nagle, 1953; Thorndike, 1949; Toops, 1944) is based on an overall success criterion, which is a linear composite of the dimensional job criteria. In contrast, the Multiple Criterion Model (Ghiselli, 1956; Guion, 1961, 1965, 1976; Dunnette, 1963) is based on maintaining the multiple, independent dimensions underlying the performance criteria. James (1973) points out that these two criterion models are similar with respect to criteria collection; both models require the collection of multiple criteria. The key difference therefore is not "what to measure" but rather what is done with the criteria after they have been obtained (James, 1973, p. 75). The Ultimate Criterion proponents assume an underlying general factor which accounts for all of the important variance in work behavior, thus "justifying" the use of the overall composite performance criterion. Yet the multiple criterion proponents make no such assumption; they argue that job performance dimensions are factorially independent (Ghiselli, 1956; Turner, 1960) and that the "dimensionalities" of criteria are a means for understanding the underlying job behaviors and subsequent performance (Dunnette, 1963; Ghiselli, 1956; Wallace, 1965). Although these two models appear to be opposites, and have in fact been the basis for an on-going controversy in the selection literature, they are not dissimilar with respect to criterion development. Yet both models are seriously deficient in terms of identifying "what to measure" (James, 1973). As a result, both models neglect the substantive nature of the criterion parameter altogether.

James (1973) describes a General Criterion Model, which was originally developed by Campbell, Dunnette, Lawler, and Weick (1970) to examine the determinants of managerial effectiveness/performance. The key feature of this criterion model, as compared to the other two, is the explicit focus on "what to measure" for performance criteria. This model (see Campbell et al., 1970, for full details) depicts performance criteria on three measurement levels: job behaviors, job performance, and organizational outcomes. These distinct ***levels*** of measurement provide a framework for studying the different dimensions of job performance. This framework (multiple levels) is equivalent to Weitz's argument for examining and establishing boundary conditions. Different "types" (levels) of performance criteria may represent different performance constructs which might alter the adequacy of our selection models. By investigating performance along these different levels, we are making no assumptions regarding the equivalence of criteria. In fact, this criterion model stipulates a direct examination of the equivalence of alternative performance measures. It also examines the range of criterion (performance) conditions under which a selection model will hold.

In order to fully examine the construct validity of performance criteria, James (1973) integrated the General Criterion model with the Multiple Criterion model, thus suggesting a framework for developing future predictor-criterion "theories." This integration requires that multiple criteria be obtained on each of the different measurement levels, to be investigated either individually or as independent dimensions. Furthermore, various methods of measurement should be used for collecting criterion data. This suggestion of multiple methods is a necessary condition for examining the construct validity of measures (i.e., Campbell and Fiske, 1959) in addition to the degree of measurement contamination (James, 1973; Bernardin and Beatty, 1984).

In summary, the model (paradigm) for selection research outlined by James (1973) contains:

- multiple levels of criteria measurement (therefore "what to measure");
- multiple criteria per level; and
- multiple methods of measurement.

The result is a model aimed at developing job performance criteria, examining the relationships between different levels of performance, and investigating the constructs underlying performance criteria (James, 1973).

Whereas James' paradigm for selection research targets the criterion parameter as the focal point for selection research, Borman, Rosse, and Abrahams (1980) advocate a paradigm focusing on the "linkages" between predictors and criteria as the focal point. Citing the arguments, and need, for a construct validation approach to selection research, Borman et al. (1980) suggest a process of discovering-understanding-confirming these relationships. The features of this research strategy are:

1. Criterion development efforts to identify the underlying performance constructs of success,
2. Predictor development efforts that generate constructs related to these criteria, and
3. Linkage analyses that seek to replicate and confirm the predictor-criterion model.

It is interesting to note that both James (1973) and Borman et al. (1980) are advocating a construct validity approach to selection, yet these authors have identified a different focal point for "theory development." James (1973) argues for directed efforts at the construct validity of the criterion parameter, much like the suggestion of Weitz regarding "laws of criteria" which will provide a basis for "cleaner hypothesis" development. On the other hand, Borman et al. (1980) have directed their efforts at the construct validity of the predictor-criterion linkage as a means of understanding the constructs. This is apparent from examining the criterion development phase of his study. Although a multidimensional framework for criteria was employed, thus preserving the dimensionality of performance, the criterion measures were all performance ratings. Criterion development consisted of examining different performance appraisal formats (anchored and unanchored) and different performance raters (peer and supervisory), which was "replicated" with different samples of similar individuals (two samples of Navy recruiters). However, in terms of construct validity, only one method was employed thus reducing the confidence in the criterion measures. James (1973) has pointed out that "ratings from different sources could have acceptable convergent and discriminant validity solely on the basis of similar incorrect inferences of behavior" (p. 80). The lack of independent methods hampers the determination of adequate operational definitions (the corre-

spondence between observables and constructs) as well as "constitutive" definitions (the correspondence between constructs) (James, 1973, p. 80).

Summary: The foregoing models represent recent conceptual developments that are aimed at "understanding" the constructs of selection. There have been arguments for a new paradigm for selection research, one that is concerned with the construct validity of the parameters themselves as well as the parameter relationships. Weitz's (1961) concern for establishing "boundary conditions" via the criterion parameter and Guion's (1976) call for "rational hypotheses" via criterion conceptualization have been met, in part, by the conceptual efforts of James (1973) and Borman et al. (1980). To date, there has been no concerted effort to test the models described above.

Construct Validation Research

There is a vast amount of empirical research in the selection area, yet the majority of these studies are focused on predictor development and/or "test validation" (i.e., James, 1973; Smith, 1976; Guion, 1976). As noted by James (1973) and Prien (1966), only secondary attention has been given to criterion research; thus our understanding of ***what*** has been measured by the performance criteria is limited. Prien points out that most selection researchers spend considerable time and effort developing predictors with near-perfect reliability and demonstrated factorial multidimensionality. Yet these same researchers will evaluate these elegant predictors with "the shoddiest, partial, immediate, or proximate criterion" (Prien, 1966, p. 502). Because most empirical selection research has focused on predictors, the studies to be reviewed here are limited to those studies that are explicitly aimed at examining multiple criteria in the context of construct validation. This review does not include research efforts directed at composite (ultimate) versus multiple criteria; the purpose herein is to explicitly examine multiple criteria. Nor does this review include studies regarding the dynamic character of criteria; the focus here is on the convergence among different levels of measurement rather than the temporal stability of the constructs.

The Multitrait-Multimethod (MTMM) approach to construct validity has been applied to selection in an attempt to discover the validity of job performance criteria. The most notable application of the MTMM approach is that of Lawler (1967). Noting that judgmental performance ratings are the most frequently used criterion measure, Lawler (1967) modified the MTMM matrix approach for the purpose of learning more about the "meaning" of ratings and understanding the rating process. This approach is the Multitrait-MultiRATER (MTMR) method, where multiple raters have been substituted for multiple methods. Although several studies of the MTMR method have been undertaken (e.g., Gunderson and Nelson, 1966; Lawler, 1967; Nealey and Owen, 1970; Kavanagh, MacKinney, and Wolins, 1971; Zedeck and Baker, 1972), only two exemplar studies will be reviewed here. The first study is Lawler's MTMR approach to measuring managerial job performance (1967). The second is Borman's "Hybrid Matrix" approach for rating individuals (1974). It should be noted that because these MTMR studies utilize only one method of criteria measurement (ratings), they are not construct validation studies per se.

In 1967, Lawler developed the MTMR approach as a means of understanding what ratings mean, thus focusing solely on the rating process. In his monograph, Lawler described how this approach yields information about the construct validity of selection criteria (performance ratings) through an examination of convergent and discriminant validities. Lawler's first example of this MTMR analysis is based on research by Tucker, Cline, and Schmidt (1967) in which three rater groups (superiors, peers, and subordinates) evaluated the performance of research scientists on four different traits. Using only the peer and superior ratings, Lawler described the results of this study via the MTMR matrix analysis. By examining the data in this way, he concluded that (1) the peer and superior ratings do not correlate "high enough" to infer that convergent validity exists, and (2) the data do not "appear to satisfy the requirements for discriminant validity either" (Lawler, 1967, p. 373). The advantage of analyzing the data in the MTMR matrix is the insight gained regarding "what" the criterion is measuring. Specifically, Lawler points out that the low convergent validities indicate that the superiors and peers are "seeing quite different things." Furthermore, the multitrait data indicate a "large halo tendency" that would not have been evident if only one trait had been

examined. In summary, Lawler is making the point that the MTMR approach to analyzing this data enables a more thorough analysis of the rating results. The judgmental rating process, and judgmental rating criterion, can be examined with the intent of detecting instances of rater convergence and trait discriminability (Lawler, 1967).

In a second example, Lawler analyzeds his own data on a group of middle- and top-level managers. Multiple ratings (superior, peer, and self ratings) were obtained on three different traits (quality of job performance, ability to perform, and effort put forth on the job). The MTMR analysis indicateds that:

1. Superior and peer ratings have good convergent validity;
2. The validity diagonal (monotrait-heterorater) values are higher that the correlations found in the heterotrait-heterorater triangles; and
3. The same pattern of trait interrelationships is exhibited in all of the heterotrait triangles.

Lawler cites these findings as being suggestive that judgmental ratings of managerial job performance "can achieve a level of measurement that is aspired to, but infrequently obtained" (p. 375).

Criticisms of the MTMR approach to construct validity center around the basic concern regarding "what" has been measured. The use of multiple raters, rather than multiple methods, inhibits a determination of exactly "what" has been observed/measured (i.e., James, 1973; Bernardin and Beatty, 1984). James (1973) argues that this approach fails to establish formal connections between the observed variables and the underlying constructs, thus preventing any understanding, and evaluation, of the constructs themselves (p. 80). Bernardin and Beatty (1984) also note this possibility of "shared errors," and further criticize MTMR studies as being "one-shot" approaches for construct validation. Construct validity is a process, and additional analyses are required to identify the explanatory constructs underlying these observed performance variables.

Borman (1974) reviewed much of the research using the Multitrait-Multirater (MTMR) matrix as applied to performance ratings. Noting that an assessment of the "validity" of performance ratings is the intended purpose, Borman inferred that analyses such as these might be misleading.

Specifically, the MTMR approach for performance ratings "may be ignoring or incorrectly interpreting valid differences in perceptions between organization levels" (p. 187). Therefore Borman's "hybrid" MTMR approach subgroups the different raters according to their organization level; ***within-level*** interrater agreement is thus the index of convergent validity. Borman (1974) stated that the benefits of this "hybrid" analysis include a more realistic analysis of the quality of ratings and an "improved conceptual fit" for assessing performance rating validity (see Borman, 1974, p. 105). Borman stressed that the within-level interrater agreement is more appropriate than demanding across-organizational level agreement. Across-level agreement would imply that all raters share the same view of performance on all dimensions of the job, an assumption that is tenuous.

Borman's study (1974) focused on rating secretaries (n = 41) from five different departments. Two different rater groups were used: department instructors (supervisor ratings) and the secretaries themselves (peer ratings). Each organizational level (instructors and secretaries) developed their own behavioral expectations rating scale, thus identifying dimensions and anchors that were relevant to that group's perception of secretarial performance. The performance rating process consisted of raters from both levels rating the secretaries on ***all*** of the performance dimensions; i.e., those developed by their own group as well as those dimensions prepared by the other group. The findings, as related to construct validity, are that:

1. Rater agreement was significantly higher on the raters' own dimensions than on the other level's (group's) dimensions,
2. There is a "reasonable amount of convergent validity", and "discriminant validity is also encouraging" (p. 115).

It should be noted that the results of this study were analyzed in two ways. First, the criteria matrix was analyzed using the hybrid matrix. These results were stated above. Second, the matrix was also analyzed on a relative basis by comparing the hybrid analysis with the traditional MTMR analysis, which collapses across levels. In this comparison between the "hybrid" matrix analysis and the MTMR analysis, "five of the seven convergent validity indices in the hybrid matrix were higher than their counterparts in the MTMR matrix validity diagonal," and "only 46% of the off-diagonal correlations in the MTMR matrix were smaller that the diagonal coefficients, compared to the 66%

figure computed for the hybrid matrix" (p. 116) The conclusion drawn here by Borman (1974) is that "the hybrid analysis was superior for this set of data" (p. 118).

Borman concluded that convergent and discriminant validity might be more easily established using the hybrid matrix because the raters are subgrouped into homogeneous groups based on their relationships with ratees. Furthermore, he suggested that this approach can be generalized to most situations in which two or more raters are available in each of two or more organizational levels. Because this approach is based on within-level agreement, the "conceptual fit" for examining construct validity is improved (Borman, 1974). This improved fit is based on the fact that raters at each level evaluate ratee performance on only those dimensions in which they are ***a priori*** in good position to make judgements. Borman further suggested that because raters at the same organizational level often observe ratee behavior under the same conditions, a performance dimension under this approach can be considered a "single construct" more so than a dimension on which members of different levels provide ratings. The raters at different organizational levels presumably "see" different samples of ratee performance--and this disagreement in ratings across organizational levels may result from "honest differences in orientation and perspective" (Borman, 1974, p. 118).

Borman's analysis has refined the MTMR approach, recognizing the impact of different "levels" on the meaning of a construct. By examining the impact of differences in opportunity to observe, disagreement (error) among raters can be partitioned into "honest" bias (differences in perspective) and "error" bias (within-level disagreement). However, this hybrid approach to construct validity is still "method-bound" and addresses only the question of performance rating validity rather than the more general criterion validity.

Summary: The foregoing studies of construct validity indicate that while efforts have been made to assess the validity of judgmental criterion measures, the results are inconclusive. First, the determination of convergent validity within the context of multiple-rater analyses is debatable. Lawler undertook across-level convergence as a means of establishing convergent validity while Borman argued for within-level convergence as the more conceptual analysis. Second, the analysis for

discriminant validity is loosely defined, both conceptually and analytically. Campbell and Fiske (1959) provide no definitive rules for determining discriminant validity, and the MTMR approaches have not refined this validity assessment either. Third, the substitution of multiple raters for multiple methods prevents any conclusive assessment of performance criteria convergence and construct validity. This criticism, which is detailed by James (1973), is based on the fact that multiple **methods** are a necessary condition for examining the convergence of measures on a single construct. Independent methods of measuring the same performance construct provide a more solid basis for inferring whether in fact the construct of interest was measured, or whether "shared errors" among performance raters were measured.

Because the foregoing studies utilize only one method of performance measurement (judgmental ratings), an understanding of the relationship between the observable variables and the underlying constructs is limited. Further, an understanding of the relationship between the underlying constructs themselves is hindered. Referring back to the intent of performance construct validity--an understanding of "what to" measure and "what has been" measured--it is apparent that these studies have not directly addressed the validity of job performance criteria.

Performance Rating Criteria

The most widely-used criterion measure in the context of selection research is the judgmental performance rating. However, these measures are subject to biases which restrict determinations of not only the predictor-criterion validity but also the criterion validity (i.e., Landy and Farr, 1983; Bernardin and Beatty, 1984). Guion (1976) suggests that this reliance on judgmental performance ratings has led to situationally-specific criterion measures that cannot be generalized across settings. It is apparent that if performance is measured by only this one method, we have neglected the need for understanding ***what*** has been measured. This statement is supported by the previous discussion of the Multitrait-Multirater (MTMR) approach; convergent validity assessments are often com-

posed of different referents (i.e., expectations, perspectives) for performance evaluation and carry no assurance that "true" performance levels have been observed and/or rated. One means of assessing the validity of judgmental performance ratings is to establish an objective criterion for the relevant rating criterion, assess the convergence between these different performance measurements, and continue this cyclical process for the construct validation of criterion measures utilizing different methods of performance measurement (Bernardin and Beatty, 1984).

Research efforts in the performance appraisal literature have largely focused on the cognitive aspects of the appraisal, investigating such issues as the judgmental rating process, rater training, and performance feedback. In addition, a great deal of attention has been devoted to the technical aspects of performance appraisal; i.e., instrumentation and bias. An emerging area of performance appraisal research deals with the more fundamental issue of performance rating validity. It is this area of the performance appraisal literature that is relevant to this study. The validity of performance appraisal data is an important component of the construct validation approach for establishing job performance criteria. James (1973) makes this point when he recommends that different methods of measuring performance be used, if for no other reason than to directly investigate sources of criterion contamination. Therefore, the appraisal literature reviewed here is limited to research that focuses on the validity of performance appraisal data; i.e., the convergence between judgmental rating criteria and other, "objective" measures of performance.

Performance Appraisal Validation Strategies

Bernardin and Beatty (1984) advocate establishing criteria for the criterion. These authors note that while such attempts have been suggested by others (i.e., Weitz, 1961), and numerous lists of evaluative criteria for performance measures have been developed (i.e., Blum and Naylor, 1968), there is still a great deal of ambiguity regarding the validity of appraisal data. In fact, these authors present a list of the most frequently cited "criteria for criteria," providing both the literal and operational definitions. One of these criteria is "validity," which the authors define as "the extent to

which ratings on an appraisal instrument correspond to actual performance levels for those who are rated" (p. 143). A construct validity strategy outlined by Bernardin and Beatty (1984) follows the concepts recommended by Campbell (1976) and incorporates the necessity of content-domain sampling. Bernardin and Beatty stress that this more comprehensive approach to validating performance appraisal data (and raters) is preferable to "one-shot" construct validity studies using the Multitrait-Multimethod, or rather Multitrait-Multirater, approach.

The process for the validation of judgmental performance ratings begins with ***content-domain sampling.*** (Bernardin and Beatty, 1984). For this first step, a job analysis is undertaken in order to determine the full scope/domain of each job and identify the salient dimensions upon which to measure performance; i.e., "what to measure." Kane (1980) developed a job-mapping procedure which provides the framework for job analysis, on the one hand, and simultaneously provides a guide for criterion specification and the development of performance criterion dimensions. This means of job analysis is a first-step toward developing a model (theory) of job performance, as evidenced by the explicit focus on the salient performance dimensions and the different levels of analysis. The job map is a hierarchical job analysis which begins with the identification of those broad functional components that define the most relevant aspects of a job. These functional components are independent job dimensions, and each of these broad functional dimensions is further defined in terms of subcomponents (tasks and duties). The resultant hierarchy of the job content (domain) is composed of different levels of the job. The top-level is the overall job, which is composed of, and defined as, a number of functions. These functions represent the different dimensions of the job. The functions are composed of, and defined as, a number of tasks and duties. These tasks/duties represent the different behaviors/behavior patterns that are carried out in order to perform the function, and hence perform the job. This content hierarchy is equivalent to James'(1973) recommendation for using different levels of performance measurement. For example, the job functions provide a basis for developing dimensional job performance criteria; the job tasks/duties provide the basis for developing the job behaviors.

The implications of this approach to job analysis for performance conceptualization are as follows. First, a job-mapping approach to job analysis defines a hierarchy of job components and their respective performance correlates. Second, the level of specificity of the job components generates different performance criteria, ranging from a job behaviors level to a global effectiveness level. Third, the interrelationships within and between these levels of performance provide a means of understanding, as well as predicting, individual performance.

The second step in the rating validation approach suggested by Bernardin and Beatty (1984) is the examination of the degree of ***convergence*** between judgmental ratings and objective data assessing the same or similar aspects (dimensions) of work. For this kind of convergent analysis, Bernardin and Beatty stress the need for hypotheses that specify the relationship between multiple measures of job performance. Whereas some validation studies have employed multiple criteria, and have correlated these criteria, there has been no conceptual foundation for such correlations and no discussion as to the meaning of significant/insignificant correlations. Therefore, these authors suggest that hypothesis development precede the mere statistical correlation of criteria. Such hypotheses would specify the expected nature of the convergence between ratings and objective performance measures and explain why these multiple measures should or should not converge.

The benefits of this type of research are that it provides evidence not only for rating validity but also for rater validity (Bernardin and Beatty, 1984). For instance, a high convergence between a matched dimension of performance, measured by both judgmental and nonjudgmental means, provides direct evidence about the degree of error (reliabilty) in ratings and the degree of accuracy of the rater. This approach to rating validity is preferable to a Multitrait-Multirater approach in that an objective criterion for the appraisal data has been established by which to validate the rating criterion. Thus a high degree of convergence among multiple measures of the same performance dimension supports the inference that ratings reflect actual performance levels on those performance factors, that ratings are "valid" measures of those aspects of performance that were rated, and that a particular rater's ratings are valid or invalid (Bernardin and Beatty, 1984).

In addition to the foregoing approach to construct validation of ratings, Kane (1982) has developed a new system of performance appraisal that provides a basis for making comparisons between the performance appraisals of job incumbents within the same job and between different jobs (Bernardin and Beatty, 1984). This Performance Distribution Assessment (PDA) method is a means of exploring, and examining, ***what to measure*** as well as what has been measured. The similarity of this method to the prior construct validity arguments made by Weitz (1961) and James (1973) warrants its inclusion here as an approach to the validation of appraisal data.

Two key features of the PDA method bear directly on the pursuit of construct validity of performance criteria. First, Kane (1982) explicitly deals with the issue of ***what to measure*** as job performance. The PDA is based on the job-mapping approach to job analysis. As a result, multiple dimensions of job performance are identified that are in fact the "conceptual criteria" foundation advocated by Astin (1964). The hierarchical structure for defining the job explicitly recognizes the level of specificity with which job dimensions are described and hence identifies the different levels for subsequent performance evaluation (i.e., from job behaviors to organizational outcomes). On a technical basis, this PDA method addresses the issue of "what to measure" by expanding the concept of performance beyond an average level of performance. Most performance appraisal methods (and most performance measurement systems) assume that the average level of performance over a period of time represents the "true" estimate of performance, and all variation around the mean is random (Kane, 1982). However, the PDA method provides a measurement system whereby the criterion parameter is viewed as a distribution of performance values. This focus on the performance criterion parameter is directly related to Weitz's argument for establishing boundary conditions for research models via the parameters of the criteria. The parameter estimates developed by Kane represent a means of examining the range of conditions under which the selection models hold and establishing "laws of criteria" upon which rational hypotheses for model testing can be based.

The second key feature of the Performance Distribution Assessment (PDA) is that it also addresses the question ***what has been measured.*** The performance distribution developed by Kane

directly defines the conceptual boundary limits for performance evaluation. An absolute zero point of performance would imply that the job incumbent is not performing at all. However, the PDA method defines this lower limit (zero point) of performance in terms of "minimally feasible" (allowable) performance levels. Furthermore, the performance distribution defines the upper limit for performance in terms of "maximum feasible" (achievable) performance levels. This system thus provides a frame of reference for the performance appraiser and avoids assessing only "typical" performance. By evaluating "feasible" performance in light of external constraints (opportunity bias), this method establishes a referent for measuring the ***distribution*** of performance; i.e., the "average" performance level achieved in addition to the "maximal" distribution of performance that was achievable during the appraisal period.

