# AN EMPIRICAL INVESTIGATION OF CRITERION DEVELOPMENT AND THE MULTIPLE CRITERIA VERSUS COMPOSITE CRITERION DEBATE

Patrick A. Dailey, B.A.

APPROVED:

Michael M. Beyerlein

Major Professor

Kurt A. Helm

Committee Member

Douglas A. Johnson

Committee Member

Chairman of the Department of Psychology

Dean of the Graduate School

# AN EMPIRICAL INVESTIGATION OF CRITERION DEVELOPMENT AND THE MULTIPLE CRITERIA VERSUS COMPOSITE CRITERION DEBATE

THESIS

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Patrick A. Dailey, B.A.

Denton, Texas

August, 1990

AB

Dailey, Patrick A., An Empirical Investigation of Criterion Development and the Multiple Criteria Versus Composite Criterion Debate. Master of Science (Industrial/Organizational Psychology), August, 1990, 66 pp., 6 tables, references, 26 titles.

The purpose of this study was to empirically examine two main areas of concern in selecting criteria for validation studies: the development of the criterion and the multiple criteria versus composite criterion debate. Evidence was found for the ability of the various weighting schemes used to generate composites that were statistically and conceptually different from one another. Knowledge of the nature of each composite, along with the multiple criteria approach, proved essential to understanding the composite criterion's relationship to the predictor and the impact of the criterion to the validation process. Selection and treatment of the criterion apparently consist of judgment and individual estimations.

# ACKNOWLEDGEMENT

The author wishes to acknowledge Dr. Kurt Helm for his guidance in this study. Without Dr. Helm's contributions, this thesis would not have been possible.

# TABLE OF CONTENTS

## LIST OF TABLES

# CHAPTER I

## INTRODUCTION

Early in the history of industrial and organizational (I/O) psychology, researchers viewed "criteria as either given of God or just to be found lying about" (Jenkins, 1946, p. 93). Many industrial psychologists now realize that neither of these cases is true. Although studies can still be found which use the most expedient criteria (those measures which are readily and easily available), more and more thought is given to the significance of careful development of the criterion.

While greater consideration, at least in theory, is being given to the significance of the criterion, relatively few studies emphasize the criterion as the main target for research. This study is designed to address the need for an empirical examination of important concerns regarding the use of the criterion in validation research.

Only a small amount of research on the criterion exists from the past decade. A review of the literature revealed that recent articles on the criterion address the same issues which were points of concern during the initial investigation into the criterion. It seems little has changed. The concerns about the criterion which were important in the 1940s and 1950s (Nagle, 1953; Thorndike, 1949; Toops, 1944) are much the same as those in the 1980s and into the 1990s (Guion, 1987).

This study has been conceived in order to call attention to the paucity of current research which investigates the impact of the criterion in validation studies. A noticeable artifact of little recent research investigating the criterion, is that the information for this study relies heavily on the use of literature from early work. There is also extended use of textbooks as sources for summarizing earlier studies on the criterion.

The term criterion has many different definitions (Astin, 1964; Dunnette, 1963; Nagle, 1953). Although there are individual differences in the definition of criterion, "common to each of the definitions is the idea that the criterion represents something important or desirable" (Astin, 1964, p. 808). Criterion development is an attempt to make meaningful and understandable, measures of those things which are considered important or desirable.

The criterion has several purposes within an organization. It is useful for validation, selection, compensation, training, motivation and satisfaction of employees, and feedback, along with other purposes (Landy, 1989). This study will focus on the use of the criterion in personnel selection test validation.

The use of the criterion in the validation process has been subject to much criticism and debate. Many authors (Astin, 1964; Dunnette, 1963; Nagle, 1953; Wallace, 1965; Weitz, 1961; Wherry, 1957) have discussed the various problems inherent in the use of criteria. Important issues which surround the criterion include: (a) the dimensional problem of criteria (Ghiselli, 1956); (b) the dynamic character of criteria (Bass, 1962; Ghiselli & Haire, 1960); (c) the role of the "ultimate" criterion (Astin, 1964; Dunnette, 1963; Nagle, 1953); (d) the difference between the use of composite and multiple criteria (Guion, 1987; Nagle, 1953; Schmidt & Kaplan, 1971; Toops, 1944); and (e) the development and evaluation of the criterion (Nagle, 1953; Toops, 1944; Wherry, 1957).