In summary, Kane's PDA is an operational procedure for addressing the construct validity of appraisal data; "what is" to be measured, and "what was" measured are two of the key features of this method. This method was developed for performance appraisal but has implications for both objective and judgmental sources of data (Bernardin and Beatty, 1984).

Method Convergence Research

The investigation of criteria equivalence is one approach to investigating the construct validity of job performance criteria. The degree of equivalence among multiple performance criteria must be established before attempting to substitute one performance criterion measure for another (Smith, 1976; Severin, 1952). Although many selection studies have employed multiple criteria for criterion development and/or validation purposes, most of them have not directly addressed the question of whether or not one criterion measure is equivalent to, hence substitutable for, another. The research gap regarding criteria substitutability is most evident in validation studies where a predictor is validated against either an expedient criterion measure and/or multiple criteria (objective and subjective) that have not been conceptually related or conceptually distinguished. For example, performance ratings are the most widely used criterion measure for validation studies, and yet the

meaning of such ratings is still in question and the situational nature of such evaluations is probable (i.e., Guion, 1976; Landy and Farr, 1980; Steel and Mento, 1986). Furthermore, when validation studies do employ multiple criteria, no attempt is made to conceptually relate these criteria and hence the question remains as to "what" has been measured (Bernardin and Beatty, 1984). The issue of criteria substitutability deals with the question as to whether "performance is performance is performance," and such global substitutions cannot be made without determining the degree of convergence among multiple measures (Smith, 1976).

Only a few studies have dealt directly with the question of the validity of performance appraisal data via convergent analyses with objective performance measures. Although some research on the relationship between objective and subjective measures of performance has been conducted in experimental situations (i.e., Scott and Hamner, 1976; Borman, 1978; DeNisi and Stevens, 1981), those studies are not included in this review for two reasons. First, the performance appraisal ratings of interest here entail a synthesis of information by one individual about the efforts, or performance of another. In this respect, it is necessary to consider a history of ratee/rater interactions during which time a long string of actions determine the performance ratings rather than a single, isolated action. Second, the performance appraisal situation of interest here is that which occurs in the natural work setting, devoid of experimental controls. Alexander and Wilkins (1982) have pointed out that the experimental studies consist of a controlled simulation of a work setting in which the performance action under study is highlighted to the rater while all other factors are held constant. While these studies are laudable insofar as influences on the rating process are concerned, they do not address the issue of rating validity as it relates to on-the-job performance assessments. The studies reviewed here consist of field research studies which have directly examined the degree of convergence among subjective and objective performance criteria.

In 1952, Severin investigated the relationships among several measures of job performance and concluded that criterion measures cannot be substituted for one another without knowledge of the degree of equivalence among those measures. The purpose of this study was, in part, to determine the equivalence of multiple measures of job performance, and his findings indicate that such

measures generally exhibit low correlations. The data for this study consisted of abstracts of prior research studies published in the ***Handbook of Employee Selection*** (1906-1948), publications in the literature since the publication of the ***Handbook,*** and military studies (total sample equals 144 studies). From these abstracts, nine job performance criteria were identified, which included both objective and subjective measures, and the results were summarized in a table of correlation coefficients. Some of the notable findings were:

1. The median correlation between training records and supervisory ratings was 0.11 (10 studies, with correlations ranging from 0.05 to 0.24);
2. The median correlation between production records and supervisory ratings was 0.47 (3 studies, with correlations ranging from 0.40 to 0.54);
3. The median correlation between subordinate ratings and supervisory ratings was 0.70 (12 studies, with correlations ranging from 0.55 to 0.84);
4. The median correlation of all correlations was 0.28 (144 studies, with correlations ranging from -0.10 to +0.84).

Based on these findings, Severin (1952) concluded that there is "strong evidence that one should never substitute one criterion for another without first determining if they are reasonably equivalent" (p. 245).

Turner (1960) undertook a multidimensional approach to criterion development. The purpose of this study was to construct job performance criteria in light of the multidimensional criterion assumptions, thus examining multiple dimensions, relevance weights, and generalizability across plants. Of interest here is the examination of the degree of equivalence among the objective and subjective measures of criterion dimensions. Turner (1960) found that there is little relationship among these different methods of job performance measurement even when they were conceptually equivalent. The data for this study consisted of 20 job performance criteria (11 objective, 9 subjective) for the job of production foreman, collected in two locations (n = 102 and 104). The objective criteria were operationalized on an operating-unit basis and include grievances, turnover, absences, scrap, efficiency, suggestions, hospital passes, disciplines, absenteeism, flexibility, expense tools, and expense processing supplies. The subjective measures were supervisory ratings (alternating rankings) on eight functional areas plus an overall performance rating. These eight areas

encompass quantity, quality, cost control, organization and planning, employee relations, cooperation with other supervision, safety, and housekeeping. The overall rating included performance on these eight areas plus "any other functions that the rater thought important" (p. 217). With regard to criteria convergence, Turner found that:

1. Objective measures had low correlations with other objective measures;
2. Objective measures had low correlations with the ratings; and
3. Objective cost measures were not significantly correlated with the cost control ratings (i.e., measures that logically cover similar kinds of job performance).

Based on these findings, Turner (1960) concluded that "ratings and objective data are not necessarily equivalent, even when they supposedly measure similar things" (p. 220); thus, "the equivalence of ratings and objective criterion measures should never be assumed" (p. 222).

Seashore, Indik, and Georgopoulos (1960) also examined the relationships among job performance criteria and found that the relationships are "generally small" and that the size and direction of these relationships is variable across similar units/organizations. The data for this study consisted of five job performance criteria for the nonsupervisory employees of a delivery service firm (i.e., loaders and drivers, with n = 975). The criteria were operationalized by four objective measures (productivity, accidents, absences, and "errors") and one subjective measure ("effectiveness"). The findings of this study that pertain to the individual level of analysis were:

1. Across stations (n = 975, collapsed across 27 stations), all criteria correlations are of a small magnitude thus indicating that there is little common variance among these performance measures (correlations range from .01 to .32);
2. Within stations (n ranges from 13 to 54 for each of the 27 stations), the criteria correlations remain low, indicating little common variance among the job performance measures (correlations range from .01 to .42).

Based on these findings, Seashore et al. (1960) concluded that the relationships among multiple criteria are small and highly variable (after taking into account measurement and sampling errors) and that these data "contradict the validity of overall job performance as a unidimensional construct" (p. 202). The authors further state that the use of a single job performance measure as a

"sample" of job performance is not justified "without prior determination of interrelations among different aspects of performance" (p. 202).

The last study reviewed here is that conducted by Alexander and Wilkins (1982). The purpose of this study was to determine the degree of validity of subjective measures of job performance--i.e., the degree to which actual performance differences account for variances in performance ratings. These authors found that there is "no relationship" between objective and subjective measures of job performance. The data for this study consisted of five job performance criteria for the job of vocational/rehabilitation counselors and communication scales which measure the "degree of liking" that exists between the rater and ratee (n = 87). The four objective performance measures were outcome measures regarding the number of client service applications, the number of client cases on active status, the number of nonseverely disabled client cases "closed" (placed in a job), and the number of severely disabled client cases closed. The subjective measure was a supervisory rating composed of seven trait-oriented items. Although it was not explicitly addressed in the analysis, it is assumed here that these authors used some kind of composite rating as the subjective evaluation. The findings of this study that pertain to the performance criteria are summarized below.

1. There is a low relationship between the objective and subjective performance measures (correlations range from 0.11 to 0.28); and
2. The objective measures do not generally predict the subjective evaluation.

Based on these findings, Alexander and Wilkins (1982) conclude that there is not a clear relationship between performance ratings and actual performance, which raises questions about the validity of subjective performance evaluations. Furthermore, these authors suggest that "ratings are potentially biased by the quality of the relationship between the supervisor and the worker" (p. 485).

Summary

Overall, these studies found that there is no clear relationship among alternative measures/methods of job performance. The low magnitude of the convergence between the ob-

jective and subjective criteria indicates that these job performance criteria are not equivalent, thus questioning the substitutability of criterion measures. On the surface, these findings suggest that selection models are not generalizable across different performance measures; the method of measurement for the criterion parameter appears to represent a "boundary condition" for selection research theories. However, a closer examination of the underlying data used in these studies indicates that the basis for selecting relevant criteria measures and the level of specificity for comparing criteria measures confound the nonequivalent results reached in these studies.

Each of the foregoing studies examined the degree of equivalence among objective and subjective performance measures, yet the criterion measures selected for subsequent convergent analyses were not conceptually congruent. First, the basis for criterion measurement can be criticized on the grounds of relevance. For example, Seashore et al. (1960) state that their criterion variables were chosen "because they have face validity, objectivity, and either measured or estimated high reliability" (p. 200). Alexander and Wilkins (1982) state that the "outcome measures represent the critical steps in the vocational rehabilitation process" (p. 488), yet the performance rating is focused on factors such as reliability, adaptability, and overall quality of job performance. The point here is that it is unclear whether all of the job performance criteria are "conceptually relevant" to actual performance on the job.

Second, the basis for judgmental performance ratings is ambiguous regarding the choice of rating dimensions and the composition of the resultant rating criterion measures. In Severin's study (1952), no information was provided about the underlying sample characteristics, criterion characteristics, or the use of a composite versus overall (one item) rating criterion. Seashore et al.'s (1960) performance ratings were, in fact, a rank-ordering of foremen on the overall quality of their job performance. Alexander and Wilkins (1982) used a seven-item rating scale which was not even based on the job in question (i.e., "borrowed" from prior research). Furthermore, these authors used a composite measure for the subjective evaluation without specifying the basis for combining items.

Finally, the job performance measures are not "matched" with regard to performance specificity. For example, Turner (1960) compared alternative performance measures within the same dimension of performance (i.e., objective and subjective measures of cost control), yet he used different levels of analysis for the objective and subjective performance measures. The objective job performance criteria were based on the performance of the foreman's operating unit whereas the subjective criteria were based on the individual foreman's performance. Seashore et al. (1960) did not even match the job performance dimensions in their analysis. These authors state that the relationships ***among*** the different aspects of job performance are generally small; yet if the measures are matched regarding similar performance dimensions, the data indicate a higher degree of convergence than the authors claim. For example, the subjective "effectiveness" criterion refers to the overall quality of job performance, and the objective "errors" criterion also represents the quality of job performance. For both the across-station and within-station analyses, the relationship between these measures is significant (-0.32 and -.042, respectively), thus indicating a moderate degree of convergence among "matched" job performance dimensions. Alexander and Wilkins (1982) also failed to match job performance dimensions. In their study, the subjective performance evaluation is based on traits, yet the objective performance evaluation is based on the "critical steps" of the vocational rehabilitation counselor. Severin's study (1952) summarizes prior research, and no information is provided regarding whether or not the criterion measures used for comparison were initially based on conceptual congruence.

Bernardin and Beatty (1984) have summarized the performance appraisal validity studies very aptly in the their critique of convergence studies. First, many of these studies have simply correlated the alternative performance measures without providing a rational basis (hypothesis) for even expecting any convergence. Second, many of the studies found a "low convergence" among the measures but failed to "match" these measures with regard to the specific dimensions of performance being represented. Third, many of the studies failed to provide information about the behavioral specificity of the performance appraisal items.

Quality of Prediction

In this section, selection research methodology is reviewed and critiqued, and the statistical techniques for prediction analyses are presented. Of importance here is the distinction between quality of fit and quality of prediction as it relates to research findings in selection. While the concepts presented here are not new, the statistical techniques developed to compare these qualities warrant discussion here.

Classic Validation

Dunnette (1963) characterizes the "classic validation model" of selection research as one in which simplistic, bivariate relationships are postulated that link predictors with a criterion. As a result, the correlation coefficient is used as almost the sole statistic of validation research (Dunnette, 1963). However, this reliance on simple correlations is insufficient in light of the multiple dimensions of performance (Ghiselli, 1956) and the likelihood of nonlinear selection models (Dunnette, 1963). In order to examine the nature of the hypothesized relationships, and the linkages between selection model variables, a more complex methodology is required (Dunnette, 1963; Guion, 1976). The transition from simple correlations to more "robust" analyses shifts the research focus from "test evaluation" to performance prediction (i.e., Guion, 1976).

Fit versus Prediction

The "quality of fit" refers to how well the data fit the postulated model. The underlying assumption is that the goal of such research is to explain, from the fitted model, which regressor (predictor) variables influence the criterion variable. In this sense, the model is used merely as an instrument to detect the degree of importance of each variable in explaining the variation in re-

sponse (Myers, 1986). Selection research has relied almost solely on the use of the correlation coefficient; the implicit goal of this research is thus "system explanation" via quality of fit procedures. Pearson's correlation coefficient assumes a linear relationship between the variables and thus describes the degree of linear association between them; i.e., the strength of the relationship. However, a model that satisfactorily describes the data will not necessarily yield the best prediction; quality fit and quality prediction do not necessarily coincide (Myers, 1986).

In contrast to the above, the "quality of prediction" refers to how well the postulated model performs, or predicts criterion responses. Further, an assessment of the prediction capability of a model encompasses the notion of "model validation"; i.e., predicting response values that are independent of the data that built the model (Myers, 1986). In this sense, the criterion variable is of utmost importance; the goal of this research is to predict the criterion adequately and assess the quality of future predictions. The implication of this methodology for selection research is the explicit focus on ***prediction*** and ***model validation.*** Many selection researchers have criticized the reliance on the correlation coefficient as "the validity coefficient," stressing that this practice assumes research aimed at "test evaluation" rather than performance prediction (e.g., Guion, 1976; Dunnette, 1963; Seashore et al. 1960). Therefore a more robust methodology would entail a direct examination of the prediction capabilities of selection models, placing less attention on correlation and more attention on "validation."

Prediction Variation

The term "prediction" implies estimating a response at future data locations, hence model validation (Myers, 1986). One method for validating the model is data splitting in which the data are partitioned into a "fitting sample" and a "validation sample." Another method of validation is the use of the Prediction Sum of Squares ***(PreSS statistic).*** One of the advantages of this method as opposed to data splitting is that the PreSS statistic does not require two samples; instead, validation is achieved by "setting aside" each of the sample observations one at a time, estimating the model

coefficients using the remaining (n - 1) observations, and thus producing *n* validations. Myers (1986) has devoted considerable attention to describing this validation statistic. His work is summarized below, and the reader is referred to his book for a more thorough explanation of this topic.

The PreSS statistic is defined as the sum of squared PreSS residuals, with a separate PreSS residual for each observation. These PreSS residuals are similar to ordinary residuals. An ordinary residual is the difference between the estimated (model) response and the observed response at each data point. The difference between ordinary and PreSS residuals is that ordinary residuals are not based on independent responses; the estimated (model) response is based on, and drawn toward, the observed response. In contrast, the PreSS residuals are "true prediction errors;" the observed response was not simultaneously used for both model fit and assessment (Myers, 1986, p. 106). The process for generating these PreSS residuals is as follows. Consider a data set with *n* observations (i.e., sample size = n). To begin, the first observation is "set aside" and the remaining (n - 1) observations are used to estimate the coefficients of the model. This observation is then replaced, the second observation is withheld, and the coefficients of the model are estimated again using the other (n - 1) observations. This process continues such that each observation is removed, one at a time, and therefore the model is estimated *n* times. In essence, the model has been internally validated *n* times. The PreSS residuals refer to the estimated response each time an observation is withheld. These residuals are true prediction errors because, for each observation in the model, the predicted value is independent of the observed value. This is a true test of validation (Myers, 1986, p. 106).

The PreSS residuals are used to compute the following statistics. It should be noted that these statistics are not "test statistics"; they are predictive statistics used to compare various models.

PreSS This statistic is the sum of the squared PreSS errors associated with the model. This is analogous to the Residual Sum of Squares used in regression. The distinction here centers on the composition of the above residuals. The ResSS (Residual Sum of Squares) uses "ordinary residuals" which, by the very nature of least-squares regression,

will be minimized, thus smaller than the true prediction errors. Because the estimated response is not independent of the observed response, these residuals estimate quality of fit, not quality of future prediction. As explained previously, the PreSS utilizes "prediction residuals."

$R^2(p)$ This statistic is an R^2-type statistic which reflects the overall prediction capability of the model. Computationally: R^2_p = 1 - (PreSS/Total SS). This is analogous to the Coefficient of Determination: R^2 = 1 - (ResSS/Total SS). These statistics are distinguished in terms of the goal of the analysis: R^2 estimates the "quality of fit" of the model whereas R^2_p estimates the "quality of prediction" of the model.

Present Analyses

The purpose of the present research is to examine the validity of selection decisions. As noted in Chapter 1, this validity assessment is composed of two goals. First, ***criterion validity*** is examined: what are the relationships among alternative measures of job performance? In this assessment, the degree of convergence among job performance criteria is estimated using correlation analyses. In addition, these criteria relationships are compared to one another using PreSS analysis. Whereas the correlation measures the strength of the association between the variables for each hypothesized model (criteria relationship), the PreSS analysis provides additional information about how these multiple criteria models compare to each other in terms of predictive capability. For instance, each of the hypothesized relationships among job performance criteria is characterized by a correlation coefficient that indicates the strength of that relationship. These criteria models (relationships) are then compared in terms of prediction errors; i.e., model validation using independent data.

Second, ***predictive validity*** is examined using a criterion-related validation design: what is the nature of the relationship between predictors and criterion variables? In this assessment, the quality of fit is examined for each predictor-criterion model using the correlation (r) and Coefficient of

Determination (R^2). In addition, the quality of prediction for each predictor-criterion model is compared against the other hypothesized models in an effort to determine which model yields the best prediction capability with regard to prediction errors. For example, a postulated model using ability tests to predict overall job performance is compared to a model using ability tests to predict specific output performance. This comparison is based on measures of fit (i.e., R^2) and also on measures of prediction (i.e., R^2_p). This dual comparison (fit and prediction) highlights the prediction capability of each selection model and provides a basis for determining the "best" selection model in terms of both fit and prediction.

Conceptual Model

Given that the aim of selection research is the prediction of job performance, the need becomes apparent for a thorough conceptualization of the performance criteria. In the preceding literature review, the issues of criteria validity and equivalence were raised as they relate to a selection research paradigm. The model for job performance criterion validation (and subsequent performance prediction) proposed herein is based on the criterion model recommended by James (1973). The thrust of this model is the explicit inclusion of multiple levels, referents, and sources of performance criteria, with a primary focus on the dimensional criterion level. The outcome of such research is a step toward the validation of performance criteria as well as the validation of predictors.

According to this model, "performance" is seen in terms of a hierarchy of criterion levels. The relevant criterion elements of the model for this study are described below.

Global Performance

James (1973) refers to global criteria as organizational outcomes which represent global indices of job effectiveness. Similarly, Dunnette (1963) used the term "consequences" to refer to global

measures of occupational effectiveness--an all-encompassing measure of occupational success. This level of criterion performance is similar to a notion of an ultimate criterion which reflects overall success. The sole reliance on such a global level of performance measurement has been criticized because (a) it assumes an underlying general factor for all facets of job performance, (b) it hampers an understanding of the underlying job behaviors and performance, and (c) these measures are contaminated to an undetermined extent (James, 1973). Examples of global performance criteria are results-oriented measures such as promotion rate, productivity indices, and overall ratings of effectiveness.

Dimensional Performance

James defined dimensional criteria in terms of job behaviors that are directly related to organizational goals. As noted by James, this level of performance criteria recognizes the identification of organizational goals and the explicit relationship between these goals and individual job behaviors. In essence, the dimensional performance criteria represent a sort of "link" between individual activities and organizational goals; thus, these criteria should be the primary focus of criterion development and validation (James, 1973; Campbell et al., 1970). Examples of job performance criteria are behavioral results measures such as absenteeism, sales performance, production output, and dimensional ratings of job performance.

Dimensional performance measures can be defined along two parameters: method of measurement and type of referent. The ***method*** (source) of assessing this level of performance might be judgmental (subjective ratings) and/or nonjudgmental (objective counts). The judgmental measures are multiple rating dimensions that are ideally independent criterion dimensions which tap into the many facets of "job success." These dimensions are similar to what Kane (1980) refers to as the functional components of the job; i.e., the broad components that, collectively, define the full scope of the job. Similarly, the nonjudgmental measures are quantifiable/observable measures that also represent independent criterion dimensions. For example, a relevant component of job perform-

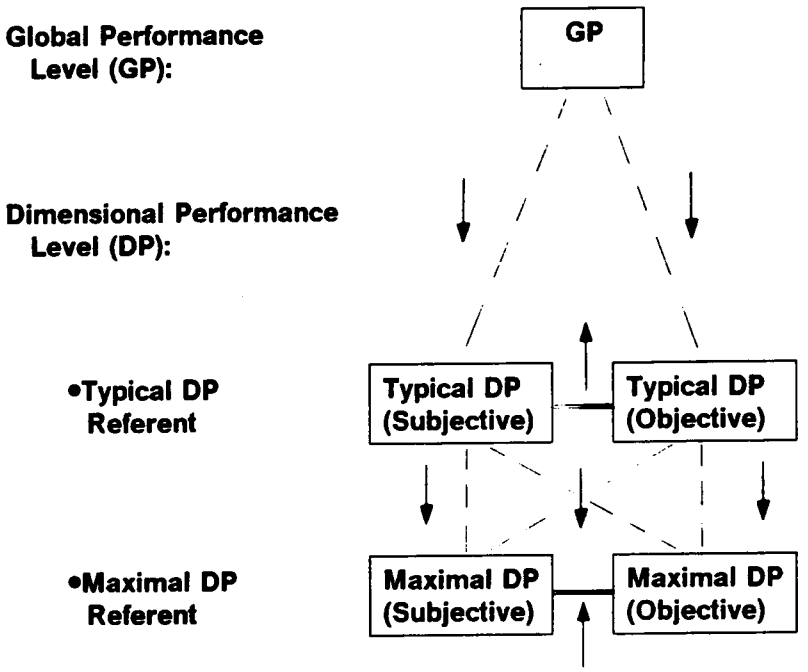
ance might be "dependability." This dimension of performance could be measured subjectively (i.e., a supervisor's rating of dependability/absence behavior) and/or objectively (i.e., personnel records of absence frequency, or absence "episodes").

The second parameter of job performance measurement refers to the ***referent*** point of performance assessment. Dimensional performance might be assessed in terms of "typical" performance and/or "maximal" performance. The "typical" performance referent specifies that performance be measured in terms of an average level of job performance over some period of time. This referent implies that all performance variation around the average is random; thus, the average level is the "true" estimate of performance achieved. For instance, an assessment of typical employee dependability might be gauged as the average level of attendance (objectively) and/or the typical attendance behavior over some period of time (subjectively).

In contrast, the "maximal" performance referent specifies that job performance be measured in terms of feasible performance (highest level of performance achieved). This concept of maximal performance focuses on the variation (distribution) of performance. The maximal (peak) level is an estimate of performance that is achievable in light of the organizational/external factors present on the job on a day-to-day basis. Following with the above example, an assessment of maximal dependability might be measured subjectively using a performance rating asking "what level of attendance is this employee capable of attaining?" On the other hand, maximal (attainable) attendance levels exhibited over a period of time might be measured objectively using attendance reports and measuring the highest level of attendance during a specific period of time.

Summary: The purpose of this study is to extend previous validation research by examining the relationships among multiple performance criteria as well as the relationships between criteria and predictors. The model used for this study focuses on multiple performance criteria and the interrelationships among these alternative criterion measures. This Job Performance Model is shown in Figure 1.

FIGURE 1
JOB PERFORMANCE MODEL



NOTES: "↓" denotes a positive yet "weak" relationship (i.e., low (convergence)

"↑" denotes a positive and "strong" relationship (i.e., high convergence)

The criterion variables of interest for this research study refer to two levels, two referents, and two methods of performance measurement. Specifically, these constructs are:

1. Global performance
2. Dimensional performance
   - two referents: typical and maximal
   - two methods: judgmental and nonjudgmental

Conceptual Hypotheses

The following research hypotheses are drawn from the Job Performance Model depicted in Figure 1. These hypotheses focus on the different levels and referents of performance criteria, utilizing different methods of performance measurement. In addition, the relative predictability of these multiple performance criteria is hypothesized in terms of their relationship with individual ability constructs. For each hypothesis, a summary notation is introduced in order to abbreviate the predicted relationships and the relative strengths of these relationships.

Relationships Among Performance Levels

The criterion model (Figure 1) postulates that:

1. There will be a low convergence among performance criteria that are measured on different levels (i.e., global performance versus dimensional performance), and
2. There will be a high convergence among performance criteria that are measured on the same dimensional performance level, regardless of the method of measurement.

With regard to the first postulate, the model depicts two levels of performance measurement. The relationship ***between*** these two different levels of performance criteria is positive due to the shared "criterion space" of the performance dimension; i.e., the global performance is a summary criterion measure that incorporates the dimensional performance element. Yet this relationship is

weak due to the lack of congruence in the underlying performance constructs; the global performance is an overall performance measure whereas the Dimensional Performance is a specific performance result. Conceptually, this low convergence between performance levels is supported by arguments from Smith (1976) regarding different levels of specificity, from James (1973) regarding different levels of results, and from Weitz (1961) regarding different "types" of parameter measurement. Empirically, there is also support for this low convergence (i.e., Seashore et al., 1960; Severin, 1952; Alexander and Wilkins, 1982).

Although the relationship between the dimensional and global levels of performance criteria is hypothesized as a low convergence, there is evidence that common methods of criterion measurement produce shared method variance. In fact, this common method variance (shared errors) is one of the basic criticisms of the Multitrait-Multirater approach to criterion construct validity (James, 1973). Because most global performance criteria are judgmental (subjective) ratings, it is further hypothesized that the relationship between judgmental ratings of both global and dimensional performance will be stronger than the relationship between different methods of measuring these two levels. Thus:

H1(a) There will be a positive relationship between a objective measure of Dimensional Performance and a subjective measure of Global Performance (DPO-GP).

H1(b) There will be a positive relationship between a subjective measure of Dimensional Performance and a subjective measure of Global Performance (DPS-GP).

H1(c) There will be a stronger relationship between subjective measures of Dimensional Performance and Global Performance than between an objective measure of Dimensional Performance and the Global Performance rating (DPS-GP > DPO-GP).

The second postulate states that the relationship of criteria measured ***within*** a specific Dimensional Performance level is expected to be strong. The relationship is positive due to the congruent dimensional factor space and is strong due to the matched specificity of the underlying constructs (i.e., Smith, 1976; James, 1973). Furthermore, this high convergence will be evident regardless of the method of measurement utilized. By matching criteria in terms of the specific aspect of performance of interest, a high convergence between different methods of measuring the same dimen-

sion aids in our understanding of the construct under study. Whereas previous empirical studies have not found a strong relationship between objective and subjective measures of performance, it was argued that these studies failed to "match" these different measures with respect to a specific, common construct of performance.

Based on the postulates above, the following research hypotheses are presented.

H2(a) There will be a positive relationship between a subjective measure of Dimensional Performance and an objective measure of Dimensional Performance (DPS-DPO).

H2(b) There will be a stronger relationship between a subjective and objective measure of Dimensional Performance than between the subjective measures of Dimensional Performance and Global Performance (DPS-DPO > DPS-GP).

These findings will have an impact on our understanding of the relationships among performance criteria in two ways. First, the equivalence (hence substitutability) of alternative performance criteria can be examined. A recognition, and understanding, of the different levels of the performance parameter addresses the basic question of whether "performance is performance is performance." The inclusion of different methods of Dimensional Performance measurement relates to the construct validity of the underlying performance criterion; a strong convergence within the Dimensional Performance level adds to our understanding of the construct and the accuracy of performance ratings. Second, the linkages between the different levels of the performance parameter addresses the issue of "what has been measured." According to the model in Figure 1, the different levels refer to different levels of specificity, and different dimensions of goal attainment. Consider that a Dimensional Performance measure is an indicator of a specific, unidimensional performance result. In contrast, a Global Performance measure is indicative of both performance behaviors and results, thus multidimensional in nature. While the issue here is not a matter of a single versus composite criterion, or "the best" criterion, it is important to know whether different criterion levels represent different performance constructs. The outcome of this examination of performance levels is a means of establishing a range of criteria, or "laws of criteria" (Weitz, 1961), in order to validate the subsequent selection models.

Relationships Among Performance Referents

The criterion model further postulates that:

1. There will be a low convergence among dimensional performance criteria that are measured in terms of different performance referents (i.e., typical versus maximal), and
2. There will be a high convergence among dimensional performance criteria that are measured in terms of the same performance referent, regardless of the method of measurement.

A performance referent relates to the frame of reference for performance evaluation. Recall that a typical level of performance assumes that all performance variation around the average is random (Kane, 1982). Yet this referent does not account for "feasible" performance levels; hence it is affected by constraints placed on an individual, constraints which might prevent a maximal level of performance achievement (Kane, 1982; Bernardin and Beatty, 1984). In contrast, a maximal (feasible) level of performance accounts for "opportunity bias" and refers to the highest performance level attained by the individual at least once during a period of time. These different performance referents also relate to different expectations of performance. Landy and Farr (1983) identified the need for an optimization parameter as a frame of reference for performance evaluation which defines the "best" level of performance as opposed to an ideal level of performance (which may be unattainable). These authors, and others (Kane, 1982; Bernardin and Beatty, 1984; Borman et al., 1980) have stressed that the frame of reference chosen for performance evaluation is a criterion parameter in itself and that different referents of performance conceptualization are not necessarily equivalent. By establishing different expectations for performance measurement, the underlying selection goals (hence performance constructs) are also different. This referent parameter of criterion performance represents a potential boundary condition for subsequent selection models, and as a result might limit the usefulness of a particular predictor. Thus the first postulate states that the relationship ***between different*** referents of dimensional performance, measured on the same Dimensional Performance level, will be positive due only to the shared factor space; i.e., the same dimension of performance. Yet this is a low convergence due to the different expectations for measured performance (i.e., different performance constraints). The hypotheses state that:

H3(a) There will be a positive relationship between a subjective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance (DPS-MaxS).

H3(b) There will be a positive relationship between an objective measure of "typical" Dimensional Performance an objective measure of "maximal" Dimensional Performance (DPO-MaxO).

H3(c) There will be a positive relationship between a subjective measure of "typical" Dimensional Performance and an objective measure of "maximal" Dimensional Performance (DPS-MaxO).

H3(d) There will be a positive relationship between a subjective measure of "maximal" Dimensional Performance and an objective measure of "typical" Dimensional Performance (DPO-MaxS).

The second postulate states that the relationship ***between matched*** referents of dimensional performance will be a strong, positive relationship due to congruent factor space and also congruent performance expectations, hence more equivalent performance constructs. In addition, because the Dimensional Performance criteria are matched according to the performance referent, this relationship will outweigh the relationship between Dimensional Performance criteria with dissimilar referents (i.e., typical with maximal) due to a common frame of reference. Furthermore, the relationship between referent-matched Dimensional Performance criteria will outweigh the relationship between Dimensional Performance criteria measured by common methods (objective and subjective) due to a higher degree of specificity.

The specific research hypotheses which can be drawn from this model are as follows. For ***matched*** referents:

H4(a) There will be a positive relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance (DPS-DPO).

H4(b) There will be a positive relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance (MaxS-MaxO).

Regarding the relative convergence of matched referents versus different referents with different measurement methods:

H5(a) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Perform-

ance than between an objective measure of "maximal" Dimensional Performance and a subjective measure of "typical" Dimensional Performance (MaxS-MaxO > DPS-MaxO).

H5(b) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance (MaxS-MaxO > DPO-MaxS).

H5(c) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between an objective measure of "maximal" Dimensional Performance and a subjective measure of "typical" Dimensional Performance (DPO-DPS > DPS-MaxO).

H5(d) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance (DPO-DPS > DPO-MaxS).

Regarding the relative convergence of matched referents versus different referents with common measurement methods:

H5(e) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and an objective measure of "maximal" Dimensional Performance (MaxS-MaxO > DPO-MaxO).

H5(f) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between a subjective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance (MaxS-MaxO > DPS-MaxS).

H5(g) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and an objective measure of "maximal" Dimensional Performance (DPS-DPO > DPO-MaxO).

H5(h) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between a subjective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance (DPS-DPO > DPS-MaxS).

The conceptualization of the referent parameter suggests the following implications. First, the equivalence among different dimensional performance referents can be empirically examined. On a conceptual basis, this distinction between performance referents might be obvious; different expectations for performance relate to different contexts, thus constructs, of performance. Yet on a practical level, these different performance referents refer to different criteria for selection, hence

different selection models. Specifically, if a particular predictor is validated against a typical performance criterion, this does not necessarily yield information as to what performance levels "could be" achieved. Recall that a maximal frame of reference for performance evaluation explicitly accounts for the extraneous constraints placed on performance, whereas a typical performance evaluation does not. Second, the linkages between the different referents addresses the question of "what to measure." The different performance referents relate directly to different selection goals, and these goals represent the "conceptual criteria" of importance for a particular situation.

Relationships Between Ability and Performance

Although the Job Performance Model presented in Figure 1 is a criterion model for performance measurement, it has implications for subsequent predictor-criterion validation models. The foregoing hypotheses state that the different levels and different referents of job performance conceptualization represent different constructs of performance. These different performance constructs imply different selection models (i.e., Guion, 1976; Dunnette, 1963); i.e., one predictor does not necessarily predict different levels of job performance criteria with equal "power." As discussed in the literature review, selection research has been criticized due to the lack of conceptual specification of the criterion parameter (James, 1973; Prien, 1966). Numerous validation studies for a particular predictor have produced inconsistent results (Guion, 1976), yet it is difficult to determine whether the inconsistent findings are due to statistical artifacts (i.e., Schmidt and Hunter) or to different underlying performance constructs. Thus, the relative predictability of these performance constructs can be examined using the performance model in Figure 1.

The "basic" performance model used throughout organizational literature reads: Performance = f(ability x motivation). Although in this sense the term "motivation" is merely a summary label that identifies a *class* of independent variable/dependent variable relationships (Campbell and Pritchard, 1976), the model is used here to illustrate the importance of ability as a component of job performance. The foregoing arguments distinguish "performance" in terms of different levels

and different referents; however, ability is hypothesized to be related to each of these performance criteria. Thus:

H6(a) There will be a positive relationship between individual abilities and "maximal" Dimensional Performance (Abilities-MaxO/MaxS).

H6(b) There will be a positive relationship between individual abilities and "typical" Dimensional Performance (Abilities-DPO/DPS).

H6(c) There will be a positive relationship between individual abilities and Global Performance (Abilities-GP).

Of interest here is the relative importance of ability constructs with respect to the different criterion constructs. Thus while abilities are related to Maximal Dimensional Performance, Typical Dimensional Performance, and Global Performance, they are assumed to be differentially predictive of these alternative criteria. First, it is hypothesized that abilities are more strongly predictive of Maximal Dimensional Performance than of Typical Dimensional Performance, regardless of the method of measurement employed. Consider that in the case of Maximal performance, performance is conceptually defined as a feasible (peak) referent, measured on a Dimensional Performance level. It has been argued that this construct accounts for organizational (external) constraints that affect individual performance. Thus the functional relationship specifies:

Maximal Dimensional Performance = f(ability, motivation)

Now consider that a Typical performance referent, measured on a Dimensional Performance level, is a result of not only an individual's ability and motivation but also the organizational constraints on the job that prevent an individual from consistently performing at "peak" performance levels. Thus the typical frame of reference does not explicitly control for "opportunity bias;" i.e., external factors that constrain day-to-day performance. The functional relationship can be stated such that:

Typical Dimensional Performance = f(ability, motivation, opportunity bias)

Thus Maximal Dimensional Performance is more closely related to the individual's abilities, and Typical Dimensional Performance is a broader construct that incorporates the effects of external constraints on individual performance. As a result, abilities are predicted to be more congruent

with the Maximal Dimensional Performance referent than the Typical referent. This hypothesis is stated here as follows:

H7(a) There will be a stronger relationship between individual abilities and "maximal" Dimensional Performance than between individual abilities and "typical" Dimensional Performance (Abilities-MaxO/MaxS > Abilities-DPO/DPS).

Second, it is hypothesized that abilities are more predictive of Maximal Dimensional Performance than of Global Performance, regardless of the method by which Dimensional Performance is measured. This hypothesis is based on the functional relationships inherent in these two criterion parameters. The Global Performance measure is a summary performance measure that includes multiple dimensions of performance (i.e., performance behaviors as well as results). It is possible that, given this broad, overall performance construct, abilities are of lesser importance to overall job success. In addition, the frame of reference (context) of a Global Performance assessment is usually undefined and contains factors that may be unrelated to an individual's ability. In functional form:

Global Performance = f(ability, motivation, opportunity bias + "Other")

Whereas:

Maximal Dimensional Performance = f(ability, motivation)

Thus:

H7(b) There will be a stronger relationship between individual abilities and "maximal" Dimensional Performance than between individual abilities and Global Performance (Abilities-MaxO/MaxS > Abilities-GP).

Third, it is hypothesized that abilities are more strongly predictive of Typical Dimensional Performance than of Global Performance, regardless of the method by which Typical Dimensional Performance is measured. Although both measures place the performance construct within a "typical" frame of reference (i.e., no explicit control for external constraints), the Typical Dimensional Performance is confined to only one aspect of performance results. As noted above, the Global Performance measure is multidimensional and refers to a summary evaluation of overall job success. Thus, while Typical Dimensional Performance refers to only one dimension of perform-

ance results, Global Performance refers to a variety of performance behaviors and results, some of which may not be controllable by an individual. Thus:

H7(c) There will be a stronger relationship between individual abilities and Typical Dimensional Performance than between individual abilities and Global Performance (Abilities-DPO/DPS > Abilities-GP).

Summary

The review of prior validation research reveals that:

- Contruct validity research has focused on the meaning of ratings yet has been limited to one method of job performance measurement, and
- Performance appraisal validity research has focused on the convergence between ratings and "objective" criteria yet has compared these multiple measures across different levels of job performance measurement (specificity).

Heneman (1986) conducted a meta-analysis of criterion studies with the intent of determining the degree of convergence between supervisory ratings and results-oriented measures of job performance. His three major conclusions included:

1. There is a weak relationship between multiple methods of job performance measurement;
2. The rating method (relative v. absolute comparisons) has a moderating effect on the degree of convergence; and
3. The alternative measures of job performance are not equivalent.

However, this meta-analysis does not account for the different levels of criterion specificity. For example, some of the objective results measures are global indices of job success whereas some of these measures are specific indices of dimensional performance. Furthermore, the type of rating was not specified; i.e., behavioral ratings versus trait ratings. Because this study did not examine the impact of the different levels for criterion measurement, the "weak relationships" between different methods of performance measurement may in fact be a result of different levels of specificity.

The research model suggested by James (1973) and developed in this study explicitly examines the degree of convergence among multiple methods of job performance measurement. Furthermore, this model posits potential boundary conditions on criteria convergence in the form of levels and referents for job performance measurement. Whereas Heneman (1986) and others have concluded that there is a weak relationship between different methods of criterion measurement, this model accounts for the conceptual congruence between the levels of specificity with which performance is measured and the expectations by which performance is evaluated. The outcome of this research is the establishment of boundary conditions for job performance measurement and the introduction of different expectations for job performance evaluation.

Chapter 3: Methodology

Overview

In the preceding chapter, the conceptual model for criterion validation was presented as a framework for this study. This model postulated a series of hypotheses about the interrelationships among alternative job performance criteria and about the pursuant predictor-criterion relationships using each of the criteria. The basic design for this research is a criterion-related validation study using a predictive strategy (longitudinal). Within this approach, the construct validity of criterion measures is examined in terms of multiple levels, referents, and methods of performance measurement. The predictor variables (ability tests) were measured at time of employment application and the criterion variables (job performance) were collected throughout the study. Specifically, the subjective performance criteria were measured after approximately six months on the job, and the objective performance criteria were measured on a weekly basis throughout the period of this study.

This chapter describes the methodology employed in conducting the study. First, a description of the research setting is presented. Second, the variable measurement procedures are discussed. Third, the conceptual hypotheses are restated in operational terms together with the data analysis procedures employed.

Research Sample

The research site for this study is a garment manufacturing plant that produces a variety of sportswear styles. The parent company operates ten sewing plants located in Virginia and employs approximately 8,000 people. The Human Resource function is a centralized department for all

plants, maintaining individual employee records regarding selection, performance, and termination data. The plant chosen for this employs approximately 640 sewing machine operators. This sewing machine operator job was selected for this study due to the large number of employees and the availability of both judgmental and nonjudgmental performance data. These operators are paid on a piecework basis with a guaranteed minimum wage, and production reports on these employees are developed on a weekly basis.

The sewing machine operator jobs in the plant were organized into production lines, with each responsible for the completion of a specific style of garment. For example, a sweatshirt line consisted of 6 separate sewing operations (i.e., seam sleeves, sew collar) that are required to produce the finished garment. This meant that a single operator would sit at a sewing machine stitching the same seam (i.e., operation) throughout the course of the day. Although the operations were designed as production lines, the sewing maching operators did not work interdependently per se. The "line" collectively produced a garment, yet each operator independently stitched bundles of garment pieces, and their raw materials were not dependent on the work pace of the rest of the "line."

The longitudinal nature of this study meant that the final sample size would be affected by turnover, transfers, and the availability of the multiple performance measures. For purposes of analysis, this study used two samples. The first sample was used for the performance criteria analyses (H1a - H5h) and consisted of those employees with multiple performance data (i.e., both judgmental and nonjudgmental measures). Data collection procedures were started in January, 1986, and during the period under study, 338 people were hired as sewing machine operators. Due to turnover, transfers, and missing performance data, the sample size was reduced to 153. The demographic characteristics of this sample reveal that all of the operators are female, of which 60% are white and 40% are black. The average age of this group is 27, ranging from 18 to 49 years old. With respect to prior experience, 65% of the operators had no previous sewing machine operator experience, 16% had 1 to 12 months experience, and 19% had more than one year experience.

The average experience level is 10 months. All of these employees were hired between January and October, 1986, and are divided among 13 supervisors.

The second sample was further reduced in order to examine the ability-performance relationships (H6a - H7c). This sample reduction was due to the fact that not all of the employees in the first sample had predictor (ability test) measures. Thus, the subsample used for the ability-performance analyses consists of 117 sewing machine operators. The demographic characteristics of the two samples are similar: 63% are white, the average age is 27, and the average experience level is 8 months. Out of the 117 operators, 67% had no previous experience, 15% had 1 to 12 months experience, and 18% had more than one year.

Variable Measures

Criterion Measures

Five measures of job performance are relevant to this study. These variables reflect the relationships between different levels, different referents, and different methods of performance measurement. In the following paragraphs, the different methods of performance measurement are discussed first; each of the criterion variables is then presented in operational terms.

Measurement Methods: In this study, the ***nonjudgmental*** (objective) measures of job performance were collected on the dimensional level but not on the global performance level. Furthermore, only one dimension, Production Output, was examined in this study. The reason for this restriction was the lack of nonjudgmental performance data on the other performance dimensions. Management identified production output as a critical dimension of job success for the sewing machine operators. As a result, the primary focus of this study was on one dimension, Production Output.

The data for nonjudgmental production measures were obtained from the plant's weekly production earnings report. For each employee, these reports showed the average hourly output for each week, multiplied by the unit pay that was determined for each operation by the Method Time Measurement procedure. Unit pay was established for each operation on the basis of the time required to complete one cycle of work. Thus, within the limits of error of the industrial engineering studies employed, two employees working at the same output rate, but on different operations, would receive the same pay. By using average production earnings, jobs were thus equated along a common pay scale, and differences in earnings reflected differences in the rate with which operations were performed. It should be noted that if an employee's average production earnings fell below the guaranteed minimum wage, the entry in the weekly production earnings report showed the wage equivalent of the number of units actually produced even though the employee received the guaranteed wage. Thus, the average weekly "earned wages" constitute the data for this nonjudgmental measure of Dimensional Performance. These data were collected on a weekly basis throughout the period of the study.

The ***judgmental*** (subjective) measures of performance were used to measure both the Global Performance and Dimensional Performance levels. A performance appraisal instrument was designed in cooperation with management. This instrument (see Appendix A) was based on a job analysis of the position of sewing machine operator and consisted of performance ratings on five dimensions of job performance. Five performance dimensions were thus identified as necessary components of successful job performance: Quantity of work (production output), quality of work, flexibility, receptiveness to training and instruction, and dependability. In addition, the Quantity of Work dimension was rated a second time using the Performance Distribution method (Kane, 1980). This second rating focused on the distribution of performance outcomes and was used to determine "maximal" (peak) output levels for the period. Each operator was rated on each of these dimensions by her immediate supervisor.

The data on judgmental performance measures were obtained from this performance appraisal. The five dimensional ratings were measured by graphic-type scale scores for each separate per-

formance dimension. These rating scales were constructed by defining each of the dimensions and providing written objective anchors for each response scale. An additional dimensional rating was obtained on the production output dimension to reflect the "maximum" quantity of work level achieved during the period. The format for this item was based on Kane's Performance Distribution Method (1980). These data were collected twice in order to rate each individual after approximately six months on the job.