For the purpose of this study, two main areas are examined: the development and evaluation of the criterion and the composite criterion versus multiple criteria debate. The role of the ultimate and conceptual criterion is discussed initially to assist in an understanding of the theoretical foundation of the criterion.

## Criterion Development

As evident by the wide range of concerns surrounding the criterion, criterion development is an important topic. Inherent in the selection of criteria is choice. Several choices must be made concerning the utilization of criteria in a validation study. Certainly, one of the most important of those choices is which criteria to include. Weitz (1961) emphasizes the significance of choosing criteria by asking ". . . does the choice of the criterion have any effect on the results (or lack of them) . . . . and if we are to evaluate our conclusions, do we not need to understand the effect of choosing a particular criterion?" (p. 228).

How should researchers choose a specific criterion? Many authors (Astin, 1964; James & Ellison, 1978; Nagle, 1953; Thorndike, 1949) suggest the use of a conceptual criterion or an ultimate criterion to serve as a theoretical basis for making decisions about what aspects to include in the criterion or the choice of the criterion itself.

The conceptual criterion represents the lowest level of abstraction in the developmental hierarchy of relevant goals (Astin, 1964). The ultimate criterion can be thought of in a similar manner, except it is a more encompassing and more abstract criterion. The ultimate criterion represents the highest level of abstraction and includes everything that ultimately defines success on the job (Thorndike, 1949). Although defined in a slightly different manner, the ultimate criterion and conceptual criterion are often discussed in the literature as being interchangeable. For the purposes of this study the ultimate and conceptual criterion will be treated as interchangeable concepts.

It is important to realize that neither the conceptual nor the ultimate criterion exist in any empirical terms. Although they are considered intangible, the ultimate and conceptual criterion are beneficial in the conceptualization of the criterion. Both are levels of abstraction used to understand the criterion.

The conceptual and ultimate criterion derive their value by requiring the researcher to understand what the criterion measure is going to represent and the different ways in which the performance criterion can be represented. The two abstract concepts assist in examining the many different, potential measures which can be used to satisfy the purpose of the ultimate criterion in its abstract form.

The ultimate and conceptual criterion are concepts which assist the researcher in including or not including particular measures or certain elements in a criterion. When the researcher has an understanding of what constitutes the criterion space, an empirical criterion measure or performance criterion (Astin, 1964) can be chosen which is representative of the total conceptual criterion.

There are guidelines which have been developed in order to assist the researcher in choosing the most effective measures of the conceptual criterion or determining the effectiveness of empirical measures of the criterion which has already been chosen. The following section examines some of these standards.

The significance of the criterion selection to test validation studies has not gone without remark. Krug (1961) points out that "Clearly a program of personnel selection can be no better than the criteria which define it" (p. 107). Nagle (1953) states "Research can be no better than the criteria used" (p. 273). Perhaps, the most appropriate remark about the significance of the criterion to test validation was made by Toops (1944) early on in the investigation of the impact of the criterion. He commented, "Possibly as much time should be spent in devising the criterion as in constructing and perfecting the tests . . . If the criterion is slighted the time spent on the tests is, by so much, largely wasted" (p. 290). In theory, most researchers agree that the necessity of carefully choosing, developing and evaluating the criteria used in selection validation studies is important. In application, often times, the criteria are considered secondary to the development of the test.

Frequently, after an extensive effort has been put into devising a test, criteria for validation purposes are chosen haphazardly or by means of expediency. Such disregard for the significance of the criteria can lead to research concluding that there is a weak relationship between the predictor and the criteria, even though a strong relationship exists.

When the validity coefficient(s) show a weak relationship to the criteria, the test is often scrutinized for reconsideration, when in actuality the test may have few deficiencies. It is possible that the criteria used in the validation study contained the deficiencies which accounted for the poor correlation. This reality has forced researchers to take a closer look at the relationship of the criterion to the predictor, to examine the criterion on its own, and to specify standards to follow for criterion development.