Operational Variables: The Job Performance Model presented in Chapter 2 depicts two levels for performance evaluation. The Global Performance level refers to an overall job effectiveness construct, and the Dimensional Performance level refers to a more specific performance results construct. Further, two referents for performance evaluation were identified for the dimensional performance level. The typical and maximal performance constructs refer to different performance referents within the Dimensional Performance level. Thus five performance criteria are used in this study and are described below. These measures are:

1. Global Performance: judgmental rating
2. Typical Production Output (dimensional level): judgmental and nonjudgmental measures
3. Maximal Production Output (dimensional level): judgmental and nonjudgmental measures

Global Performance was measured subjectively using the performance appraisal data. A single-item response to an overall performance rating scale was used for the global performance rating. This item asked the supervisor to consider all of the identified factors (dimensions) involved in the sewing machine operator job for successful performance (see Appendix A).

Typical Production Output was measured both subjectively and objectively on the dimensional performance level. The subjective measure of Production Output was a single-item rating on the performance appraisal dimension "Quantity of Work," using a typical performance level as the frame of reference for evaluation. The objective measure of the typical production output was based on the mode of the hourly "earned" wages over a 20-week period (5 months), which omits the first 3-4 weeks on the job. This modal production score reflects the actual weekly production

of the employees; thus it does not include time-not-worked (vacations, absenteeism), rework (quality rejects), or guaranteed wages (supplemental pay). The modal level was chosen, rather than the average, based on the concept of typical--the most frequently attained performance level. The use of an arithmetic average is heavily influenced by extreme performance levels and thus does not provide a meaningful measurement for "typical." In the case of a bi-modal (or multi-modal) performance distribution, the average of these modes was taken as the typical production level.

Maximal Production Output was also measured both subjectively and objectively. The subjective measure was a single-item rating on the appraisal dimension "Quantity of Work" which utilized the "maximal" frame of reference for evaluation. This rating was obtained from the Performance Distribution method. The objective measure was based on the "peak" level of average hourly earned wages during the 20-week period. This peak production level reflects the highest weekly production level achieved by each employee.

Predictor Measures

Ability constructs were chosen as predictors for two reasons. First, the jobs included in this study provide for no formal training; all "learning" takes place on the job in the natural work setting. As a result, cognitive ability was an important component of "learnability" (Ghiselli, 1973; Dreher and Sackett, 1983). Second, the jobs are composed of complex psychomotor tasks which range from low to high difficulty levels. As a result, psychomotor ability was identified as a significant factor for successful performance on the job. Both cognitive ability and psychomotor ability were measured using the General Aptitude Test Battery (GATB), which was developed by the U.S. Department of Labor and the U.S. Training and Employment Services. The GATB consists of 12 tests, which are grouped into 9 aptitude scales. Six of these scales are relevant to this study; these scale scores were designed to measure cognitive ability and psychomotor ability. Cognitive ability is composed of three scales: General intelligence, Verbal aptitude, and Numerical aptitude. Psychomotor ability is composed of Coordination, Finger dexterity, and Manual dexterity scales.

These ability scores are based on research conducted by John Hunter in conjunction with the U.S. Employment Service (Hunter, 1982).

The GATB was administered by the local Employment Service office under controlled conditions. Thus, the ability test scores were obtained from that office for the sample in this study.

Measurement Reliability

Reliability of a measurement procedure refers to its freedom from unsystematic (random) errors of measurement, or its consistency under different conditions that might introduce error into the scores (Cascio, 1982). Traditionally, reliability is estimated via one of three methods: test-retest, alternative forms, or internal consistency. A fourth method, "scorer" reliability, estimates the degree of scoring consistency among different scorers (raters). These estimates involve the notion of ***repeated*** performance measurements and assume that a "true score" can be estimated from the observed score. According to this view, the observed score is composed of "true" and random error components. Thus reliability is estimated from a coefficient of stability (test-retest), a coefficient of equivalence (alternate forms), a Kuder-Richardson coefficient (internal consistency), or interrater agreement (scorer consistency).

Criterion Reliability: For this study, the above-mentioned estimates of reliabilty are not appropriate, or feasible, for the constructs of performance. Consider that Test-Retest reliability assumes the consistency of performance over time, an assumption that is problematic and, possibly, inappropriate (i.e., Ghiselli, 1956; Smith, 1976; Prien, 1966). In fact, the very issue of the consistency, or dynamics, of performance is an area of research in itself. Furthermore, an Internal Consistency method of estimating performance reliability requires that multiple items are measured for each hypothesized dimension of performance, and that these dimensions are to be collapsed into some kind of overall composite measure of job performance. This method is more appropriate for subjective performance measurement than for objective performance measurement. However, the use

of a composite measure of job performance is the basis for an on-going controversy in the criterion development (and selection research) field. Specifically, if the multiple dimensions of performance are conceptually independent factors for overall job success, what is the ***meaning*** of a composite measure that combines supposedly independent job performance dimensions? And what weighting procedure should be employed in order to obtain the linear composite? In this study, five dimensions of job performance were identified as necessary components for sewing machine operator job success. The intercorrelations of these five dimensions range from -.03 to +.54, suggesting that these dimensions do in fact relate to different factors of overall job performance. A Cronbach's alpha was in fact computed on the five dimensions, yielding a value of 0.64 for internal consistency. This relative low value, combined with the observed range of intercorrelations, provides some evidence that the five dimensions are not directly comparable and that a linear composite measure of overall job performance would obscure the interactions, and independence, of the multiple dimensions of job performance.

The use of an Interrater reliability estimate is perhaps one of the most commonly used reliability estimates when performance appraisal data is used for performance measurement. Yet the use of this method is based on two features of the rating situation: multiple raters and no objective performance data. First, interrater reliability requires that multiple raters for each employee are not only available but also have the same opportunity to observe each employee's performance behavior and the same perception (frame of reference) for what dimensions are important for successful job performance (i.e., Borman, 1974). In this study, as in most industrial settings, there was only one immediate supervisor for each employee. Based on this limiting factor, the use of interrater agreement reliability was not feasible. A second study feature that warrants the use of interrater reliability is the absence of an objective measure of job performance. In most cases, interrater agreement is used for either a reliability estimate and, sometimes, a validity estimate due to the fact that objective, "hard" criteria for job performance are not available (i.e., Bernardin and Beatty, 1984). In this study, objective measures of job performance were available, and the degree of convergence between these dual performance measures is in fact a major focus of this study.

Bernardin and Beatty (1984) suggest that the convergence between these multiple methods provides evidence not only about the validity of performance ratings but also about the reliability of the subjective ratings.

The last method of reliability estimation, Alternate Forms, focuses on the consistency of performance measurement over different measurement procedures--hence a coefficient of equivalence. In this study, a kind of alternate forms reliability is examined, whereby the alternate forms are in fact alternate ***methods*** of job performance measurement. Note that the "coefficient of equivalence" investigated herein involves the degree of convergence--thus equivalence--between alternative measures of job performance.

In the following paragraphs, each of the performance measures are discussed in terms of measurement reliability. While no reliability estimates are computed, reliability is discussed in terms of the nature of the measurement procedure.

The **Global Performance** measure used in this study was a single-item rating of overall job performance. The performance evaluation procedure asked that the raters first evaluate their employees on five separate dimensions of job performance and, second, evaluate each employees' overall performance based on the stated dimensions of job performance. The single-item global measure was used instead of a linear composite due to the low intercorrelations among the dimensions and the relatively low value of internal consistency. However, the data collected from the appraisal session did allow for a policy-capturing examination of supervisory ratings; specifically, which of the identified dimensions of job performance were important determinants of the overall rating? Regression analysis was used, regressing the overall rating item against the five dimensions for job performance. First, the "full model" was examined, this model postulating that the global rating is a function of all five job performance dimensions. Second, all possible models were examined to determine whether a subset of dimensions could "explain" the overall rating item as well as the use of all five dimensions. The results of this analysis revealed the following:

1. The "full model" yielded an R-Square of 53%, thus indicating that the five dimensions collectively explain approximately 53% of the variation in the overall rating item.

2. A reduced model, using production quantity, production quality, and dependability, yielded an R-Square of 50%, thus indicating that these three dimensions alone explain half of the variation in the overall rating item. Each of the regression coefficients was significant (at $p < .01$), and the relative importance of each dimension, based on the slope estimates, indicates that quantity of production produced the strongest effect ($b = 0.42$), followed by quality of production ($b = 0.32$) and then dependability ($b = 0.11$). Upon internal cross-validation (i.e., PreSS analysis), the degree of error in the model increased slightly such that the proportion of variation in the overall rating that is predictable by these three dimensions is 47%.

3. A comparison of the full and reduced models suggested that, on the average, a supervisor's evaluation of overall job performance is strongly influenced by only three of the five identified dimensions. Although for each model the quantity of production produced the strongest influence, this dimension alone explains only 39% of the variation in the global rating. As a result, the reduced model indicates that three dimensions are important determinants of the overall rating and that a linear composite of all five dimensions might not capture the implicit importance of the independent items, importance as perceived by the immediate raters.

It should be noted that, at best, only 53% of the variation in global performance ratings could be explained by the five identified job dimensions. This finding supports some of the basic criticisms regarding the reliance on subjective ratings for selection research: performance ratings are heavily influenced by "outside factors" (i.e., ratee and rater characteristics, context of evaluation, etc.) and the degree of rating contamination is largely unknown. Although the global rating item examined here, and used throughout this study, is not determined solely by the explicit rating dimensions, it is doubtful that it is contaminated any more or less than ratings used in previous research. Although the policy capturing analysis indicated that approximately 50% of the global item variation could be accounted for from a knowledge of three dimensional ratings, it is the purpose of this study to determine the nature of the convergence between global performance and more specific, and objective, measures of job performance.

The **Dimensional Performance** measures used in this study were measured by both subjective and objective methods. The subjective dimensional performance measures are single-item ratings on one dimension of job performance--Quantity of Production--using first a typical frame of reference and, second, a maximal (feasible) frame of reference. Although a concrete estimate of reliability is not available, the use of detailed definitions and objective performance standards for rating anchors provides an objective basis for performance evaluation. For this study, graphic-type rating

scales were used (see Appendix A), and the selection (choice) of job dimensions was in fact based on an analysis of the job under study.

The objective dimensional performance measures were based on records of production output. The data used for these measures are actually the average hourly production earnings, and the degree of reliability, or unreliabilty, is thus a function of the error in the industrial engineering studies. The company under study employs an Industrial Engineering staff, and piece rates are continually under examination. No piece rate modifications were made during the time of this study, and no major changes in piece rates were under advisement. Thus the reliability of the objective measures was assumed to be high, and any errors in measurement were a direct function of the underlying engineering standards.

Predictor Reliability: The performance predictors used in this study are the aptitude test scores of newly-hired sewing machine operators obtained from the General Aptitude Test Battery (GATB). The GATB was developed and administered by the U.S. Employment Service and has been the extensive validation research (i.e., Hunter, 1983b; Hunter, 1983c). Because the aptitude scales studied herein are drawn directly from this professionally developed test, the reliability of measurement is based on the previous research program undertaken by U.S. Employment Service.

Operational Hypotheses

In this section, the conceptual hypotheses discussed in the previous chapter are restated in terms suitable for statistical analyses.

Relationships Among Performance Levels

The relationships between the different performance levels were hypothesized to be positive because the global performance construct refers to multiple dimensions of job performance. Yet the conceptual model used in this study predicts that these different levels are not equivalent. Although there will be a positive relationship between the different levels of performance, there will be a stronger convergence between measures that are "matched" with regard to the specific level of performance.

Between Performance Levels: The first conceptual hypothesis was:

H1(a) There will be a positive relationship between an objective measure of dimensional performance and a subjective measure of global performance (DPO-GP).

This can be restated as:

OH1(a) The correlation between the modal production level and the overall rating score will be greater than zero.

This relationship was hypothesized to reflect the common criterion, production output, shared by the dimensional performance measure and the global rating score. The modal production level refers to the 20-week production distribution ("earned" wages). This is the nonjudgmental measure of the performance dimension Production Output, collected from the weekly production earnings report. The overall rating score refers to the overall performance rating item. This is the judgmental measure of the global performance level, reflecting total job success.

To test this relationship, the modal production level was correlated with the overall rating by means of Pearson's product-moment correlation. This correlation coefficient was then tested to see if it was significantly different from zero.

The next hypothesis stated that:

H1(b) There will be a positive relationship between a subjective measure of dimensional performance and a subjective measure of global performance (DPS-GP).

In operational terms:

OH1(b) The correlation between the "Quantity of Production" rating and the overall rating score will be greater than zero.

This relationship is believed to reflect not only the shared criterion dimension (Production Output) but also common method variance. The "Quantity of Production" score refers to the judgmental rating of performance on the dimensional performance level, whereas the overall score refers to the judgmental rating of performance on the global level. These rating scores (on dimensional and global performance) were correlated via the product-moment correlation and tested for a significant difference from zero.

It was argued above that the subjective measures of different performance levels contain common method variance which may inflate the relationship between different levels. Conceptually, the hypothesis stated that:

H1(c) There will be a stronger relationship between a subjective measure of dimensional performance and global performance than between an objective measure of dimensional performance and the global performance score (DPS-GP > DPO-GP).

This hypothesis was rephrased as:

OH1(c) The relationship between the "Quantity of Production" rating with the overall rating will be greater than the relationship between the modal production level with the overall rating.

This relationship reflects the differential influence of measurement methods on global performance. While global performance is conceptually a function of multiple dimensions, the common method variance (H1b) provides a further explanation for the association between these levels. To

test this hypothesis, two analyses were conducted. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from these two simple linear regression models. One model postulates that Global Performance is a function of the judgmental measure of production output; the other states that Global Performance is a function of the non-judgmental measure of production output. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

Within Performance Levels: Regarding the relationship of multiple criteria within the dimensional performance level, it was stated that:

H2(a) There will be a positive relationship between a subjective measure of dimensional performance and an objective measure of dimensional performance (DPS-DPO).

Thus:

OH2(a) The correlation between the "Quantity" rating and the modal production level will be greater than zero.

This relationship is hypothesized to be a reflection of the matched specificity of performance constructs. These intra-level performance measures were correlated (Pearson's product-moment correlation) and then tested for a significant difference from zero.

The final hypothesis regarding relationships among performance levels was:

H2(b) There will be a stronger relationship between a subjective and objective measure of dimensional performance than between the subjective measures of dimensional performance and global performance (DPS-DPO > DPS-GP).

Thus:

OH2(b) The relationship between the "Quantity" rating with the modal production level will be greater than the relationship between the "Quantity" rating with the overall rating.

This difference was based on the argument that the convergence between different methods of measuring the same performance construct (same level) will outweigh the relationship between common methods of measuring different constructs. In order to test this hypothesis, two analyses were conducted. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from these two simple linear regression models. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

From the foregoing analyses, the equivalence of different performance constructs can be examined. Although all of the pairwise correlations are predicted to be positive [H1(a), H1(b), H2(a)], the strength of this relationship differs regarding the level of specificity with which the constructs are defined. By matching the definitions of the construct domain, the equivalence of different operational measures of "performance" can be ascertained.

Relationships Among Performance Referents

The Dimensional Performance level was hypothesized to consist of sub-levels of performance constructs, these sub-levels being construed as different referents for dimensional performance evaluation. Furthermore, these different referents were believed to be nonequivalent in the sense that they refer to a different frame of reference for measuring job performance. Thus it was argued that within the performance dimension Production Output, there will be a high convergence between performance measures that are "matched" on the referent construct and a low relationship

between performance measures that are referring to different reference points for performance evaluation.

Between Performance Referents The following four hypotheses relate to the relationships between different performance referents (H3a - H3d).

H3(a) There will be a positive relationship between a subjective measure of "typical" dimensional performance and a subjective measure of "maximal" dimensional performance (DPS-MaxS).

Thus:

OH3(a) The correlation between the typical "Quantity" rating score and the maximal "Quantity" rating score will be greater than zero.

Both of the measures refer to a shared dimension of performance (output) and are measured using the same method of measurement (judgmental ratings). The performance referents are different, however, with regard to the context of the appraisal. This relationship is expected to be positive based on the fact that each variable is capturing the same dimension of performance as well as common method variance. The relationship between these variables was determined by correlating the measures (product-moment correlation) and then testing this correlation to see whether it was significantly different from zero.

In addition:

H3(b) There will be a positive relationship between an objective measure of "typical" Dimensional Performance an objective measure of "maximal" Dimensional Performance (DPO-MaxO).

Thus:

OH3(b) The correlation between the modal production level and the maximum production level will be greater than zero.

These measures also reflect a common method of performance measurement yet different referents for measurement. The relationship between these variables was determined by correlating the measures (product-moment correlation) and then testing this correlation to see whether it was significantly different from zero.

It was also hypothesized that:

H3(c) There will be a positive relationship between a subjective measure of "typical" dimensional performance and an objective measure of "maximal" dimensional performance (DPS-MaxO).

In operational terms:

OH3(c) The correlation between the typical "Quantity" rating score and the maximum production level will be greater than zero.

These measures refer to different referents of performance measurement as well as different methods of measurement. Thus the relationship is expected to reflect only the shared dimension of performance; i.e., the Production Output dimension. To determine the degree of association between these measures, the performance measures were correlated via the product-moment correlation and the correlation was tested for a significant difference from zero.

H3(d) There will be a positive relationship between a subjective measure of "maximal" dimensional performance and an objective measure of "typical" dimensional performance (DPO-MaxS).

Thus:

OH3(d) The correlation between the maximal "Quantity" rating and the modal production level will be greater than zero.

This hypothesis also refers to the relationship between different referents and different methods of performance measurement. To determine the degree of association between these measures, the

performance measures were correlated via the product-moment correlation and the correlation was tested for a significant difference from zero.

Within Performance Referents: The following hypotheses were made with respect to the convergence between multiple methods of measuring the same dimension of performance, with the construct defined in terms of both a common performance dimension (Production Output) and a common performance referent. First:

H4(a) There will be a positive relationship between an objective measure of "typical" dimensional performance and a subjective measure of "typical" dimensional performance (DPS-DPO).

In operational terms:

OH4(a) The correlation between the modal production level and the typical "Quantity" rating will be greater than zero.

This is the same hypothesis made above (OH2a). This hypothesis is repeated here because these measures also refer to "typical" levels of performance on the Production Output dimension. To assess the convergence between these measures, they were correlated by means of the product-moment correlation and then tested for a statistically significant difference from zero.

Also:

H4(b) There will be a positive relationship between an objective measure of "maximal" dimensional performance and a subjective measure of "maximal" dimensional performance (MaxS-MaxO).

In operational terms:

OH4(b) The correlation between the peak production level and the maximal "Quantity" rating will be greater than zero.

The maximum production level refers to the peak level of production output achieved by each individual during the 20-week period. The maximal "Quantity" score refers to the appraisal dimension score which utilized a maximal (feasible) frame of reference for rating employees. The convergence between these multiple measures of maximal performance referents was determined by correlating the peak production level with the maximal production dimension score from the performance appraisal, and then testing this correlation for a significant difference from zero.

It was argued previously that multiple measures of the same performance referent construct (Typical or Maximal) will show convergence due to a common context of performance evaluation. It was also hypothesized that measures reflecting different performance referents for dimensional performance evaluation will be spuriously related due to common method variance and/or the shared dimension of performance (output). The following hypotheses were thus stated to compare the relationships of matched referent constructs versus different referent constructs with a common dimension (H5a - H5d) and with a common method (H5e - H5h). The comparison of matched referents versus a common dimension are stated first. With regard to the "maximal" dimensional performance construct:

H5(a) There will be a stronger relationship between an objective measure of "maximal" dimensional performance and a subjective measure of "maximal" dimensional performance than between an objective measure of "maximal" dimensional performance and a subjective measure of "typical" dimensional performance (MaxS-MaxO > DPS-MaxO).

Restated, this relationship is:

OH5(a) The relationship between the maximal production level and the maximal "Quantity" rating will be greater than the relationship between the maximal production level and the typical "Quantity" rating score.

This relationship states that the convergence of multiple measures of the same maximal production dimension will outweigh the relationship between different performance referents that may be related due only to the shared dimension of performance, production output. To test this hy-

pothesis, two analyses were conducted. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from these two simple linear regression models. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

Additionally:

H5(b) There will be a stronger relationship between an objective measure of "maximal" dimensional performance and a subjective measure of "maximal" dimensional performance than between an objective measure of "typical" dimensional performance and a subjective measure of "maximal" dimensional performance (MaxS-MaxO > DPO-MaxS).

Restated, this relationship is:

OH5(b) The relationship between the maximal production level and the maximal "Quantity" rating will be greater than the relationship between the modal production level and the maximal "Quantity" rating.

To test this hypothesis, two analyses were conducted. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from these two simple linear regression models. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

The same relationships stated above also relate to the "typical" dimensional performance construct. Thus:

H5(c) There will be a stronger relationship between an objective measure of "typical" dimensional performance and a subjective measure of "typical" dimensional performance than

between an objective measure of "maximal" dimensional performance and a subjective measure of "typical" dimensional performance (DPO-DPS > DPS-MaxO).

Operationally:

OH5(c) The relationship between the modal production level and the typical "Quantity" rating will be greater than the relationship between the maximal production level and the typical "Quantity" rating.

In addition:

H5(d) There will be a stronger relationship between an objective measure of "typical" dimensional performance and a subjective measure of "typical" dimensional performance than between an objective measure of "typical" dimensional performance and a subjective measure of "maximal" dimensional performance (DPO-DPS > DPO-MaxS).

Restated, this relationship is:

OH5(d) The relationship between the modal production level and the typical "Quantity" rating will be greater than the relationship between the modal production level and the maximal "Quantity" rating.

Both of these hypotheses will be analyzed in terms of correlation and prediction errors. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from the simple linear regression models. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

The following hypotheses refer to the comparison of the matched referents versus common methods of measurement. With regard to the "maximal" dimensional performance construct:

H5(e) There will be a stronger relationship between an objective measure of "maximal" dimensional performance and a subjective measure of "maximal" dimensional performance than between an objective measure of "typical" dimensional performance and an

objective measure of "maximal" dimensional performance (MaxS-MaxO > DPO-MaxO).

Restated, this relationship is:

OH5(e) The relationship between the maximal production level and the maximal "Quantity" rating will be greater than the relationship between the modal production level and the maximal production level.

This relationship states that the convergence of different methods of measuring the same performance referent will outweigh the relationship between the different referents which were measured via the same method.

Also

H5(f) There will be a stronger relationship between an objective measure of "maximal" dimensional performance and a subjective measure of "maximal" dimensional performance than between a subjective measure of "typical" dimensional performance and a subjective measure of "maximal" dimensional performance (MaxS-MaxO > DPS-MaxS).

Therefore:

OH5(f) The relationship between the maximal production level and the maximal "Quantity" rating will be greater than the relationship between the typical "Quantity" rating and the maximal "Quantity" rating.