Researchers invariably cite a lengthy list of requirements for criteria. A thorough list suggested by Blum and Naylor (1968) identifies the following as important standards for criteria:

1. Reliable,
2. Realistic,
3. Representative,
4. Related to other criteria,
5. Acceptable to the job analyst,
6. Acceptable to management,
7. Consistent from one situation to another,
8. Predictable,
9. Inexpensive,
10. Understandable,

11. Measurable,
12. Relevant,
13. Uncontaminated and bias-free, and
14. Discriminatory.

In an additional list, Landy (1989) included concern with freedom from contamination, relevancy, freedom from deficiency, freedom from bias, acceptability by management, cost and predictability. Landy suggested that all of the above considerations are important. He reduced the considerations to three major concerns: reliability, validity, and practicality. Muchinsky (1983) agreed with the summary of concerns but calls the factors stable (reliable), appropriate (valid), and practical. The following sections examine these three factors and their significance to the development of a criterion.

## Reliability

The reliability of a criterion is concerned with the criterion's stability and freedom from unsystematic variance. In validation procedures the reliability of a measure is looked upon as the most concrete of the three requirements. Reliability is favored by psychometricians because, in most cases, a numerical value can be derived. When a reliability coefficient for a criterion can be derived, there is concern for an over-emphasis of the importance of the resulting value. Nagle (1953) comments, "it is not to imply that it [reliability] is not important, but . . . high reliability, while desirable is not sufficient" (p. 276). This suggests that other things, such as validity and practicality, contribute to the effective development of the criterion.

Many different methods exist for empirically estimating the reliability of a measurement (Ghiselli, Campbell, & Zedeck, 1981). The average methods assume that the target of reliability estimation is a test or in the form of a test. Two types of criteria which are most often used as criterion in validation studies are ratings and objective measures. Generally, these types of criteria do not take the form of a test.

The reliability of ratings has received much attention because of their common use as criteria. In order to estimate the reliability of ratings, the ratings must have either different raters rating the performance or multiple ratings by a single rater. If either of these two conditions (or both) exist, empirical methods for determining a reliability coefficient can be utilized. A problem which is often encountered is the use of a single rating on a single occasion. In this case a reliability coefficient cannot be determined mathematically and the ratings must be treated in the same manner as objective measures which are discussed next.

After an extensive review of the literature this author was unable to find any methods for empirically estimating the reliability of objective criterion measures. Examples of objective measures which are often used as criteria are generally in the form of reported performance and include such things as scrap rates, absenteeism, and sales volume. Because these performance criteria are merely recorded and do not take the form of a test, traditional methods for statistically determining reliability are not applicable. Usually, reliability in the traditional sense, is not addressed with regard to objective criterion measures. Instead, objective criterion measures are commonly assumed to have perfect reliability (a statistical reliability coefficient of one).

Caution is advised in assuming measures have perfect reliability. Performance-based criteria and ratings with reliability coefficients that can only be assumed to be one are subject to the same concerns of unreliability as criteria for which reliability coefficients can be mathematically derived. It is plausible that there are sources of unreliability in criteria assumed to be perfect and in fact the true reliability is less than one.

There are many sources for criterion unreliability. Nagle (1953) identifies a number of possible sources: (a) the size of the sample of performance; (b) the range of ability among the subjects; (c) ambiguity in instructions; (d) variation in conditions during measurement period; and (e) the amount of aid provided by instruments. Good experimental design can control for many of the sources of unreliability.

Validity

A second requirement is that the criterion be valid. As mentioned above, high reliability is desirable but not sufficient for criterion development. In fact, a statistical formula exists which is known as the correction for attenuation in the criterion variable (Cascio, 1987). This allows researchers to use criteria with relatively low reliability for validation purposes.

Why would researchers choose unreliable criteria over more reliable alternatives? The answer to this question deals with the concern for the validity of a criterion. The validity of a criterion, in this instance, is not the numerical relationship between the predictor and a criterion, but rather how relevant and representative the criterion is of the desired performance. A criterion could be highly representative of the designated performance but have low reliability.

The relevance of a criterion is rationally rather than mathematically derived (Astin, 1964). Thorndike (1949) describes the relevance of the criterion as the extent to which an index of success (criterion) is related to the "true" order of success in the given activity. If there were a way to ascertain the true order of success, as Thorndike describes it, then it would be possible to choose criteria which are completely relevant and representative of the work performance. Because there is some difficulty in choosing criteria which are completely relevant and representative, we must continue to use and be aware of the shortfalls of fallible measures of criteria.