Both of these hypotheses will be analyzed in terms of correlation and prediction errors. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from the simple linear regression models. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

These same relationships also pertain to the "typical" referent of dimensional performance.

H5(g) There will be a stronger relationship between an objective measure of "typical" dimensional performance and a subjective measure of "typical" dimensional performance than between an objective measure of "typical" dimensional performance and an objective measure of "maximal" dimensional performance (DPS-DPO > DPO-MaxO).

This was restated as:

OH5(g) The relationship between the modal production level and the typical "Quantity" rating will be greater than the relationship between the modal production level and the maximal production level.

The final hypothesis regarding differences in performance referents states that:

H5(h) There will be a stronger relationship between an objective measure of "typical" dimensional performance and a subjective measure of "typical" dimensional performance than between a subjective measure of "typical" dimensional performance and a subjective measure of "maximal" dimensional performance (DPS-DPO > DPS-MaxS).

Operationally:

OH5(h) The relationship between the modal production level and the typical "Quantity" rating will be greater than the relationship between the typical "Quantity" rating and the maximal "Quantity" rating.

Both of these hypotheses will be analyzed in terms of correlation and prediction errors. First, differences in the quality of fit were examined by means of comparing (1) the magnitude of fitted errors and (2) the proportion of residual (fitted) error to total model variance (R-Square). Second, differences in the quality of prediction were examined by means of comparing the prediction errors resulting from the simple linear regression models. For each prediction model, the PreSS residuals were computed and then compared in terms of the (1) magnitude of prediction errors, and (2) the percent of prediction error to total model variance.

Relationships between Ability and Performance

From the foregoing explication of performance constructs, the question arises as to the predictability of these various criteria. The conceptual model for this study posits that a measure of maximal dimensional performance may in fact be different from a measure of typical dimensional performance. If these criteria represent different constructs (or parameters) of performance, are they predicted equally well by the same predictor constructs? Given that the different levels and referents of job performance may not be equivalent (interchangeable), the need to examine the relative predictability of these measures is apparent.

The relationships between individual abilities (cognitive and psychomotor) and performance were hypothesized to vary, in terms of strength and power, depending upon the construct (i.e., parameter) of the performance measure. While all of the ability-performance relationships are predicted to be positive, the predictive power of the ability measures is expected to be greater for the Maximal dimensional performance level criteria than for the Typical dimensional performance measures. This differential prediction is based on the distinction between these performance constructs with respect to the constraints placed on the observed criterion. In addition, the ability measures are expected to be more predictive of the Dimensional performance criteria that of the Global Performance criteria due to the level of specificity with which the criteria are measured. By matching the level of specificity of both the criterion and the predictor, a stronger predictive relationship is expected. These hypotheses are described below.

The positive nature of the relationships between ability-performance were conceptually stated as follows:

H6(a) There will be a positive relationship between individual abilities and Maximal dimensional performance (Abilities-MaxO/MaxS).

H6(b) There will be a positive relationship between individual abilities and Typical dimensional performance (Abilities-DPO/DPS).

H6(c) There will be a positive relationship between individual abilities and Global Performance (Abilities-GP).

These relationships were restated as:

OH6(a) The correlation between the maximum production level and the cognitive and psychomotor abilities will be greater than zero. The correlation between the maximum "Quantity" rating and these abilities will also be greater than zero.

OH6(b) The correlation between the modal production level and the cognitive and psychomotor abilities will be greater than zero. The correlation between the typical "Quantity" rating and these abilities will also be greater than zero.

OH6(c) The correlation between the overall rating score and the cognitive and psychomotor abilities will be greater than zero.

Each of these relationships states that "performance" (whether it be global performance, typical dimensional performance, or maximal dimensional performance) is a function of both cognitive and psychomotor abilities. These hypotheses were tested by examining the product-moment correlations. Correlation coefficients were computed for each performance measure as it was related to the cognitive ability scales and to the psychomotor ability scales. These relationships (bivariate correlations) were then tested for a statistically significant difference from zero.

The relative predictive quality of these linear models was hypothesized as follows:

H7(a) There will be a stronger relationship between individual abilities and Maximal dimensional performance than between individual abilities and Typical dimensional performance (Abilities-MaxO/MaxS > Abilities-DPO/DPS).

This hypothesis was restated as:

OH7(a) The maximum production level (and maximal "Quantity" rating) will be more accurately predicted by the ability tests than the modal production level (and typical "Quantity" rating).

It was argued previously that the maximal performance measures refer more to an individual's potential level of performance. In contrast, typical performance on the dimensional performance level is construed as a function of not only abilities but also motivation and opportunity bias. To test this relationship, regression analyses were used in order to compare these performance criterion models in terms of quality of fit and quality of prediction. Quality of fit was assessed by comparing the models in terms of R-Square, SSE, and the Coefficient of Variation. The quality of prediction was assessed via R-Square of Prediction, PreSS, and the Coefficient of Variation of Prediction.

The following two hypotheses were made regarding the prediction of different levels of performance.

H7(b) There will be a stronger relationship between individual abilities and Maximal dimensional performance than between individual abilities and Global Performance (Abilities-MaxO/MaxS > Abilities-GP).

Therefore:

OH7(b) The maximum production level (and maximal "Quantity" rating) will be more accurately predicted by the ability predictors than the overall rating score.

This hypothesis refers to the predictability of two performance measures that differ in terms of both the level of performance measurement (dimensional versus global performance) and the referent of performance measurement (maximal versus typical). The analyses here consisted of the same model comparisons made for the previous two hypotheses. Quality of fit was assessed by comparing the models in terms of R-Square, SSE, and the Coefficient of Variation. The quality of prediction was assessed via R-Square of Prediction, PreSS, and the Coefficient of Variation of Prediction.

In addition:

H7(c) There will be a stronger relationship between individual abilities and Typical dimensional performance than between individual abilities and Global Performance (Abilities-DPO/DPS > Abilities-GP).

This hypothesis was restated as:

OH7(c) The modal production level (and typical "Quantity" rating) will be more accurately predicted by the ability predictors than the overall rating score.

In this hypothesis, both performance measures refer to typical performance evaluations (Typical output and typical overall performance). However, the Typical dimensional performance measures are confined to one dimension (production output) whereas the Overall rating score is a measure of the overall job success. To test this relationship, the same procedures for model comparison were used here as those used to test OH7(a). Quality of fit was assessed by comparing the models in terms of R-Square, SSE, and the Coefficient of Variation. The quality of prediction was assessed via R-Square of Prediction, PreSS, and the Coefficient of Variation of Prediction.

Summary

The table on the next few pages summarizes the foregoing research hypotheses according to the comparisons being made.

FIGURE 2
Summary of Research Hypotheses

Relationships Between Performance Levels

H1(a) There will be a positive relationship between a objective measure of Dimensional Performance and a subjective measure of Global Performance. **(DPO-GP)**

H1(b) There will be a positive relationship between a subjective measure of Dimensional Performance and a subjective measure of Global Performance. **(DPS-GP)**

H1(c) There will be a stronger relationship between subjective measures of Dimensional Performance and Global Performance than between an objective measure of Dimensional Performance and the Global Performance rating. **(DPS-GP > DPO-GP)**

Relationships Within Performance Levels

H2(a) There will be a positive relationship between a subjective measure of Dimensional Performance and an objective measure of Dimensional Performance. **(DPS-DPO)**

H2(b) There will be a stronger relationship between a subjective and objective measure of Dimensional Performance than between the subjective measures of Dimensional Performance and Global Performance. **(DPS-DPO > DPS-GP)**

Relationships Between Performance Referents

H3(a) There will be a positive relationship between a subjective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance. **(DPS-MaxS)**

H3(b) There will be a positive relationship between an objective measure of "typical" Dimensional Performance an objective measure of "maximal" Dimensional Performance. **(DPO-MaxO)**

H3(c) There will be a positive relationship between a subjective measure of "typical" Dimensional Performance and an objective measure of "maximal" Dimensional Performance. **(DPS-MaxO)**

H3(d) There will be a positive relationship between a subjective measure of "maximal" Dimensional Performance and an objective measure of "typical" Dimensional Performance. **(DPO-MaxS)**

Relationships Within Performance Referents

H4(a) There will be a positive relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance. **(DPS-DPO)**

H4(b) There will be a positive relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance. **(MaxS-MaxO)**

Figure 2 continued

Comparisons of Matched Referents v. Common Dimension

H5(a) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between an objective measure of "maximal" Dimensional Performance and a subjective measure of "typical" Dimensional Performance. **(MaxS-MaxO > DPS-MaxO)**

H5(b) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance. **(MaxS-MaxO > DPO-MaxS)**

H5(c) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between an objective measure of "maximal" Dimensional Performance and a subjective measure of "typical" Dimensional Performance. **(DPO-DPS > DPS-MaxO)**

H5(d) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance. **(DPO-DPS > DPO-MaxS)**

Comparisons of Matched Referents v. Common Method

H5(e) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and an objective measure of "maximal" Dimensional Performance. **(MaxS-MaxO > DPO-MaxO)**

H5(f) There will be a stronger relationship between an objective measure of "maximal" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance than between a subjective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance. **(MaxS-MaxO > DPS-MaxS)**

H5(g) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between an objective measure of "typical" Dimensional Performance and an objective measure of "maximal" Dimensional Performance. **(DPS-DPO > DPO-MaxO)**

H5(h) There will be a stronger relationship between an objective measure of "typical" Dimensional Performance and a subjective measure of "typical" Dimensional Performance than between a subjective measure of "typical" Dimensional Performance and a subjective measure of "maximal" Dimensional Performance. **(DPS-DPO > DPS-MaxS)**

Figure 2 continued

Relationships Between Ability & Performance

H6(a) There will be a positive relationship between individual abilities and "maximal" Dimensional Performance. **(Abilities-MaxO/MaxS)**

H6(b) There will be a positive relationship between individual abilities and "typical" Dimensional Performance. **(Abilities-DPO/DPS)**

H6(c) There will be a positive relationship between individual abilities and Global Performance. **(Abilities-GP)**

Relative Predictability of Predictor-Criterion Models

H7(a) There will be a stronger relationship between individual abilities and "maximal" Dimensional Performance than between individual abilities and "typical" Dimensional Performance. **(Abilities-MaxO/MaxS > Abilities-DPO/DPS)**

H7(b) There will be a stronger relationship between individual abilities and "maximal" Dimensional Performance than between individual abilities and Global Performance. **(Abilities-MaxO/MaxS > Abilities-GP)**

H7(c) There will be a stronger relationship between individual abilities and Typical Dimensional Performance than between individual abilities and Global Performance. **(Abilities-DPO/DPS > Abilities-GP)**

Chapter 4: Results

Overview

In the preceding chapter, the research methodology for this study was described. In this chapter, the empirical results of this research study are presented. The first section, ***Criterion Validity,*** presents the research findings of Hypotheses 1(a) through 5(h). These hypotheses were analyzed using both correlational and regression techniques, emphasizing statistical measures of quality of fit and quality of prediction. H1(a-b), H2(a), H3(a-b) and H4(a-b) predicted positive relationships among the alternative measures of job performance. These results, which appear in Table 1, are based on bivariate correlations between the criterion variables. H1(c), H2(b), and H5(a-h) predicted the relative strengths between these criteria relationships. These results, found in Tables 2-4, are based on regression analyses which incorporate both quality of fit and quality of prediction comparisons.

The second section, ***Predictive Validity,*** presents the research findings of Hypotheses 6(a) through 7(c). H6(a-c) predicted positive relationships between abilities (cognitive and psychomotor) and each of the job performance criteria. These results, found in Table 5, are based on bivariate correlation coefficients for each of the postulated predictor-criterion relationships. H7(a-c) predicted differential prediction of the alternative job performance criteria using the same ability predictors. These results, found in Tables 6, 7, and 8 are based on multiple regression analyses which examine both the qualities of fit and prediction.

Performance Characteristics

In Chapter 3, the demographic characteristics of the research samples were described and the similarity between the two samples was noted. In this section, the performance characteristics of the two samples are listed. The performance characteristics of the Criteria Validation sample (n = 153) are as follows:

Measure	Mean	S.D.	Range
Global Performance	3.0	0.88	1 - 5
Dimensional Performance:			
Subjective Typical	2.1	1.07	1 - 5
Objective Typical	$3.44	1.17	$1.25 - $8.00
Subjective Maximal	$4.78	1.19	$2.99 - $7.36
Objective Maximal	$4.29	1.31	$2.12 - $10.14

The performance characteristics of the Predictive Validation sample (n = 117) are approximately equal to the above sample. The similarity between the two samples in terms of performance as well as demographic characteristics indicates that these two samples are equivalent.

Measure	Mean	S.D.	Range
Global Performance	3.0	0.89	1 - 5
Dimensional Performance:			
Subjective Typical	2.1	1.01	1 - 5
Objective Typical	$3.40	1.08	$1.25 - $8.00
Subjective Maximal	$4.75	1.16	$2.99 - $7.36
Objective Maximal	$4.24	1.14	$2.12 - $8.46

Criterion Validity

This portion of the research findings provides a test of the Job Performance research model shown in Figure 1. This model identified two different levels of job performance measurement--global and dimensional. Within the dimensional level, two different referents of job performance measurement were defined--typical and maximal. The purpose of this model, and this study, was

to examine the interrelationships among different levels and referents of job performance measurement, using different methods of measurement (judgmental and nonjudgmental methods). Table 1 contains the correlations among all of the job performance criteria. Table 2 describes the relationships between and within the different ***levels*** of job performance; i.e., global and dimensional performance variables. Tables 3 and 4 describe the relationships between and within the different ***referents*** for measuring dimensional job performance. In discussing these results, the term "model" is used to refer to the postulated criteria relationships. This terminology was used because these relationships are in fact based on a conceptual model which relates these criteria; rational hypotheses regarding the strength and direction of each relationship have been developed as a foundation for the criteria models.

Relationships Among Performance Levels

The correlations in Table 1 reveal positive, and significant, relationships among all of the job performance criteria ($p < .0001$). These correlations, which reflect the strength of the linear relationship between the alternative job performance measures, range from 0.49 to 0.91. With regard to the convergence between ***different methods*** of measurement, the correlations between the subjective Global Performance rating (GP) and both of the objective production criteria are 0.53 and 0.58. In addition, the correlations between the Typical (dimensional) production rating (DPS) with both of the objective criteria are 0.74 and 0.77. Lastly, the correlations between the Maximal (dimensional) production rating (MaxS) with the objective criteria are 0.69 and 0.72.

Of particular interest here are the relationships among ***different levels*** of job performance measurement. As shown in Table 1, the convergence between the objective and subjective measures of dimensional job performance, using the Typical frame of reference (DPS-DPO), is greater than the convergence of criteria measured on different levels. Table 2 further explores these relationships in terms of quality of fit (i.e., how well does the data fit the postulated criterion model) and quality of prediction (i.e., will the dimensional criteria predict global responses adequately).

TABLE 1
Job Performance Criteria
Correlations

Correlations among five alternative measures for job performance, using two different methods of measurement.

	GP	DPS	DPO	MaxS	Mean	Standard Deviation
Subjective Global Performance--"Overall" (GP)	---	---	---	---	3.0	0.88
Subjective Dimensional Performance-"Typical"(DPS)	.63	---	---	---	2.1	1.07
Objective Dimensional Performance-"Typical"(DPO)	.53	.74[a,b]	---	---	$3.44	1.17
Subjective Dimensional Performance-"Maximal" (MaxS)	.49	.56	.69	---	4.78	1.19
Objective Dimensional Performance--"Maximal" (MaxO)	.58	.77	.91	.72[b]	4.29	1.31

NOTE: Sample size = 153.
All correlations are statistically significant @ $p < .0001$.

"a" denotes "matched" levels of job performance measurement: i.e., DPS-DPO are Matched Levels -- Dimensional.
"b" denotes "matched" referents of dimensional job performance measurement: i.e., DPS-DPO are Matched "Typical" Referents and MaxS-MaxO are Matched "Maximal" Referents.

This analysis involves a comparison of different criteria models with respect to the statistical measures of fit and prediction, a comparison which is conceptually-oriented more so than "testing"-oriented. The statistical measures shown in Table 2 are calculated as ratios in order to allow a direct comparison between the models. Because some of the criteria models possess different amounts of inherent response variability, it is not always possible to compare the absolute value of the statistics.

The Quality of Fit (QOF) comparisons involve R-Square (the proportion of total variation in the response data that is "explained" by the model), Coefficient of Variation (the proportion of the mean response of the model that is "fitting" error), and the ratio of "Signal-to-Noise" (measured by the Regression variation divided by the "error" variation of the model). The Quality of Prediction (QOP) comparisons involve the R-Square of Prediction (the proportion of total variation in the response data that is "predictable" by the model), Coefficient of Variation for Prediction (the proportion of the mean response of the model that is prediction error), and the predictable ratio of Signal-to-Noise (measured with the prediction error variation of the model rather than the standard fitting error variation of the model). While many of these statistics provide the same information and might appear to be redundant, each statistical ratio is actually providing additional information as to fit and prediction. For example, the R-Square refers to the variance explained by the model relative to the total variation in the model. Yet the R-Square statistic may be artificially high due to either a large regression slope and/or a large spread in the data (Myers, 1986). The Coefficient of Variation focuses on the spread in the data and expresses the fitted error relative to the mean response for the model. Further, the Signal-to-Noise ratio focuses on the slope of the model, relative to the standard error variation. In essence, each of these statistics are expressing the quality of fit as it is gauged by different model characteristics.

A comparison of Columns 1-2 reveals that the relationship between Global Performance and the Subjective measure of Dimensional Performance (GP-DPS) is stronger than the relationship between Global Performance and the Objective measure of Dimensional Performance (GP-DPO). In terms of fit, the former model possesses a higher R-Square, thus "explaining" a larger proportion

TABLE 2
Performance Level Comparisons:
Global & Dimensional Relationships

Summary of statistical qualities of Fit and Prediction, using the alternative job performance criteria.

Models:	**GP &**	**GP &**	**DPS &**
Quality of Fit:	**DPS**	**DPO**	**DPO**
R^2	0.39	0.28	0.54
Signal-to-Noise	0.64	0.38	1.18
Coefficient of Variation	22.99%	25.10%	34.54%
Quality of Prediction:			
R^2 of Prediction	0.38	0.25	0.53
Prediction Signal-to-Noise	0.60	0.34	1.13
Prediction Coefficient of Variation	23.24%	25.38%	34.83%
Model Characteristics:			
Average Response (y)	3.00	3.00	2.09
Total Model Variation (SST)	118.99	118.99	172.72
Fitting Variation (SSE)	72.19	85.97	78.80
Prediction Variation (PreSS)	74.34	88.72	81.09

NOTE: Sample size = 153.
All criteria models (relationships) are statistically significant @ $p < .0001$; i.e., the F-tests indicate that the ratio of variance explained by the model-to-variance due to model (fitting) error is statistically significant.

of the variation in Global Performance responses. The Signal-to-Noise ratio for GP-DPS is also larger, suggesting that this relationship is marked by a strong regression slope relative to the standard error of the fitted regression. Lastly, the Coefficient of Variation is only 23% of the mean GP rating, as opposed to approximately 25% for the GP-DPO model. The model differences in fit and prediction can be seen directly by comparing the Model Characteristics section. Although both of the models in Columns 1 and 2 contain the same amount of total model variation, and the same mean response, the GP-DPS model has a smaller residual variation than that of the GP-DPO model and a smaller prediction variation as well.

With respect to prediction, the GP-DPS model has a higher prediction capability; i.e., Global Performance is more predictable based on a knowledge of Subjective dimensional ratings than on a knowledge of Objective dimensional production levels, as evidenced by the higher R-Square of Prediction, higher prediction Signal-to-Noise, and lower Coefficient of Variation for Prediction. Furthermore, the difference between fit and prediction for the GP-DPS model is small: the R-Square statistic declines only 1% upon internal cross-validation. For the GP-DPO model, this decline in R-Square due to cross-validation is 3%.

A comparison of Column 3 with Column 1 refers to the comparison of "matched levels" using different methods versus different levels using the same method of measurement (subjective). As predicted, the criteria convergence (correlation) within the dimensional performance level is stronger than either of the between-level relationships. These two models (GP-DPS and DPS-DPO) possess different model characteristics. For instance, the total model variation (SST) is greater with the matched dimensional model (45% more total variation), yet the increase in error variation is not as large (only 9% more error variation). In fact, both the quality of fit and the quality of prediction are strengthened by the convergence of job performance criteria that are matched according to the level of measurement, thus matched specificity. The R-Square type statistics indicate that the DPS-DPO model explains a higher proportion of the variance in GP responses and also is capable of predicting a higher proportion of response variance. The ***fitted*** signal-to-noise ratio for the within-level model is 1.18, which means that the explained variance in the model is almost one and

a half times greater than the standard error of the regression. Upon internal cross-validation, this ratio is approximately the same, suggesting that the prediction errors are not significantly affecting the explanatory nor predictability powers of the model. These findings emphasize the fact that although the total model variation (SST) is larger for the matched dimensional model, the amount of error (fitted and prediction) is minimized. However, the Coefficient of Variation, for both fit and prediction, is larger for this "matched" model--a finding which indicates that the dispersion (error) around the regression line is greater than either of the GP-DP models. This increased Coefficient of Variation is based largely on the increase in SST for the matched dimensional model.

These findings are now stated as they relate to H1(a) through H2(b) (refer to Figure 2, Chapter 3).

H1(a) is Supported: As seen in Table 1, the relationship between Global Performance and Objective Dimensional Performance (GP-DPO) is positive and significantly different from zero ($r = 0.53$). This hypothesis reflected the presence of a shared criterion dimension; although measured on different levels, both of these criteria capture the specific dimension "Quantity of Production", though in different degrees.

H1(b) is Supported: As seen in Table 1, the relationship between Global Performance and Subjective Dimensional Performance (GP-DPS) is positive and significantly different from zero ($r = 0.63$). This hypothesis was based on the presence of a shared criterion dimension in addition to a common method variance.

H1(c) is Supported: As shown in Table 2, there is a stronger relationship between GP-DPS than between GP-DPO. The basis for this hypothesis was that although both relationships are between different performance measurement levels, the former will be stronger due to a common method variance which acts to inflate the global-dimensional relationship. Because of the commonality of the subjective measures, this relationship was predicted to be higher.

Table 2, Columns 1 and 2, shows that the GP-DPS relationship is marked by a higher degree of fit.

H2(a) is Supported: As seen in Table 1, the relationship between Subjective Dimensional Performance and Objective Dimensional Performance (DPS-DPO) is positive and significantly different from zero ($r = 0.74$). This hypothesis refers to the higher degree of convergence within the dimensional levels of performance measurement. This relationship between matched conceptual criteria is strong despite the different methods of measurement.