How does a researcher determine the relevance of a criterion for a given job performance? One method which assists in determining both the relevance and the representativeness of a criterion is the use of a job analysis. A job analysis is a procedure useful in identifying the criteria or performance dimensions of a job (Muchinsky, 1983). When done correctly, a job analysis identifies all of the essential elements necessary for successful performance in the job. The use of job analysis has been recommended for many years (Nagle, 1953). Recently, job analysis methods have received renewed support and emphasis from the U.S. Supreme Court and are included in the Uniform Guidelines on Employee Selection Procedures (1978) as a necessary requirement when validation studies are conducted (Cascio, 1987).

## Practicality

Whether or not it is possible to identify a criterion that meets the requirements of reliability and validity, one additional concern should be addressed: the practicality of the criterion. The practicality of the criterion refers to the ease with which the measures can be gathered and the cost-to-benefit analysis of choosing one specific criterion over another. Practicality has also been defined as accessibility and cost (Wherry, 1957).

The significance of a practical criterion is a result of moving from the theoretical to the applied. After a criterion has been identified which meets the requirements of reliability and validity, it must be determined if the criterion to be used is practical.

Many researchers have successfully chosen reliable and valid criteria, only to discover that the cost of such criteria would outweigh the benefits of the measure or that the measure is inaccessible. An example of a measure that is inaccessible is a person's hospitalization records for psychiatric care. In most states, such records are covered under right to privacy and confidentiality acts which would disallow access of such records to most individuals.

Reliability, validity, and practicality are not the only requirements for criterion development (Blum & Naylor, 1968). They do serve as useful, if not minimal, standards in identifying potential criteria. More importantly, careful consideration of the reliability, validity, and practicality of a criterion leads to the selection of an effective criterion.

## Multiple Criteria Versus Composite Criterion

The definition of criterion suggests a single measure of job success. Yet, because of the dimensional characteristics of performance in a job (Dunnette, 1963; Ghiselli, 1956), no matter how simple or how involved, the use of a single criterion for success has given way to the utilization of several criteria. When more than one criterion is used, a decision must be made as to how the criteria should be treated.

The treatment of the criteria, specifically whether the criteria should be combined or left separate, has been an issue of debate since the beginning of the study of the criterion. Toops (1944), Nagle (1953), Guion (1961, 1987), and Dunnette (1963) are among the many authors who have confronted the problems associated with the multiple criteria versus the composite criterion debate.

Before the specifics of multiple versus composite criterion are discussed, an example will be used to illustrate the difference between the two approaches. An individual's college grade point average (GPA) is an often cited case which contrasts multiple criteria with a composite criterion approach (Muchinsky, 1983; Landy, 1989).

A traditional form of reporting a person's academic success in college is the student's GPA. An overall GPA consists of the person's performance reported by grades (A

= 4.0, B = 3.0, C = 2.0, and D = 1.0). The grades are weighted by the number of credits for each class and then summed across different subject matter in order to form an overall GPA. The overall GPA is considered to be a composite criterion for success in college.

A multiple criteria approach to a student's academic success in college would disregard the overall GPA. Instead, grades in each class or each subject matter area, would be examined in order to determine the student's specific strengths and weaknesses. A single indicator of success is not available when the multiple criteria method is used.

Reasonable evidence has been presented in the literature for both the multiple and composite criterion approach. It is apparent that each approach contributes something unique to the understanding of the criterion. The following sections summarize the arguments for each. The discussion includes a proposed means for resolving the multiple versus composite criterion controversy (Schmidt & Kaplan, 1971). Special attention is given to the methods for combining criteria to form a composite.

## Argument for Composite Criterion

Originally, psychologists emphasized the need to combine the criteria into a single score when more than one criterion was used. In early work on the use of criteria, Toops (1944) stated emphatically: "In all test work, and in making predictions generally, there must be . . . a unitary, general success score, or criterion score, for each person of the experimental group by whose aid the tests are constructed, combined, and validated" (p. 271). Proponents of the composite criterion approach are not concerned with whether several criteria should be combined but rather how the criteria should be combined. The use of criteria in test validation strategies early on, was heavily influenced by the necessity of a single score for validation purposes.

The composite criterion method focuses on the derivation of an overall measure of success or an estimable worth in economic terms of an individual's performance. The argument for a composite criterion focuses on the concern that when multiple criteria are used for selection purposes, the criteria must be collapsed in some form or manner for the decision to be made. The composite criterion method performs this combination objectively. The multiple criteria method is subjective and the method of combination is often times not communicated to others by the person making the judgment.

### Argument for Multiple Criteria

While much is to be gained in using a composite criterion, there are many concerns with its use in some situations. The most important of which is that not all criteria variables can be combined. Cattell (1957) points to a lesson learned early in algebra: "ten men and two bottles of beer cannot be added to give the same total as two men and ten bottles of beer" (p. 11). The same thing applies to combining several different criteria with low or negative correlations amongst each other. The criteria are representing different variables. To weight them into a composite leads to a result in scores so ambiguous as to be uninterpretable (Schmidt & Kaplan, 1971).

An additional concern with the use of an overall measure of success is the information which is lost when combining several criteria. The use of multiple criteria would enable the selection decision to be made by examining several different measures of performance. By keeping the criteria separate, it also assists the decision maker in understanding the components used in making the decision, those based on the applicant's specific strengths and weaknesses as opposed to a measure of overall worth.

Schmidt and Kaplan (1971) offer an excellent review of the composite versus multiple criteria controversy. In addition to citing evidence for both sides of the arguments, the authors discuss the assumptions underlying the two positions, evaluate the proposed arguments, and offer a resolution for the ongoing controversy.

As mentioned, a composite criterion is an attempt to determine an individual's overall success or ultimate worth to the organization. Advocates of a composite criterion see the criterion as an economic, rather than a behavioral, construct (Schmidt & Kaplan, 1971). The concept of an underlying economic nature is evident in many of the authors' definitions of a composite criterion.

Brogden and Taylor (1950) are the strongest proponents of using the composite criterion in an economic form. This becomes apparent in their statement "the criterion should measure the contribution of the individual to the overall efficiency of the organization" (p. 134). They suggest, if feasible, to turn the targeted performance behavior into monetary values.

The multiple criteria camp view the criterion as a representation of a behavioral or psychological construct. Advocates of multiple criteria simply can not justify combining different criteria into a composite. A composite of separate criterion elements would support the assumption of a general factor in job performance. The composite is used to determine a general measure of success. Ghiselli (1956) states with support from research (Bass, 1962; Ghiselli & Haire, 1960) that no such general factor exists, even in the simplest of tasks. Therefore, Ghiselli feels the criteria should be left separate instead of being combined.

One exception is when the criteria elements appear to be loading on a general factor, that is, when there are high intercorrelations among the criteria. Then it is "psychologically sensible" (Guion, 1965) to form a composite criterion. Guion notes that the resultant composite is in essence behaviorally homogeneous anyway.

Schmidt and Kaplan (1971) point out that the two groups also differ in the assumptions about the primary goals of the validation process itself:

> Those favoring the composite criterion are assuming, usually implicitly, that the validation process is initiated and carried out only for practical and economic reasons . . . None of those favoring composite criteria mention the attainment of increased understanding of the psychological and behavioral processes involved in various tasks as a goal of the validation process. By contrast, the advocates of multiple criteria view increased understanding as an important goal of the validation process, perhaps co-equal with the practical and economic goals. (p. 425)

In evaluation of the arguments, Schmidt and Kaplan (1971) focus primarily on what effect the differences have between using a composite criterion versus a multiple criteria method for validation purposes. When a multiple criteria approach is used for decision making, Schmidt and Kaplan point out that "the adequacy of the selection program depends heavily on the adequacy of the decision maker in assigning subjective criterion weights" (p. 425). This suggests that the decision maker must, in essence, be validated along with the selection device for consistency and reliability in the decision making process. When a composite criterion is used for assigning weights, the procedure is objective.

Finally, Schmidt and Kaplan (1971) address the different goals of the validation process: practical and economical for the composite method, and behavioral and psychological for the multiple criteria method. Their conclusion is that both composite criterion and multiple criteria are important to the validation process. Inherent in the article's resolution of the controversy of the composite criterion versus multiple criteria debate is that both methods should be used. Schmidt and Kaplan state that:

> From the point of view of the criterion end of the prediction equation, the implication of this fact is that he [sic] should, ideally, weight criterion elements, regardless of their intercorrelations, into a composite representing an economic construct in order to achieve his [sic] practical goals, and, at the same time, he [sic] should analyze the relationships between predictors and separate criterion elements in order to achieve his psychological goals. (p. 432)

By utilizing both methods together, validation research satisfies both the practical and theoretical goals of applied psychology.