H2(b) is Supported: As shown in Table 2, there is a stronger relationship between DPS-DPO than between GP-DPS. This hypothesis was based on the argument that the convergence between different methods of measuring a single job performance construct (i.e., same level of specificity) will be greater than the relationship between different levels of performance that share a common method of measurement. Table 2 indicates that the matched specificity of DPS-DPO results in a marked increase in both the quality of fit (i.e., the R-Square increases 15%) as well as the quality of prediction (i.e., the R-Square of Prediction also increases 15%). Yet the dispersion around the regression line, as measured by the Coefficient of Variation statistics, is higher for the DPS-DPO model, thus indicating that there is more "noise" in this matched model, measured as a percent of the mean response.

Overall, the findings detailed in Table 2 indicate that the degree of convergence between job performance criteria is in fact affected by the ***level*** of measurement. At this point, these findings suggest that the relationship--and hence equivalence--between alternative measures of job performance is contingent upon the level of specificity with which criteria are compared. Whereas previous studies have generally concluded that ratings and objective performance measures are not equivalent (see Chapter 2), it was argued that these criteria comparisons did not account for the impact of different levels of measurement; typically, these studies compared subjective Global Performance measures with specific Objective dimensional measures. In this study, the convergence between the objective data and global ratings was significant (GP-DPO), yet the convergence be-

tween the objective data and a conceptually equivalent subjective rating (DPS-DPO) outweighed the former relationship and additionally outweighed the convergence between common methods (GP-DPS).

Relationships Among Performance Referents

Tables 3 and 4 compare the criteria relationships within and between the referents for dimensional job performance measurement. Column 1 of each table refers to the "matched" referent model, measured using different methods. Columns 2-3 present criteria models tapping different performance referents yet using different methods of measurement. Columns 4-5 represent criterion models tapping different referents yet using a common method of measurement. The Qualities of Fit and Prediction statistics are the same as those used in the foregoing comparisons (Table 2).

First, a comparison of the ***matched typical*** criteria (Table 3) with the "mixed" models indicates that two of the predictions regarding the relative strength of criteria relationships were not supported by the data. Column 2, Subjective Typical with Objective Maximal criteria (DPS-MaxO), and Column 4, Objective Typical with Objective Maximal criteria (DPO-MaxO), show that the relationships between these different referents were stronger than the matched Typical criteria in terms of both fit and prediction. The DPS-MaxO model (column 2) contains the same amount of total variation as the matched typical model yet a smaller amount of error variation (both fitted and prediction error). The qualities of fit and prediction are better for the DPS-MaxO model than for the DPS-DPO model, as evidenced by the difference in R-Square and Signal-to-Noise. For the DPO-MaxO model (column 4), the proportion of variation in DPO (the modal production level) that is "explained" by the maximal production level is 0.83. The Signal-to-Noise ratio is markedly stronger (higher) for the DPO-MaxO model, or approximately four times larger for this model that for the "matched" model. The Coefficient of Variation is the smallest for the DPO-MaxO model--the dispersion around the regression line, as measured by the standard "fitted error", is approxi-

TABLE 3
Performance Referent Comparisons:
"TYPICAL" Referents v. Mixed Models

Summary of statistical qualities of Fit and Prediction, using the alternative job performance criteria.

Quality of Fit:	DPS & DPO	DPS & MaxO	MaxS & DPO	DPO & MaxO	DPS & MaxS
R^2	0.54	0.60	0.48	0.83	0.31
Signal-to-Noise	1.18	1.44	0.87	4.74	0.42
Coefficient of Variation	34.54%	32.35%	17.97%	14.16%	42.33%
Quality of Prediction:					
R^2 of Prediction	0.53	0.59	0.47	0.82	0.29
Prediction Signal-to-Noise	1.13	1.41	0.87	4.62	0.41
Prediction Coefficient of Variation	34.83%	32.80%	18.06%	14.30%	42.80%
Model Characteristics:					
Average Response (y)	2.09	2.09	$4.78	$3.44	2.09
Total Model Variation (SST)	172.72	172.72	2,136,748	2,086,034	172.72
Fitting Variation (SSE)	78.80	69.12	1,116,826	359,181	118.36
Prediction Variation (PreSS)	81.09	71.22	1,142,791	371,011	122.17

NOTE: Sample size = 153.
All criteria models (relationships) are statistically significant @ $p < .0001$; i.e., the F-tests indicate that the ratio of variance explained by the model-to-variance due to model (fitting) error is statistically significant.

mately 14% of the mean response in DPO. In contrast, the Coefficient of Variation for the "matched" model is approximately 35% with regard to fit and prediction.

The other two predictions are supported--the "matched typical" performance model is characterized by a higher degree of fit and prediction when compared with Columns 3 and 5. Although the R-Square statistics and Signal-to-Noise measures indicate that the matched typical model is a better model in terms of fit and prediction, this matched model contains a higher degree of "noise" around the regression line, measured as a percent of the average DPS response, than the MaxS-DPO model. The comparison of the matched Typical model against the Subjective Typical Performance with Subjective Maximal Performance (DPS-MaxS) model is particularly interesting. The DPS-MaxS represents a relationship between matched methods of performance measurement that relate to different performance referents. Note that the matched typical model explains 23% more of the response variation in DPS, and the signal-to-noise ratio is almost three times as large.

Second, a comparison of the ***matched maximal*** performance model with the "mixed" models (Table 4) also produces mixed results. While this model also has a higher degree of fit and prediction than the matched Subjective relationship (i.e., DPS-MaxS), the matched Maximal relationship is outweighed by the matched Objective model (i.e., DPO-MaxO). In addition, the Subjective Typical Performance with Objective Maximal Performance model (DPS-MaxO) is stronger than the matched maximal relationship. In this case, the DPS-MaxO model is characterized by a higher degree of fit (R-Square is 8% higher) and a higher degree of prediction. These findings would suggest that while the models that relate typical performance with Subjective Maximal Performance (Columns 3 and 5) are "weaker" that the matched Maximal model, the influence of Objective Maximal Performance on both of the typical referents is stronger than the influence of Objective Maximal Performance on Subjective Maximal Performance (Columns 2 and 4).

The results shown in Tables 1,3, and 4 are now put into the context of the research hypotheses from Chapter 3 (see Figure 2).

TABLE 4
Performance Referent Comparisons:
"MAXIMAL" Referents v. Mixed Models

Summary of statistical qualities of Fit and Prediction, using the alternative job performance criteria.

Quality of Fit:	**MaxS & MaxO**	**DPS & MaxO**	**MaxS & DPO**	**DPO & MaxO**	**DPS & MaxS**
R^2	0.52	0.60	0.48	0.83	0.31
Signal-to-Noise	1.03	1.44	0.87	4.74	0.42
Coefficient of Variation	17.27%	32.35%	17.97%	14.16%	42.33%
Quality of Prediction:					
R^2 of Prediction	0.51	0.59	0.47	0.82	0.29
Prediction Signal-to-Noise	1.02	1.41	0.87	4.62	0.41
Prediction Coefficient of Variation	17.38%	32.80%	18.06%	14.30%	42.80%
Model Characteristics:					
Average Response (y)	$4.78	2.09	$4.78	$3.44	2.09
Total Model Variation (SST)	2,136,748	172.72	2,136,748	2,086,034	172.72
Fitting Variation (SSE)	1,030,476	69.12	1,116,826	359,181	118.36
Prediction Variation (PreSS)	1,057,682	71.22	1,142,791	371,011	122.17

NOTE: Sample size = 153.
All criteria models (relationships) are statistically significant @ $p < .0001$; i.e., the F-tests indicate that the ratio of variance explained by the model-to-variance due to model (fitting) error is statistically significant.

H3(a) through H3(d) are Supported: As seen in Table 1, the relationships (correlations) between "mixed" referent models of dimensional job performance criteria are positive and significantly different from zero (see Table 1). H3(a) and H3(b) refer to different referents of dimensional performance measurement that use a common method of measurement. The results show correlations of 0.56 for Subjective Typical with Subjective Maximal Performance (DPS-MaxS) and 0.91 for Objective Typical with Objective Maximal Performance (DPO-MaxO). H3(c) and H3(d) refer to different referents using different methods of measurement. The observed correlation between Subjective Typical with Objective Maximal Performance (DPS-MaxO) is 0.77, while the Subjective Maximal with Objective Typical Performance (MaxS-DPO) relationship has a correlation of 0.69.

H4(a) and H4(b) are Supported: As seen in Table 1, the relationships (correlations) between "matched" referent models of dimensional job performance criteria, using different methods of measurement, are positive and significantly different from zero (see Table 1). The matched Typical criteria convergence (correlation) is 0.74, and the matched Maximal convergence (correlation) is 0.72.

H5(a) and H5(b) produce Mixed Support: As shown in Table 4, there is a stronger relationship between matched Maximal criteria than between Subjective Maximal with Objective Typical (MaxS-DPO), *yet* the matched maximal relationship does not outweigh the convergence between Subjective Typical with Objective Maximal (DPS-MaxO). These hypotheses were stated in order to show that the degree of convergence between matched maximal referents will be stronger than that of different referents that share only the dimensional space (i.e., quantity of production). Although the relationship between the modal production level with a rating of maximal production (MaxS-DPO) was not as strong as the matched maximal model, there is a marked influence of the peak production level on the pursuant ratings of typical production levels (DPS-MaxO). However, both of the Coefficient of Variation measures reveal that the amount of "error," measured as a percentage of the mean response, is actually higher for the DPS-MaxO model.

H5(c) and (d) produce Mixed Support: As shown in Table 3, there is a stronger relationship between matched Typical criteria than between the MaxS-DPO model, yet the matched typical relationship does not outweigh the convergence between Subjective Typical with Objective Maximal (DPS-MaxO). The prediction comparisons reveal a 6% higher R-Square of Prediction for the DPS-MaxO model, a higher prediction signal-to-noise, and a lower degree of variation relative to the mean response.

H5(e) and H5(f) produce Mixed Support: As shown in Table 4, there is a stronger relationship between matched Maximal criteria than between Subjective Typical with Subjective Maximal (DPS-MaxS), *yet* the matched maximal relationship does not outweigh the convergence between Objective Typical with Objective Maximal (DPO-MaxO). Although these hypotheses were stated to show that the matched referent relationship would outweigh common method variance coupled with the shared dimension space, the high relationship between typical and maximal production, measured objectively, is much stronger. In fact, the difference in both R-Square statistics is 31%. The strength of the DPO-MaxO model is also evidenced by the higher Signal-to-Noise ratios and lower Coefficient of Variation ratios.

H5(g) and H5(h) produce Mixed Support: As shown in Table 3, there is a stronger relationship between matched Typical criteria than between Subjective Typical with Subjective Maximal (DPS-MaxS), *yet* the matched typical relationship does not outweigh the convergence between Objective Typical with Objective Maximal (DPO-MaxO). Just as in the foregoing hypotheses, these two hypotheses were intended to distinguish between common referent convergence using different methods of measurement as opposed to common method relationships referring to different referents of job performance. In this case, the DPO-MaxO model produces R-Square measures that are 29% higher than that of the matched typical criteria. In addition, the signal-to-noise ratios are substantially higher with the DPO-MaxO model, and the Coefficient of Variation ratios are approximately 40% less than that of the matched typical model (i.e., less "noise").

Summary for Criteria Validation

Overall, these results suggest that the relationships between alternative measures of job performance are positive and significant. Furthermore, the relationship of within-level job performance measurement (DPS-DPO) is stronger than the models referring to different levels of job performance (GP-DP). These findings suggest that by matching the level of performance measurement (hence specificity), a higher degree of criteria convergence can be attained in terms of both model fit and prediction. This relatively high convergence between matched levels of measurement is possible even when different methods of measurement are employed.

With respect to the concept of referents for dimensional job performance, the findings herein produced only partial support for the model. First, the convergence between ***matched typical criteria*** is statistically significant, the quality of fit is relatively high, and the quality of prediction is strong. However, this relationship is not as strong as the relationship between the objective measures of both Typical and Maximal performance (DPO-MaxO) nor the relationship between Subjective Typical with Objective Maximal Performance (DPS-MaxO). Second, the convergence between the ***matched maximal referents*** for job performance is relatively strong in terms of both fit and prediction. However, this convergence is also overpowered by the convergence between DPO-MaxO and DPS-MaxO.

Predictive Validity

Based on the foregoing criterion validation findings, the question arises as to the predictability of the alternative measures of job performance. According to the Job Performance Model in Figure 1, there are different levels and different referents for performance measurement. If these alternative job performance measures represent different performance constructs, there is no reason to believe that one selection procedure will predict these measures equally well (i.e., Guion, 1976). Therefore,

this portion of the research results examines the possibility of differential prediction of the alternative criteria.

Table 5 contains the bivariate correlations among the criteria with cognitive ability and psychomotor ability scale scores. Recall that cognitive ability is composed of three scales: General Intelligence, Verbal Aptitude, and Numerical Aptitude (G, V, and N). Psychomotor ability is composed of Coordination, Finger Dexterity, and Manual Dexterity (K, F, and M) scales. Tables 6 and 7 describe the relative predictability of the alternative criteria, using the quality of fit and quality of prediction statistics described earlier. Differential prediction is examined using, first, the three scales for cognitive ability and, second, for psychomotor ability. Table 8 displays the "best" predictor model for each job performance criteria. These models were determined from a comparison of all possible models, using the six ability/aptitude scales. The choice of "best" model was based on the highest degree of fit (R-Square) combined with the highest degree of prediction (PreSS analysis).

Predictor-Criteria Correlations

The correlations between the separate aptitude scales and the five job performance criteria (Table 5) suggest that only two criteria are significantly related to the singular ability scales. The Global Performance criterion is positively, and significantly, related to two of the cognitive ability scales, General Intelligence (G) and Numerical Aptitude (N). Although Global Performance (GP) is positively related to all of the predictors, none of the psychomotor scales produces a statistically significant relationship. In addition to Global Performance (GP), the Subjective Maximal criterion (MaxS) is also significantly related with the chosen ability scales. In this case, MaxS is positively, and significantly, associated with two of the cognitive ability scales and all of the psychomotor scales.

TABLE 5
Predictor-Criteria Correlations

Correlations between the five alternative job performance criteria with the six ability tests (predictors).

Criteria:	GP	DPS	DPO	MaxS	MaxO
Ability Scales:					
General Intelligence (G):	.25**	.04	.07	.21**	.06
Verbal Aptitude (V):	.15	-.11	-.07	.13	-.09
Numerical Aptitude (N):	.26**	.10	.09	.26**	.11
Coordination (K):	.06	.05	.11	.20**	.15
Finger Dexterity (F)	.03	.10	.05	.26**	.10
Manual Dexterity (M):	.14	.10	.07	.16*	.14

Note: Sample size is 117
*denotes significance @ $p < .10$
**denotes significance @ $p < .05$

Two notable patterns are evident in Table 5. First, the magnitude of the correlations between the cognitive ability scales with Global Performance (GP) and Subjective Maximal (MaxS) performance are in high agreement. For instance, both of these criteria are significantly related to General Intelligence (G) and Numerical Aptitude (N); furthermore, these two criteria are the only measured criteria that are positively related to V (Verbal aptitude). With regard to the cognitive ability scales, the Typical dimensional referents (DPS and DPO) and the Objective Maximal criterion (MaxO) follow similar patterns of relationships with the cognitive ability scales.

Second, the correlational patterns for the psychomotor ability predictors indicate that Subjective Maximal (MaxS) is the only criterion that is significantly related to these predictor scales. Furthermore, the impact of psychomotor tests on the other criteria is not consistent, although all three of these scales are components of psychomotor ability. For example, K (Coordination) is highly related to the Maximal criteria (MaxS and MaxO) and the Objective Typical dimensional performance criteria (DPO); F (Finger dexterity) is highly related to the Maximal criteria (MaxS and MaxO) and the Subjective Typical dimensional performance criteria (DPS); and M (Manual dexterity) is highly related to the Maximal criteria (MaxS and MaxO) and the Global Performance criterion (GP). With regard to the psychomotor ability scales, it appears that the Maximal criteria are highly related to all of the scales. The Typical criteria are differentially related to the three psychomotor abilities, with Subjective Typical Performance (DPS) showing a relatively high relationship with F and M while Objective Typical Performance (DPO) shows a high relationship with K and a low relationship with both F and M.

The findings in Table 5 are now restated in terms of H6(a) through H6(c); refer to Figure 2, Chapter 3 for the original hypotheses.

H6(a) is Partially Supported: As seen in Table 5, the correlations between the Maximal Subjective criterion and five of the postulated predictors are statistically significant; the correlation between the Maximal Subjective criterion and Verbal aptitude (V) is relatively high but not significant. In contrast, none of the correlations between the Objective Maximal criterion and

the ability scores are significant; in fact, there is a negative relationship between MaxO and Verbal Aptitude.

H6(b) is Not Supported: As seen in Table 5, none of the correlations between the ability scales and either of the Typical job performance criteria are significant. The pattern of correlations for both of the typical criteria with the three cognitive ability scales is similar. However, the correlational pattern for these criteria with the three psychomotor ability scales is the opposite: K (coordination) is highly related to DPO but not DPS, and both F and M (finger and manual dexterity) are highly related to DPS but not DPO.

H6(c) is Partially Supported: As seen in Table 5, the correlations between Global Performance (GP) and only two of the six ability predictors are significant. The relationships between GP with both G and N (intelligence and numerical aptitude) are similar and statistically significant; the relationships between GP with V (verbal aptitude) and M (manual dexterity) are high but not significant; and the relationships between GP with K (coordination) and F (finger dexterity) are low.

Criterion Predictability

Overall, the correlations between ability scales and job performance criteria are not significant nor meaningful with respect to prediction. In order to examine the predictability of the alternative job performance criteria, regression analyses were performed that used multiple predictors rather than simple bivariate correlations. In this section, each of the criteria were regressed on the three cognitive ability scales (G, V, and N), and these "models" were compared with regard to the quality of fit (R-Square and Coefficient of Variation) and the quality of prediction (same). In addition, the regression coefficients for each of the cognitive scales were examined for significance. This same procedure was conducted for the three psychomotor ability scales. The outcome of this analysis is a more detailed examination of the prediction capabilities of each predictor-criterion model.

TABLE 6
Relative Predictability of Job Performance Criteria
Using COGNITIVE ABILITY Tests

Summary of statistical qualities of Fit and Prediction, using the alternative job performance criteria.

Model:	GP & GVN	DPS & GVN	DPO & GVN	MaxS & GVN	Max0 & GVN
Overall Significance:	.0190**	.0295**	.0770*	.0400**	.0465**
Quality of Fit:					
R^2	.08	.08	.06	.07	.07
Coefficient of Variation	28.38%	47.81%	31.15%	23.95%	26.19%
Quality of Prediction:					
R^2p	.02	.00	.00	.00	.00
CV of Prediction	28.85%	48.69%	31.63%	24.37%	26.66%
Model Characteristics:					
Average Response (y)	3.03	2.06	$3.40	$4.75	$4.24
Total Model Variation (SST)	90.92	118.58	1,342,561	1,572,325	1,493,406
Fitting Variation (SSE)	83.31	109.58	1,264,399	1,461,771	1,392,475
Prediction Variation (PreSS)	89.42	117.77	1,349,549	1,566,223	1,494,557
Regression Coefficients:					
General Intelligence (G)	0.02	0.02	2.84*	1.08	2.45
Verbal Aptitude (V)	-0.01	-0.04**	-3.82**	-1.13	-4.18**
Numerical Aptitude (N)	0.01	0.01	0.46	1.86	1.07

Note: Sample size is 117
*denotes significance @ $p < .10$
**denotes significance @ $p < .05$
Regression coefficient significance based on partial F-tests.

In Table 6, the differential impact of ***cognitive ability*** predictors on each of the job performance criteria was examined. Overall, the scales are significantly predictive of all of the job performance criteria. In terms of quality of fit, these multiple linear regression models produced a significant R-Square; the scales explain 6-8% of the variance in performance. Upon internal cross-validation, the proportion of performance variation that is "predictable" by cognitive ability drops substantially and, in all cases but one, is ***zero.*** This zero R-Square of Prediction means that the sum of squared prediction errors is approximately equal to the total variation in the performance model. In fact, the linear model for Global Performance is the only model that produces a positive R-Square of Prediction (2%). Furthermore, an examination of the regression coefficients reveals that Verbal Aptitude is ***negatively*** related to each of the criteria, and in some cases this regression slope is statistically significant. The partial F-tests on the three regression coefficients indicate that none of the coefficients are statistically significant (in the presence of the other variables) with regard to prediction of either Global Performance or Subjective Maximal dimensional performance. The Verbal aptitude scale produces a significant negative relationship with respect to Typical referents and the Objective Maximal referent, and the Intelligence (G) scale is significant with respect to the Typical Objective referent.

In Table 7, ***the psychomotor ability*** scales are postulated as the relevant performance predictors. Overall, the multiple predictor models are not statistically significant; the only significant relationship is represented by Subjective Maximal performance (MaxS) as a function of K, F, and M. In terms of quality of fit, the psychomotor ability scales explain only 1% of the variation in Typical dimensional job performance, only 2% in Global Performance, and only 3% in Objective Maximal performance (MaxO). However, four of these predictor-criterion relationships produce a negative R-Square of Prediction. This "negative predictability" is due to the fact that, upon internal cross-validation, the sum of squared prediction errors was actually greater than the total variation in the performance model. The R-Square of Prediction for the Typical dimensional criteria is -.05, for GP is -.04, and for Objective Maximal criterion (MaxO) is -.03. The only positive R-Square of Prediction is with the Subjective Maximal performance model (+ .02). It should be noted that the

TABLE 7
Relative Predictability of Job Performance Criteria
Using PSYCHOMOTOR ABILITY Tests

Summary of statistical qualities of Fit and Prediction, using the alternative job performance criteria.

Model:	GP & KFM	DPS & KFM	DPO & KFM	MaxS & KFM	Max0 & KFM
Overall Significance:	.4901	.6624	.6932	.0181**	.3214
Quality of Fit:					
R^2	.02	.01	.01	.08	.03
Coefficient of Variation	29.33%	49.38%	31.89%	23.77%	26.70%
Quality of Prediction:					
R^2p	-.04	-.05	-.05	+.02	-.03
CV of Prediction	29.71%	50.0%	32.32	24.12	27.03
Model Characteristics:					
Average Response (y)	3.03	2.06	$3.40	$4.75	$4.24
Total Model Variation (SST)	90.92	118.58	1,342,561	1,572,325	1,493,406
Fitting Variation (SSE)	88.99	116.92	1,325,283	1,439,139	1,448,121
Prediction Variation (PreSS)	94.83	124.95	1,409,855	1,534,465	1,536,613
Regression Coefficients:					
Coordination (K)	.0004	.0003	0.62	1.01	0.71
Finger Dexterity (F)	.0020	.003	0.11	1.29**	0.26
Manual Dexterity (M)	.0060	.003	0.10	0.01	0.39

Note: Sample size is 117
*denotes significance @ $p < .10$
**denotes significance @ $p < .05$
Regression coefficient significance based on partial F-tests.