Schmidt and Kaplan's (1971) resolution has not proven satisfactory to all. Smith (cited in Guion, 1987) offers an alternative solution to using both methods for validation purposes. She sees the choice of multiple criteria or composite criterion as contingent upon the researcher's purpose. Smith's conclusion is that you "should use a rifle for small targets, a cannon for big ones, and a shotgun if you can't aim very well" (Guion, 1987, p. 205). Smith's metaphor relates to the strengths of each approach. If the employer's goal is to select an all-around good employee, utilize a composite criterion which can serve as a measure of overall success. If the employer has very specific needs, the criteria should be left separate. The multiple criteria approach enables the employer to identify the candidate's precise strengths and weaknesses.

Whether the multiple criteria or composite criterion approach is chosen, both methods use different means to arrive at the same end. The goal of either method should be a meaningful, rational, and statistically sound development of the criteria used. Detailed attention to the use of criteria deserves careful consideration in the validation process.

## Methods for Weighting Criteria into a Composite

Any time multiple criteria are combined to form a composite, a decision must be made as to what method will be used for the weighting. As mentioned in the discussion above, the most relevant concern to those who favor the use of a composite is not whether the criteria should be combined, but how the criteria should be combined.

A number of methods have been developed for weighting a composite. In early work by Edgerton and Kolbe (1937) six methods were identified for weighting multiple criteria into a composite. These include:

1. Weighting the several variates proportionally to their importance as judged by "experts",
2. Weighting proportional to the reliability of the variate,
3. Weighting proportional to the average of the correlations of the one variate with the remaining variates,
4. Weighting the criterion variates so as to obtain a maximum correlation with several independent predictor variates,
5. Extracting by factor analysis those parts of the criterion variates which may be ascribed to the same factor,
6. Weighting the criterion variates so that the discrimination between all possible pairs of individuals in the sample will be as great as possible. (p. 183)

In an analysis of the various methods, Nagle (1953) points out that the weighting schemes are characterized by three different aspects of what are important considerations for a composite criterion: relevancy, reliability, and the presence of a general factor. The following critique is based on Nagle's work in examining the alternative methods for combining criteria into a composite.

The "expert" judgment weighting method (1) weights the criteria on their judged relevancy. Nagle (1953) describes the "expert" judgment method as "the only defensible method of combining criteria," and "since the relevancy must be judged, so the weights of the criteria must be assigned by judgments" (p. 281). A concern with the use of experts lies in the question of how the true value of the expert is determined.

The reliability weighting method (2) is an attempt to obtain the most reliable combination of criteria for the composite. Nagle (1953) argues that this method is a consequence of the emphasis on reliability as the single most important part of criterion development. This method increases the reliability of the criterion with complete disregard for the concerns of relevancy.

The maximum correlation weighting method (4) optimizes the relationship between an independent predictor and the criteria. Nagle (1953) points out that this method represents a sort of backwards conceptualization of the validation process. In a traditional validation model, the predictor or predictors' relationship to the criterion is optimized in order to minimize the number of predictors needed to successfully predict the criterion. The maximum correlation weighting method reverses the process and attempts to minimize the number of criteria, while optimizing the relationship of criteria to the predictor.

The average correlation weighting method (3), the factor analysis weighting method (5), and the discrimination weighting method (6) are each designed around the belief that a general underlying factor of performance exists. The average correlation weighting method tends to exaggerate the influence of any general factor that might exist and ignores the issue of relevancy.

The factor analysis weighting method attempts to separate by factor analysis the variance which is due to the same factor. The purpose of this approach is to include in the composite only one measure of each factor. Even after the analysis, Nagle (1953) points out that there still remains a concern with how the factors should be weighted.

The discrimination weighting method, favored by Edgerton and Kolbe (1936), assumes that each of the criterion is a measure of the same concept in a statistical sense. Problems arise in trying to address the criteria as a measure of the same thing in a conceptual sense. Nagle (1953) points out that this method may be useful when success is actually composed of one factor.

Additional weighting methods identified by Nagle (1953) include equal weighting of the separate criteria and the use of multiple cutoffs. These two methods are often utilized to combine the criteria for decision making purposes when a multiple criteria approach is used.