R-Square statistic drops from a +8% for "explained variance" to a +2% for "predictable variance" for the MaxS model. An examination of the regression slopes indicates that the only significant regression coefficient with regard to Subjective Maximal performance is Finger dexterity (F). The partial F-tests for the other predictor-criterion models indicate that none of the psychomotor ability scales are statistically significant in the presence of the other scales.

Overall, Tables 6 and 7 indicate that (1) there is a significant linear trend between performance and cognitive ability, and (2) there is no discernible differential predictability of the postulated job performance criteria.

An attempt was made to determine the "best" predictor-criterion model for each criterion. A statistical procedure was employed to examine all possible regression models and choose the single or multiple ability scales that provided the highest degree of both fit and prediction. The result of the search for the "best" model is found in Table 8. Each of the postulated models contains both cognitive and ability scales, with the specific predictor scales differing with regard to the specific performance criterion variable.

A thorough examination of Table 8 indicates that the quality of fit, as measured by R-Square, ranges from 8% to 13%. Upon cross-validation, the proportion of model variation that is "predictable" by the relevant set of ability predictors has declined, with R-Square of Prediction ranging from 2% to 7%. Although this statistic is higher than that of the foregoing models (Tables 6 and 7), the cross-validation analysis reveals a low predictability of each criterion. Although each of the multiple predictor models signifies a significant linear relationship, the differential impact of the separate ability scales adds further insight into the nature of these predictor-criterion models. The Global Performance criterion, both of the Typical dimensional performance referents, and the Objective Maximal criterion (MaxO) models all contain the Verbal aptitude scale; yet this scale is negatively related to each criterion and, in most cases, shows a significant negative regression coefficient. Furthermore, only the Objective criteria are significantly related to all three of the relevant predictor scales; in fact, the GP model contains only one significant regression coefficient.

TABLE 8:
Relative Predictability of Job Performance Criteria
Using the "BEST" Combination of Ability Tests

Summary of statistical qualities of Fit and Prediction, using the alternative job performance criteria.

Model:	GP & GVM	DPS & GVM	DPO & GVK	MaxO & GVK	MaxS & NFK
Overall Significance:	.0167**	.0198**	.0243**	.0088**	.0017**
Quality of Fit:					
R^2	0.09	.08	.08	.10	.13
Coefficient of Variation	28.34%	47.62%	30.8%	25.76%	23.22%
Quality of Prediction:					
R^2p	0.02	0.02	.02	.04	.07
CV of Prediction	28.75%	48.36%	31.22%	26.17%	23.59%
Model Characteristics:					
Average Response (y)	3.03	2.06	$3.40	$4.24	$4.75
Total Model Variation (SST)	90.92	118.58	1,342,561	1,493,406	1,572,325
Fitting Variation (SSE)	83.09	108.73	1,235,962	1,347,802	1,374,144
Prediction Variation (PreSS)	88.77	116.14	1,315,113	1,440,396	1,467,409
Regression Coefficients:					
General Intelligence (G)	0.03**	0.03**	3.48**	3.73**	---
Verbal Aptitude (V)	-0.02	-0.04**	-4.43**	-5.05**	---
Numerical Aptitude (N)	---	---	---	---	1.70**
Coordination (K)	---	---	1.01*	1.37**	0.78
Finger Dexterity (F)	---	---	---	---	1.16**
Manual Dexterity (M)	0.005	0.005	---	---	---

Note: Sample size is 117

*denotes significance @ $p < .10$

**denotes significance @ $p < .05$

Regression coefficient significance based on partial F-tests.

Based on Table 8, two general conclusions are drawn. First, the postulated predictor-criterion models are significant, with relatively high R-Square statistics (as compared to Tables 6 and 7). However, there is a 6% drop in R-Square for each model when the internal cross-validation analysis is considered. Second, based on both fit and prediction, the Maximal criteria possess the highest degree of predictability. However, the MaxO and MaxS models are composed of different ability scale predictors.

Based on the findings in Table 8, H7(a) through H7(c) are now discussed (refer to Figure 2, Chapter 3).

H7(a) is Supported: The Maximal dimensional job performance criteria are more predictable than the Typical referents. Recall that this hypothesis was based on the argument that maximal performance is more congruent with individual abilities than typical performance; the typical referent is a function of abilities, motivation, and opportunity bias. For this sample, the highest degree of fit and prediction is attained with the maximal referents. It should be noted that the Subjective Maximal referent (MaxS) is significantly predictable with N, K, and F, or a combination of ability such that psychomotor ability is more important. In contrast, the Objective Maximal referent (MaxO) is significantly predictable by G, V, and K, or a model in which cognitive ability is more important.

H7(b) is Supported: The Maximal dimensional job performance criteria are more predictable that the Global Performance criterion. Both of the R-Square statistics are higher for the Maximal models, and the Coefficient of Variation ratios are lower for the Maximal models. The Coefficient of Variation ratio suggests that the GP model contains a higher amount of "noise," measured relative to the mean Global performance response. It is interesting to note that all three of these predictor-criterion models suggest that a different set of predictors is required for performance prediction, thus the prediction of performance is in fact dependent on the nature of job performance measurement, both level and method.

H7(c) is Not Supported: The Typical dimensional job performance criteria are not more predictable that the Global criterion. The difference in the fitted R-Square is only 1%, and there is no difference in the R-Square of Prediction. A model comparison shows that both G and V are contained in all three of these models, G consistently shows a significant regression coefficient, and V produces a negative relationship/slope, although it is not significant with regard to the GP model.

Summary for Predictive Validation

Overall, the simple bivariate correlations between criteria and ability scores indicate that only two of the job performance measures are significantly related to the single predictor instruments. While GP and MaxS reveal statistically significant correlations with the separate ability scales, the qualities of fit and prediction are enhanced by multiple predictor models using a combination of both cognitive and psychomotor abilities. Furthermore, the job performance criteria that were not significantly related with the single scales are in fact significantly predictable by a combination of ability scales.

With regard to relative predictability, several conclusions are noted. First, use of either cognitive or psychomotor ability, without accounting for the combined effects of both abilities, would suggest that cognitive ability factors are significantly related to all of the job performance criteria, indicating that there is a low distinction between the criteria in terms of prediction. As shown in Table 6, the cognitive ability scales (G, V, N) "explain" approximately 7% of the total variation in job performance, yet these scales cannot "predict" any of the variation in performance. This lack of predictability would seem to indicate that while cognitive ability alone appears to explain the variation in performance, it is not an adequate predictor of performance variation. As a result, the possibility of differences in predictability based on performance measurement cannot be determined based on an examination of either cognitive or psychomotor ability without allowing for the joint effects of these ability scales.

Second, the "best" predictor model for each performance criterion is composed of both cognitive and psychomotor ability scales, thus yielding a higher degree of fit and prediction. The relative predictability suggests that Maximal dimensional performance referents are the most predictable, followed by Global performance and then Typical referents. Although the original hypotheses regarding the relative predictability of alternative job performance measures stated that Typical dimensional performance would be more accurately predictable than Global performance, the findings herein suggest an alternative explanation. While both Global and Typical performance criteria were argued to be a function of ability, motivation, and constraints, the Global criterion measure may implicitly account for some degree of "potential" performance. In this case, GP could be more closely aligned with a notion of feasible/attainable performance than what was originally postulated.

Last, the measurement of job performance criteria does in fact have an impact on the pursuant predictability. With regard to the level/referent of measurement, Table 8 suggests that Maximal performance is more predictable that either Typical or Global performance. Furthermore, the method of measurement has an impact on prediction: the MaxO model contains different ability scales than the MaxS model, and the DPS model contains different scales than the DPO model.

Collectively, these findings suggest that:

1. Job performance criteria are more accurately explained and predicted by multiple predictors;
2. A high degree of explanatory power (R-Square) is not synonymous with a high degree of predictive power (R-Square of Prediction);
3. The alternative performance measures are representative of different constructs for job performance and, as such, require different predictors (as evidenced in Table 8); and
4. The level and method of job performance measurement represent boundary conditions for pursuant performance prediction models.

Summary of Findings

The findings in Tables 1-8 are summarized below.

FIGURE 3
Summary of Findings

Relationships Between Performance Levels

Hypothesis	*Finding*	*Table*
H1(a): DPO-GP	Support	1
H1(b): DPS-GP	Support	1
H1(c): DPS-GP	Support	2

Relationships Within Performance Levels

Hypothesis	*Finding*	*Table*
H2(a): DPS-DPO	Support	1
H2(b): DPS-DPO > DPS-GP	Support	2

Relationships Between Performance Referents

Hypothesis	*Finding*	*Table*
H3(a): DPS-MaxS	Support	1
H3(b): DPO-MaxO	Support	1
H3(c): DPS-MaxO	Support	1
H3(d): MaxS-DPO	Support	1

Relationships Within Performance Referents

Hypothesis	*Finding*	*Table*
H4(a): DPS-DPO	Support	1
H4(b): MaxS-MaxO	Support	1

Comparisons of Matched Referents v. Common Dimension

Hypothesis	*Finding*	*Table*
H5(a): MaxS-MaxO > DPS-MaxO	No Support	4
H5(b): MaxS-MaxO > DPO-MaxS	Support	4
H5(c): DPO-DPS > DPS-MaxO	No Support	3
H5(d): DPO-DPS > DPO-MaxS	Support	3

Figure 3 continued

Comparisons of Matched Referents v. Common Method

Hypothesis	*Finding*	*Table*
H5(e): MaxS-MaxO > DPO-MaxO	No Support	4
H5(f): MaxS-MaxO > DPS-MaxS	Support	4
H5(g): DPS-DPO > DPO-MaxO	No Support	3
H5(h): DPS-DPO > DPS-MaxS	Support	3

Relationships Between Ability & Performance

Hypothesis	*Finding*	*Table*
H6(a): Abilities-MaxO/MaxS	Partial	5
H6(b): Abilities-DPO/DPS	No Support	5
H6(c): Abilities-GP	Partial	5

Relative Predictability of Predictor-Criterion Models

Hypothesis	*Finding*	*Table*
H7(a): Abilities-MaxO/MaxS > Abilities-DPO/DPS	Support	8
H7(b): Abilities-MaxO/MaxS > Abilities-GP	Support	8
H7(c): Abilities-DPO/DPS > Abilities-GP	No Support	8

Chapter 5: Conclusions and Recommendations

Overview

In the preceding chapter, the empirical findings regarding both criterion validity and predictive validity were presented. At this point, it is beneficial to summarize these findings.

1. The convergent relationships (correlations) among all of the job performance measures are positive and significant;
2. The convergence between matched dimensional levels of measurement is greater than the convergence between dimensional with global levels of measurement;
3. The convergence between matched referents of dimensional performance is relatively strong ***but*** this convergence is overpowered by the convergence between Objective Maximal performance (MaxO) with both of the Typical referents (DPS and DPO);
4. The bivariate relationships (correlations) among all of the job performance criteria with ability tests is positive yet mostly non-significant;
5. The relative predictability of performance measures suggests that Maximal referents are "explained" by ability tests better than the other performance measures, and these Maximal referents are also "predicted" by ability tests better than the other criteria. In fact, the relative predictability as gauged by R-Square of Prediction is at least twice as high as that for the GP and typical criteria; and
6. There is a definite measurement impact on both criterion validity and predictive validity findings. With regard to criterion validity, the level of measurement produces a marked effect on the degree of convergence between alternative job performance criteria (i.e., "DP" v. GP-DP). With regard to predictive validity, the referent and the method of measurement produces a marked effect on the predictability of alternative job performance criteria (i.e., the objective criteria are best predicted by different ability scales than the subjective criteria).

In this chapter, the foregoing findings will be further explored and some important issues for future research will be raised. In the first section, conclusions and implications from this study are presented. In the second section, model refinements are proposed regarding the concept of Typical job performance. In the third section, the limitations of this study are noted. The last section offers recommendations for future selection research.

Conclusions and Implications

Based on the results obtained in Chapter 4, several conclusions and research implications regarding the relationships among alternative job performance criteria and predictors are discussed.

Method Effect

First, there is a high degree of convergence among ***multiple methods*** of job performance measurement when the criteria are matched with respect to the level, or specificity of measurement. This finding suggests that rating criteria for job performance do possess a significant relationship with objective performance measures when these criteria are conceptually congruent. Furthermore, this congruence between multiple methods has implications for assessing rating validity, rater accuracy, and predictive validity. With regard to rating validity, the model developed and tested here provides a guide for developing conceptual job performance criteria that are relevant to the decision at hand. By focusing attention on the question of "what to measure for performance," performance ratings take on more meaning when viewed in the context of the different levels, and referents, for measurement. Whereas most selection and validation research has relied on a single measure of overall job performance, a more meaningful approach would be to examine rating criteria on a dimensional level in order to better understand what kind of performance is being measured. Bernardin and Beatty (1984) and James (1973) have advocated such an approach, and this research provides empirical support for their construct validity recommendations. These dimensional performance ratings should be studied individually to determine: (1) their relationship with overall performance criteria, (2) their relationship with other dimensional performance criteria, and (3) their relationship with multiple predictors.

With regard to rater accuracy, the Job Performance Model provides a foundation by which the accuracy of raters can be examined via the convergence between matched levels of measurement.

Typically, rater accuracy has been studied with experimental designs using some kind of "true score" as the criterion for the rating criteria. As discussed in Chapter 2, these studies have been criticized due to the controlled nature of the research and the limited interactions between raters and ratees. Furthermore, the "true score" is determined from "expert opinion," and deviations of rater evaluations from this opinion are assumed to be "rater error" (inaccuracy). Yet the research model developed here provides a means of examining rater accuracy in field settings and avoiding a reliance on "experts" as the benchmark for true scores.

With regard to performance prediction, the Job Performance model identifies different constructs of performance, and these different constructs may or may not be relevant criteria at a given point in time. Once the level of measurement is chosen, different methods of measurement should be included in predictive validation studies. The findings in Chapter 4 suggest that, although dimensional performance criteria show a high degree of convergence, they are not equivalent with respect to specific selection instruments. Whereas most validation research has implicitly assumed that the performance construct is generalizable across different levels and methods of measurement, this study reveals the danger in validating selection procedures against one measure of performance and applying that finding to a different concept of performance.

Level Effect

There is a marked effect of the ***level of measurement*** job performance criteria on the magnitude of criteria convergence. The findings in Tables 1-4 suggest that there is a clear distinction between Global performance criteria and Dimensional performance criteria, although the distinction between different referents of dimensional performance is somewhat unclear. The implications of this level distinction relate to the establishment of boundary conditions for selection research and hypothesis development for selection decisions. With regard to boundary conditions, the criteria validation findings in Chapter 4 indicate that: alternative levels of job performance measurement represent different concepts (constructs) of job performance; these different levels of job perform-

ance criteria are not equivalent; and these different levels of job performance criteria limit the scope of performance generalization. As noted by Weitz (1961), the criterion parameter presents a potential boundary condition for selection research; i.e., "performance" is not necessarily synonymous with "performance" when different levels of measurement are employed.

With regard to hypothesis development, the predictive validation findings in Chapter 4 indicate that the manipulation of the criterion produces different results regarding the choice of selection procedures and the magnitude of the resultant predictor-criterion relationships. For instance, suppose that, based on the findings in Table 5, the Numerical aptitude scale was chosen as a selection instrument. While this scale explains a relatively high amount of the variation in Global Performance, it does not explain a commensurate amount of variation in production performance. Suppose instead that a selection battery composed of just cognitive ability scales was chosen, based on the high correlations between Global Performance with General Intelligence (G) and Numerical Aptitude (N). Table 6 describes just how erroneous a correlation-based decision can be if the overriding aim is performance prediction. While cognitive ability provides a high degree of fit regarding Global Performance and Maximal performance, it does not provide any prediction capability, as evidenced the the zero R-Square of Prediction. Thus the "manipulation" of the performance criterion has a substantial impact on the predictive validation results.

Referent Effect

The distinction between ***different referents*** for dimensional performance measurement is not as clear as the distinction between different levels of performance measurement (refer to Tables 3 and 4). However, there is a significant effect of the performance referent on the relative predictability of job performance. Maximal criteria are more predictable than either Typical or Global performance constructs. One implication of this finding relates to relevant selection goals. The use of different referents for dimensional performance refers to different expectations for performance. In the context of selection, predictor instruments are ideally chosen based on their empirical re-

lationship with on-the-job performance (i.e., criterion-related validation). If future on-the-job performance is defined as maximally feasible performance, the implication is that selection decisions are to be based on "can do" levels of performance--an individual's potential level of achievable performance. In contrast, a Typical criterion used for validation incorporates external constraints on the job that might not be constant across individuals or across time. Therefore, typical performance, at any point in time, might be more highly related to situational and motivational constraints than it is to individual abilities. The result is that by examining maximal performance referents, and incorporating such referents into validation research, selection decisions can be based on individual factors rather than on situational factors.

Model Refinements

In Chapter 4, the hypotheses regarding the degree of convergence between matched referents of dimensional performance were not upheld. The relatively high--and unexpected--convergence between Objective Maximal performance (MaxO) and both of the Typical referents for dimensional performance suggests that future applications of this model should re-evaluate the "typical" referent of job performance. Recall from Chapter 2 that the "typical" referent for dimensional job performance is synonymous with the usual level of performance, thus the most frequently attained performance levels. In contrast, the "maximal" referent for dimensional job performance represents a different expectation for performance evaluation; maximal performance refers to feasible performance levels that may not be attainable on a consistent basis. However, the findings of this study reveal that there is a marked influence of maximal job performance, measured objectively, on typical job performance. According to the Job Performance Model, these findings would indicate that supervisors' evaluations of typical job performance are not based solely on a concept of frequency of attainment.

The relationship between Subjective Typical with Objective Maximal criteria (DPS and MaxO) was found to be stronger than either of the matched-referent relationships, thus suggesting that peak production levels, measured objectively, produce a relatively strong influence on supervisor ratings of typical production output (DPS). The convergence between matched Maximal job performance criteria is the strongest of those that involve Subjective Maximal referents; i.e., the correlation between MaxS with MaxO is 0.72, whereas the correlation between MaxS with DPO is 0.69 and MaxS with DPS is 0.56. Yet this matched relationship is overpowered by the influence of MaxO on both of the Typical dimensional criteria. These findings--for this sample of employees--would seem to indicate that:

1. Supervisory judgments of "typical" production performance levels are more highly related to peak/attainable production levels than they are to modal/average production levels;
2. Supervisory judgments of "maximal" production performance levels are more highly related to peak/attainable production levels than they are to modal/average production levels; yet
3. Supervisory judgments of "typical" production performance levels are influenced by peak/attainable production levels more so than judgments of "maximal" production performance are influenced by the peak production levels.

One possible explanation for the high convergence between a subjective typical rating of performance and the peak performance levels is performance-related: a "recency effect" might be operating such that supervisors recall the peak production levels of performance more so than the modal/average levels. There is some indirect evidence that this recency effect might be causing the relatively high DPS-MaxO convergence when compared with the DPS-DPO convergence. Recall that, on the average, the modal (typical) production levels are established during the fifth week on the job, and that these modal levels are not necessarily maintained on consecutive weeks. In contrast, the peak production levels are observed during week 17 (on the average). A further examination of the data reveals that only 3% of the employee sample (4 out of 153) achieved their peak production level by the 8th week on the job, 50% had achieved their peak production by week 17, 20% achieved their peak production level between weeks 17 and 19, and approximately 20% of the operators achieved their "peak" production during week 20, indicating that these employees are still on the learning curve (i.e., their output levels are still increasing). In contrast, 50% of the sample

had established their production mode by week 4, and 100% had established this modal level by week 12. The point here is that while the modal production level refers to the highest frequency of attainment, the peak production level might refer to the more recent level of observed performance, thus producing a marked influence on production ratings.

Conceptually, a modal level of performance (DPO) refers to the most frequently attained performance level; however, on a practical basis, the mode is difficult to measure. In some instances, individuals have bimodal, or multi-modal performance distributions, thus preventing the measurement of a unique mode. In this study, these bi- or multi-modal levels were averaged. This averaging procedure might have obscured the notion of typical in that the average mode might not be an observed production level. A second difficulty with the measurement of the mode is related to the time at which the mode was observed. In this study, it was found that the modal production level for the operators was established early, approximately the fifth week on the job. However, this does not mean that the modal level was maintained throughout the remaining 15 weeks. Because the modal production level might not be observed on consecutive weeks, supervisors might not be aware of the "typical" production levels. For example, an operator might achieve a production level of, say, $2.50 for week five on the job. For the next 15 weeks, her production fluctuates above and below $2.50 such that she produces at $2.50 on weeks 7, 9, 10, 12, and 15. The measurement of typical used in this study is based purely on the frequency of attainment of a production level and not on the consistency of production performance. Because this modal level was not achieved on a consistent/systematic basis, the supervisors may use their own sort of implicit averaging scheme that is not the same measurement procedure used here.

Based on these issues regarding "typical" job performance, a refinement to the model is suggested here. The use of a mode for assessing typical job performance (objectively) is based on the frequency of performance level attainment yet neglects the consistency of performance attainment. Therefore, a better measure for the typical referent for performance, measured objectively, might be a "restricted mode." One possible form of a restricted mode is a modal performance level restricted to a shorter, more recent period of time. The rationale here is that perhaps a supervisor's

rating of Typical dimensional performance is restricted to a smaller time interval than the explicit appraisal period. Another form of a restricted mode might be a modal performance level restricted to consecutive time periods. The rationale here is that a more relevant mode might encompass the notion of consistency in conjunction with frequency. Note that these refinements are focused on the measurement of the objective typical performance criterion in order to better understand what the "typical" performance rating is capturing. The implications of this refinement to dimensional (objective) performance are that the Typical referent convergence can be re-examined for a better conceptual fit, and the relationship between supervisory ratings and congruent, observable performance standards can be better understood.

Study Limitations

There are four aspects of this study that present limitations insofar as the generality of this research is concerned: sample characteristics, measurement procedures, criteria availability, and predictor instruments.

The ***sample characteristics*** restrict the findings herein to a relatively small focal population. First, all of the employees contained in the sample are females; thus, the generality of these results to a working population of both males and females is unknown. Second, the job under study is similar to that of an autonomous job design; the sewing machine operators set their own work pace and have a minimal degree of work group interaction. As a result, the research model test herein is applicable to those jobs in the population in which the work group presents minimal constraints on performance, constraints in the form of work pace and/or production norms. Third, the piecerate incentive system produces a competitive environment in which the quantity of production represents the key to job success. A potential problem is that these results might be limited to a narrow concept of performance as opposed to a more enlarged concept of performance.