One method which was mentioned in the discussion of the composite versus multiple criteria is Brogden and Taylor's (1950) proposal for turning all of the criteria measures into dollar units. Using economically-based criteria would allow for meaningful combination of separate criteria, weighting values which are obvious, and equal unit scale differences. Brogden and Taylor's dollar unit method has received much support in theory. A major stumbling block to its use in practice is the inability to transform many measures of success into dollar values (Nagle, 1953).

## The Present Study

There are a number of articles which discuss the theoretical assumptions and rationale for choosing one method of developing criteria over another. Each time a validation study is done the researcher makes decisions on the inclusion (or exclusion) of certain criterion requirements. An additional concern is whether to combine or leave separate multiple criteria. These decisions about the criterion are of extreme importance and can consequently affect the outcome of the study.

Few studies exist (Helm, 1978; James & Ellison, 1973) which empirically examine the problem of the criterion. While many authors emphasize the significance of careful criterion development, relatively few studies expend effort on the development of the criterion or criteria used. Review of the research has revealed little recent work done into the investigation of the criterion. There still exists much controversy concerning the collection and treatment of criteria (James & Ellison, 1973; Wallace, 1965). Two specific areas will be addressed: the development of the criterion and the multiple criteria versus composite criterion controversy. A validation study on the selection of bank tellers was used to address these two issues empirically.

The present study was a concurrent design with a single predictor and multiple criteria. The criteria were chosen prior to the study. Each criterion was examined for effective criterion development. Detailed attention was given to the reliability, validity, and practicality of each criterion. The present study highlighted the significance of the reliability, validity, and practicality of the criterion as important concerns for the development of the criteria. The resolution proposed by Schmidt and

Kaplan (1971), who suggest using the multiple criteria approach in conjunction with the composite criterion approach, was enlisted in order to address the multiple criteria versus composite criterion controversy.

Several methods were used for weighting the composite. Because more than one method was used, this allowed the researcher to determine what differences exist between the selected weighting methods (James & Ellison, 1973). The use of several composites offered insight into what effect the different composites have on the relationship with the predictor and what effect the composites have on the decision making process. The use of a multiple criteria approach assisted in an understanding of the nature of the relationship of the separate criterion to the predictor.

# CHAPTER II

# METHOD

## Subjects

All subjects (N = 157) were incumbent bank tellers from institutions in New York, Texas, and Rhode Island. The sample was drawn from 35 branches of four banking organizations. The average age of the bank tellers was 30 years and average tenure was 29 months with a minimum of 3 months experience. Eighty-four percent of the sample were female. Seventy five percent of the sample were White, 15% were Black, 8% were Hispanic and 1% were Asian.

## Selection Instrument

The TELLER (Helm, 1989) was employed in the validation process. The TELLER's format is designed to replicate a typical work sample of handling money and making change that a bank teller would face in the daily performance on the job. The instrument is a paper and pencil representation of numerical reasoning problems of the type which a bank teller could encounter during a typical work day. The following is an example of the type of questions included in the TELLER:

A customer cashes a check for $25.00. You should give him or her: (a) one twenty dollar bill and one five dollar bill; (b) two twenty dollar bills; (c) two ten dollar bills; or (d) one ten dollar bill and two five dollar bills.

The TELLER was scored using a "1" for a correct response and a "0" for an incorrect response. The correct responses were added up for a total score. The higher the score, the better the performance on the TELLER. The TELLER is targeted specifically at determining a subject's ability to handle money and to make change quickly and accurately. The test is timed at 20 minutes and consists of 30 questions of increasing difficulty.

### Criteria

Five criteria were collected over a three month time frame after the test was administered. The criteria include:

1. Customer complaints--the customer's perception of the subject's inability to handle money and make change.

2. Supervisor's ratings--a measure of the supervisor's perception of the subject's performance on the ability to give change. The supervisor was asked to rate the subject on a scale of one to ten, with one depicting poor performance and ten depicting excellent performance. It is important to note that the supervisor was asked to rate the subjects only on their abilities to give change. The rating was not intended to be an overall rating of performance. In order to allow for ease of analysis, a

statistical inversion was performed on supervisor ratings which changed a rating of ten to one, one to ten, etc.

3. Cash overages--a direct performance-based measure of the subject's ability to give change represented in dollar amounts. It was a sum of the total cash overages for the 90-day period.

4. Cash shortages--a direct performance-based measure of the subject's ability to give change represented in dollar amounts. It was a sum of the total cash shortages for the 90-day period.