It should be noted that although these sample characteristics represent limitations on the one hand, they also represent naturally occurring controls. Although the sample of females-only might limit the findings, this feature has ruled out the alternative explanation that these findings might be moderated by gender differences. The autonomous nature of the job provides only a partial test of the model, yet this feature has served the purpose of examining individual performance and not work-group dominated performance. The competitive environment might limit these findings to a narrow realm of performance, yet this feature has focused on the explicit and primary criterion that links individual performance with management goals.

The ***measurement procedures*** produce two technical limitations on the findings. First, the reliability of these job performance criteria was not estimated, and the reliance on single-item ratings restricts these findings to the specific organization under study. Second, the measurement of the objective typical criterion might have restricted the concept of typical performance; the use of a mode might not be feasible in many research, or industrial settings. Although the use of modal performance levels might be limited to certain settings, this measurement of an average is more congruent with the concept of typical than a mean or median.

The use of ***objective and subjective*** performance criteria restricts the use of this model to situations in which both criteria are available. First, this feature appears to limit the generalization of these findings to more production-oriented jobs as opposed to managerial-type jobs. However, the basis for this model is a general job performance model that can, and should be applied to a range of jobs (James, 1973). A basic criticism of selection research is the reliance on performance ratings, and this model serves as a guide for examining the validity of these ratings using conceptually relevant objective data. Furthermore, this model focuses attention on the dimensional level of performance as the primary criterion of interest, thus reducing the importance of overall job effectiveness. By highlighting the importance of individual results, the focus becomes the "link" between organizational goals and individual behaviors.

Second, the requirement for multiple criteria has restricted this study to an examination of only one dimension of performance when in fact job performance is multidimensional. The quantity of production was chosen for investigation because it was a primary criterion for job success (according to management) and it was quantifiable. In the future, multiple performance dimensions, identified through job analysis, should be examined. However, many dimensions of job success are not quantifiable, and thus this model is in fact limited to only certain dimensions of performance. Yet two caveats are warranted here. First, performance standards should be explicit for any job in order to provide an explicit road map for job success. Although not all performance dimensions will have quantifiable performance standards, these dimensions can/should be linked with performance goals. These goals provide the means for developing "objective" criteria for ratings. Second, the investigation of criteria convergence for some performance dimensions can add credibility to the overall performance rating process. Consider that if there is a high degree of convergence between ratings and objective standards on some of the dimensions of job performance, the "reliability" of raters is enhanced (Bernardin and Beatty, 1984).

The ***choice of predictors*** in this study might have caused the relatively low predictive validation findings. Based on the job analysis and criteria development, the use of cognitive and psychomotor abilities appeared to be the most relevant for this job. However, additional predictors such as skills tests and/or experience might have produced more significant predictor-criterion results.

Future Research

One of the most important recommendations for future selection research is for a more thorough conceptualization and investigation of the performance criterion parameter. Many authors have stressed the importance of criterion research (i.e., Weitz, 1961; Smith, 1976; James, 1973; Guion, 1976), and this empirical study emphasizes the effect of criterion measurement on an understanding of performance. The findings herein suggest that (for this sample) the level of measurement limits

the scope of the performance construct, represents a boundary condition on the generalization of predictor-criterion findings, and identifies more conceptually relevant hypotheses for selection research.

At this point, it is interesting to recall the previous discussion of boundary conditions. Weitz (1961) "warned" that different types/levels of criterion measurement may in fact limit the usefulness (and generalizability) of our models/theories, that a manipulation of the criterion measurement may produce different results with regard to model/theory testing, and that attention should be paid to possible boundary conditions that take the form of criterion measurement. The criterion parameter itself may represent a boundary condition on the one hand, and the avenue for "cleaner hypotheses" on the other hand. These findings, taken as a whole, are in agreement with the statements made by Weitz and suggest that selection research should devote more attention to the nature of job performance measurement.

Second, future selection research should apply the Job Performance model suggested by James (1973) and developed herein. These investigations might be replications and/or extensions. Replications of this study will provide further evidence about the validity of the model. Such studies should include the refinement to typical performance discussed previously and the study of different occupational groups of employees. Extensions to the model should take the form of longitudinal studies that examine the criteria relationships over time and over multiple job performance dimensions.

With regard to **criteria validation,** several recommendations for future research are offered. First, potential moderators on criteria convergences should be examined. Factors such as different supervisors, organizational characteristics, and ratee characteristics will likely produce a pronounced effect on criteria validation findings. For example, selection research has relied on ratings of performance and has collapsed these ratings across different supervisors, assuming that different supervisors' evaluations of performance are consistent across the organization. Yet different supervisors evaluate job performance using different expectations (Borman, 1974), thus yielding

situationally-specific performance measures (i.e., Guion, 1976). Furthermore, organizational moderators might account for the span of control of each supervisor and the method of job analysis employed for criterion development. Ratee characteristics such as race and gender should also be examined for moderating effects. Henemen (1986) has suggested that differential treatment of the ratee by the rater might constitute a moderator on criteria convergent validity. He cited a study by Bass and Turner (1973) in which these authors found significant convergent validity for black but not white ratees.

Second, performance consistency over time should be incorporated into the Job Performance Model. While this research was focused on the convergence of criteria at one point in time, future research should examine these convergences at different points in time (i.e., early performance, training performance, "stabilized" performance). Ghiselli (1956) suggests that intercorrelations among criterion measurements taken at different points in time should be examined in order to ascertain the kind of pattern that holds. This study suggests that alternative criterion measures, using the model here, be used for this investigation, thus revealing the extent to which criterion specificity/dimensionality changes with time. Furthermore, the predictive validites of these alternative criteria can be examined over time, thus providing information about whether predictions of job performance refer to success early or late in employment.

Third, the concept of performance distributions should receive more attention as an informative means of understanding and measuring performance. Kane's Performance Distribution Assessment method (1980) emphasizes the need to examine additional aspects of the performance distribution, in addition to the mean level of performance. Factors such as the standard deviation of performance and negative (sub-standard) incidences of performance yield information about not only performance but also effort and situational constraints. When examined over time, the performance distribution analysis, applied to both subjective and objective measures of job performance, provides important information about performance consistency and method convergence.

With regard to **predictive validation,** there are at least three major areas of research which need to be explored. First, alternative measures for job performance need to be incorporated into predictive validation studies. These alternative measure should take the form of different levels and different methods of measurement (i.e., James, 1973; Dunnette, 1963). This study shows the fallacy of assuming that predictive validation findings are generalizable across different concepts of job performance.

Second, as with criteria validations, potential moderators need to be incorporated into selection models. One example of a predictive validation moderator is situational moderators such as the job analysis method, job performance measurement system, and different supervisors. This research is imperative given the need to examine boundary conditions on selection models (Weitz, 1961) and the simultaneous attempts to generalize validation findings (i.e., Schmidt, Hunter, and Pearlman, 1981). Other examples of predictor-criterion moderators include task complexity differences (i.e., Terborg, 1977; Fleishman, 1979; Schmidt, Hunter, and Pearlman, 1981) and employee experience differences (Schmidt, Hunter, and Outerbridge, 1986).

Third, the methodology of predictive validation needs to be extended beyond the simple correlational analyses to incorporate prediction errors. This study examined prediction in terms of internal cross-validation. Future research should continue to focus on the cross-validation of findings using the PreSS statistic and/or data splitting. Although the need for cross-validation of findings is not new to selection research, the use of cross-validation has largely been prevented due to sample size requirements. The PreSS statistic allows cross-validation without reducing the sample size and provides additional measures for gauging predictive power. Furthermore, multiple regression models should be used in order to incorporate the potential moderators on predictor-criterion relationships (i.e., Dunnette, 1963).

End Note

It is my experience that one of the overriding objectives of human resource management research is to answer the "so what" questions. As a means of concluding this research project, I feel it is instrumental to not only pose but also attempt to answer some of the more common "so what" questions. The first question posed is:

So What are the practical implications of **different levels** of performance measurement? A cursory answer is that these findings suggest that "performance is not performance is not performance." Dimensional job performance is not equivalent to global performance; the recognition of levels of measurement provides management with guidelines regarding what to measure for job performance and what has been measured as job performance.

In terms of performance appraisal, this model would ideally account for multiple dimensions for job performance, and it explicitly links these identified dimensions with observed (concrete) performance standards. As a result, the degree to which overall job performance captures each of the multiple dimensions can be examined, the relative importance of each dimension for overall job success can be determined, and the degree of independence of the multiple factors of job success can be investigated. Furthermore, rating validity can be examined on a dimensional (incremental) level thus providing information about the accuracy of both ratings and raters. In terms of prediction, this model identifies a variety of "conceptual criteria" for job performance. At one point in time, different levels of dimensional job performance might be more important than others, and this model shows that these levels and dimensions are not substitutable.

So What are the practical implications of **different referents** for job performance measurement? A rhetorical question is "Does effort have an impact on performance?" A concept of maximal performance assumes that the variation in performance is due to effort whereas a typical referent assumes that this variation is random.

In terms of performance appraisal, the distinction between maximal and typical performance levels provides management with a means for identifying the impact of situational constraints on performance and recognizing employees' for the amount of effort put into the job. In addition, the maximal frame of reference for performance evaluation focuses on the distribution of performance over time, thus providing management with information about individual performance consistency and operating unit performance consistency. This performance distribution analysis provides information about the standard deviation of performance over time. In terms of prediction, the use of different referents distinguishes "feasible" from typical performance standards and explicitly defines the performance criterion parameters of interest at a given time and/or location. The result is that validation evidence can be gathered using the different frames of reference for performance evaluation. This use of dual referents is particularly important if performance (situational) constraints are not deemed to be constant over time, locations, and/or employee groups.

Appendix A. The Performance Appraisal Instrument

List of Performance Factors

Based on an analysis of the job of Sewing Machine Operator, the following five factors were identified as the most important components of overall job success.

1. Quantity of Work: the typical, or average output of the operator over a period of several weeks.
2. Quality of Work: the overall quality of production and the frequency that repair work (re-work) is required.
3. Flexibility: the ability of operators to adjust to changes in models and/or sewing operations.
4. Receptiveness to Training/Instruction: the responsiveness of the operators to abiding by procedures and instructions.
5. Dependability: attendance and punctuality.

SUGGESTIONS TO RATERS

We are asking you to rate the job performance of the people who work under your supervision. These ratings will serve as a "yardstick" against which we can compare the test scores. If the ratings do not reflect your best and most honest judgments, this study will have little value. Please try to give the most accurate ratings possible for each worker.

These ratings are strictly confidential and will not affect your employees in any way. Neither your employees nor any member of management will ever see this rating. We are only interested in "testing the tests." Therefore, ratings are needed only for those workers who were hired during 1986, and these ratings will be available only to the Va. Tech research group.

You are being asked to rate each worker on 5 aspects of work performance. Try not to let general impressions or some outstanding trait affect your judgment. Read each of the rating scales thoroughly before rating, and rate only on that factor. Rate the workers according to the work they have done over a period of several weeks or months. Think in terms of each worker's usual, or typical performance or behavior.

The rating form is designed so that you can rate all of your employees on one scale at a time. For example, first, you will rate all of your employees on the quantity of work performed. Next, you will rate your employees on the quality of work performed. This will make it easier for you to fairly and consistently apply the rating scales to all of the employees.

Thank you for your cooperation; it is extremely important to the success of this project.

Quantity of Work

Consider the typical, or average output of the sewing machine operator over a period of several weeks. For each employee, write-in the number (1 to 6) of the statement below that best describes the employees' quantity of work.

1 = Production is at or below minimum standard

2 = Production is above minimum standard ***but*** less than the established production goal

3 = Production is at the established production goal

4 = Production is above the established production goal

5 = Production is among the best in the plant

6 = Unable to rate this employee because employee has not been under my supervision for a long enough period of time.

Quality of Work

Consider the overall quality of production and how often repairs are required.

1 = Production quality is often below minimum quality standards (often requires rework)

2 = Production quality is usually acceptable ***but*** somewhat inferior

3 = Production quality is usually acceptable ***but*** not superior

4 = Production quality is usually superior

5 = Production quality is almost always among the best in the plant

6 = Unable to rate this employee because employee has not been under my supervision for a long enough period of time.

Flexibility

Consider how well this operator adjusts to ***changes*** in operations, machines, and/or work assignments.

1 = Very low flexibility--has great difficulty adjusting to any changes

2 = Has some difficulty adjusting to any changes

3 = Adequately adjusts to most changes

4 = Adjusts well to most changes

5 = Very high flexibility--quickly adjusts to any changes

6 = Unable to rate this employee because employee has not been under my supervision for a long enough period of time.

Receptiveness to Training/Instruction

Consider how responsive this operator is to receiving and following sewing procedures and your instructions.

1 = Unacceptable (frequently disregards instructions and procedures)

2 = Minimally Acceptable (sometimes disregards instructions and procedures)

3 = Acceptable (normally follows instructions and procedures)

4 = Good (instructions and procedures seldom have to be repeated)

5 = Superior (always follows instructions and procedures)

6 = Unable to rate this employee because employee has not been under my supervision for a long enough period of time.

Dependability

Consider this operator's absence and tardiness records, and give your assessment/recollection of these records.

1 = Often absent and/or tardy without notification

2 = Often absent and/or tardy ***but*** usually gives prior notification

3 = Has an average attendance/tardiness record and usually gives prior notification

4 = Occasionally absent/tardy ***but*** gives prior notification

5 = Rarely absent or tardy

6 = Unable to rate this employee because employee has not been under my supervision for a long enough period of time.

Overall Performance Rating

Considering all of the factors involved in performing the job of sewing machine operator, how would you rate this employee's overall performance?

1 = Unacceptable (performance clearly fails to meet the minimum job requirements)

2 = Minimally Acceptable (performance occasionally fails to meet the minimum job requirements)

3 = Acceptable (performance meets the minimum job requirements)

4 = Above Average (performance usually exceeds the minimum job requirements)

5 = Superior (fully proficient in all aspects of the job)

6 = Exceptional (outstanding in all aspects of the job)

7 = Unable to rate this employee because employee has not been under my supervision for a long enough period of time.

ADDITIONAL RATING: QUANTITY OF OUTPUT

For each employee, give the percent of time (%) that her average weekly ticket earnings fall into each of the 8 categories listed below. In other words, how often (what % of the time) has each employee produced at each of the 8 earnings categories? This rating should *not* consider the initial training time on the job.

Level 0: Less than $3.00
Level 1: $3.00-$3.45
Level 2: 3.46- 3.90
Level 3: 3.91- 4.30
Level 4: 4.31- 4.74
Level 5: 4.75- 5.60
Level 6: 5.61- 6.49
Level 7: 6.50 and over

LEVEL:	1	2	3	4	5	6	7	
EX. JOHN DOE	0%	10%	30%	30%	25%	5%	0%	= 100%
1. Employee name	___	___	___	___	___	___	___	= 100%
2. Employee name	___	___	___	___	___	___	___	= 100%

Bibliography

Alexander, E.R., and Wilkins, R.D. Performance rating validity: The relationship of objective and subjective measures of performance. *Group and Organization Studies,* 1982, 7, 485-496.

Argyris, C. Problems and new directions for industrial psychology. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology.* Chicago: Rand McNally, 1976, Ch. 5.

Arvey, R.D. *Fairness in Selecting Employees.* Mass: Addison-Wesley Publishing Co., 1979.

Astin, A.W. Criterion-centered research. *Educational and Psychological Measurement,* 1964, 24, 807-822.

Bernardin, H.J. and Beatty, R.W. *Performance Appraisal: Assessing Human Behavior at Work.* Boston, MA: Kent Publ. Co, 1984.

Barrett, G.V., Caldwell, M.S., and Alexander, R.A. The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology,* 1985, 38, 41-56.

Bass, B.M. Further evidence on the dynamic character of criteria. *Personnel Psychology,* 1962, 15, 93-98.

Blum, M.L. and Naylor, J.C. *Industrial Psychology: It's Theoretical and Social Foundations.* NY: Harper & Row, 1968.

Borman, W.C. The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance,* 1974, 12, 105-124.

Borman, W.C. Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology,* 1978, 63, 135-144.

Borman, W.C., Rosse, R., and Abrahams, N. An empirical construct validity approach to studying predictor-job performance linkages. *Journal of Applied Psychology,* 1980, 65, 662-671.

Campbell, D. and Fiske, D. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, 56, 81-105.

Campbell, J.P. Psychometric theory. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally, 1976, Ch. 17.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., and Weick, K.E. *Managerial Behavior, Performance, and Effectiveness*. NY: McGraw-Hill, 1970.

Campbell, J.P. and Pritchard, R.D. Motivation theory in industrial and organizational psychology. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally, 1976, Ch. 3.

Cascio, W.F. *Applied Psychology in Personnel Management*. VA: Reston Publishing Co., 1982.

DeNisi, A.S. and Stevens, G.E. Profiles of performance, performance evaluations, and personnel decisions. *Academy of Management Journal*, 1981, 24, 592-602.

Dreher, G. and Sackett, P. *Perspectives on Employee Staffing and Selection*. Homewood, IL: Richard D. Irwin, 1983.

Dunnette, M.D. A modified model for test validation and selection research (1963). In G. Dreher and P. Sackett (Eds), *Perspectives on Employee Staffing and Selection*. Homewood, IL: Richard D. Irwin, 1983.

Dunnette, M.D. A note of the criterion. *Journal of Applied Psychology*, 1963, 47, 251-254.

Fleishman, E.A. Evaluating physical abilities required by jobs. *Personnel Administrator*, 1979, 24, 82-92.

Ghiselli, E.E. Dimensional problems of criteria. *Journal of Applied Psychology*, 1956, 40, 1-4.

Ghiselli, E.E. The validity of aptitude tests in personnel selection. *Personnel Psychology*, 1973, 26, 461-477.

Ghiselli, E.E. and Haire, M. The validation of selection tests in the light of the dynamic character of criteria. *Personnel Psychology*, 1960, 13, 225-231.

Guion, R.M. Criterion measurement and personnel judgments. *Personnel Psychology*, 1961, 14, 141-149.

Guion, R.M. *Personnel Testing*. NY: McGraw-Hill, 1965.

Guion, R.M. Recruiting, selection, and job replacement. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally, 1976, Ch. 18.

Gunderson, E.K. and Nelson, P.D. Criterion measures for extremely isolated groups. *Personnel Psychology*, 1966, 19, 67-82.

Heneman, R.L. The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology*, 1986, 39, 811-826.

Hunter, J. A causal analysis of cognitive ability, job knowledge, job performance, and supervisory ratings. In F. Landy, S. Zedeck, and J. Cleveland (Eds.), *Performance Measurement Theory*. NJ: Lawrence Erlbaum Associates, 1983a.

Hunter, J. The dimensionality of the general aptitude test battery and the dominance of general factors over specific factors in the prediction of job performance. U.S.E.S. Test Research Report #44, 1983b.

Hunter, J. Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery. U.S.E.S. Test Research Report #45, 1983c.

Hunter, J. A rebuttal of Dr. Novick's false allegations about validity generalization and the validity of general cognitive ability predicting job performance. A paper presented at the First Annual Conference of the Society of Industrial and Organizational Psychology, April 11, 1986.

James, L. Criterion models and construct validity for criteria. *Psychological Bulletin,* 1973, 80, 75-83.

Kane, J.S. Performance distribution assessment: A new framework for conceiving and appraising job performance (1980). In H.J. Bernardin and R.W. Beatty, *Performance Appraisal: Assessing Human Behavior at Work.* Boston, MA: Kent Publishing Co., 1984.

Kane, J.S. Rethinking the problem of measuring performance: Some new conclusions and a new appraisal method to fit them. A paper presented at the Fourth Johns Hopkins University National Symposium on Educational Research, 1982.

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

Landy, F.J. and Farr, J.L. Performance rating. *Psychological Bulletin,* 1980, 87, 72-107.

Landy, F.J. and J.L. Farr. *The Measurement of Work Performance.* Orlando, FL: Academic Press, Inc., 1983.

Lawler, E.E. The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology,* 1967, 51, 369-381.

Myers, R.H. *Classical and modern regression with applications.* Boston: Duxbury Press, 1986.

Nagle, B.F. Criterion development. *Personnel Psychology,* 1953, 6, 271-289.

Nealey, S.M. and Owen, T.W. A multitrait-multimethod analysis of predictors and criteria of nursing performance. *Organizational Behavior and Human Performance,* 1970, 5, 348-365.

Prien, E.P. Dynamic character of criteria: Organizational change. *Journal of Applied Psychology,* 1966, 50, 501-504.

Rambo, W.W., Chomiak, A.M., and Price, J.M. Consistency of performance under stable conditions of work. *Journal of Applied Psychology,* 1983, 68, 78-87.

Seashore, S.E., Indik, B.P., and Georgopoulos, B.S. Relationships among criteria of job performance. *Journal of Applied Psychology,* 1960, 44, 195-202.

Schmidt, F. and J. Hunter. Employment testing: Old theories and new research findings (1981). In G. Dreher and P. Sackett (Eds.), *Perspectives on Employee Staffing and Selection.* Homewood, IL: Richard D. Irwin, 1983.

Schmidt, F., J. Hunter, and A. Outerbridge. The impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology,* in press.

Schmidt, F., J. Hunter, and Pearlman, K. Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology,* 1981, 66, 166-185.

Scott, W.E. and Hamner W.C. The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. *Organizational Behavior and Human Performance,* 1975, 14, 360-370.

Severin, F. The predictability of various kinds of criteria. *Personnel Psychology,* 1952, 5, 93-104.

Smith, P. Behaviors, results, and organization effectiveness: The problem of criteria. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology.* Chicago: Rand McNally, 1976., Ch. 17.

Steel, R.P. and Mento, A.J. Impact of situational constraints on subjective and objective criteria of managerial job performance. *Organizational Behavior and Human Performance,* 1986, 37, 254-265.

Terborg, J.R. Validation and extension of an individual differences model of work performance. *Organizational Behavior and Human Performance,* 1977, 18, 188-216.

Thorndike, R.L. *Personnel Selection.* NY: Wiley, 1949.

Toops, H.A. The criterion. *Educational and Psychological Measurement,* 1944, 4, 271-297.

Tucker, M.F., Cline, V.B., and Schmitt, J.R. Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. *Journal of Applied Psychology,* 1967, 51, 131-138.

Turner, W.W. Dimensions of foreman performance: A factor analysis of criterion measures. *Journal of Applied Psychology,* 1960, 44, 216-223. 5, 348-365.

Wallace, S.R. Criteria for what? *American Psychologist,* 1965, 20, 411-417.

Weitz, J. Criteria for criteria. *American Psychologist,* 1961, 16, 228-231.

Zedeck, S. and Baker, H.T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance,* 1972, 7, 457-466.

The vita has been removed from the scanned document