5. Number of overages and shortages--a measure of total number of mistakes made by the teller for the 90-day collection period.

Only subjects with full reported criteria data were included in the study. Each criterion was transformed into standard scores in order to equate the variances and to control for the influence of differential criterion variances (James & Ellison, 1973).

An important part of the development of the criterion is having a thorough understanding of the caliber of the criteria used in the validation process. Many of the standards used to evaluate criteria lack objective procedures with which the criteria can be measured. In place of objective measures, judgments and subjective estimations must be made about the quality of the criteria. This holds true for the five criteria used in this study.

Table 1 is a summary of how well the five criteria conform to the standards for criteria suggested by Blum and Naylor (1968). It is important to acknowledge that much of the information presented in the table is the author's judgments and it is possible that others could arrive at different conclusions. Objective evidence was used, whenever available, to support the judgments.

### Weighting

Several different weighting schemes were utilized. An equal weight composite was formed by using a weight of one for each criterion and then adding the criteria together. Equal unit weights assume each criterion is equally important to the total composite.

An average intercorrelation composite was formed by taking the average intercorrelation with the four other criteria. This average was then used as a weight for the criterion which was excluded from the averaging process. The average intercorrelation composite underscores any general factor on which the criteria might be loading. The largest weight is given to the criterion with the most in common (highest intercorrelation) with the other criteria.

A weighting method which is designed to emphasize the unique information that each criterion contributes to the composite was also used. This composite was obtained by dividing the average intercorrelation of each criterion into one. The inverse of the average intercorrelation, as

Table 1

Author's Judgment of Five Criterion Using Standards for Criteria Suggested by Blum and Naylor (1968)

| Standards | Criterion | | | | |
|---|---|---|---|---|---|
| | CO[a] | CS | NOS | SR | CC |
| 1. Reliable[b] | yes | yes | yes | yes | yes |
| 2. Realistic | yes | yes | yes | yes | yes |
| 3. Representative[b] | yes | yes | yes | no | no |
| 4. Related to other criteria[c] | yes | yes | no | no | no |
| 5. Acceptable to job analyst[b] | | not available | | | |
| 6. Acceptable to management | yes | yes | yes | no | no |
| 7. Consistent over situations[b] | yes | yes | yes | yes | yes |
| 8. Predictable[c] | no | no | no | yes | no |
| 9. Inexpensive[b] | yes | yes | yes | yes | yes |
| 10. Understandable | yes | yes | yes | yes | yes |
| 11. Measurable | yes | yes | yes | yes | yes |
| 12. Relevant[b] | yes | yes | yes | no | no |
| 13. Uncontaminated and bias-free | no | no | no | no | no |
| 14. Discriminatory | yes | yes | yes | yes | yes |

[a] CO = cash overs; CS = cash shorts; NOS = number of cash overs and cash shorts; SR = supervisor ratings; CC = customer complaints.

[b] Standards addressed in the discussion.

[c] Standards addressed in the results.

TELLER in order to examine the relationship of the TELLER to each composite. This validation process was used as a procedure for identifying any differential effects the separate composites might have on a predictor.

The TELLER was correlated (Pearson's _r_) with each criterion separately. The resulting correlations represented the validity coefficients for the TELLER and each criterion considered separately. This analysis assisted in determining how the use of different weights on the separate criterion could potentially affect the composite scores and ultimately the composite's relationship to the predictor.

The data from the composite criterion analysis was examined in order to determine the consequences of the use of criterion composites for decision making in a validation study. The data were also analyzed in order to determine if any conclusions could be made about the use of a composite criterion versus multiple criteria in the validation process.

# CHAPTER III

## RESULTS

The following section is an analysis of the impact of the criterion in a validation study. The results address four specific areas: (a) how the weights for the composites were derived, (b) the relationship of the composites to the TELLER, (c) the relationship of the composite scores to one another, and (d) the relationship of the TELLER to the criteria when left separate.

Four different weighting schemes were used to form four separate composites. The weights for the composites were derived in the following manner. The equal unit weighting method consisted of simply using one as the weight for each criteria.

The average intercorrelation weights were derived from the intercorrelations of the five criteria which is illustrated in Table 2. The value was determined by adding up the absolute value of the four intercorrelations in each row and dividing the resulting figure by four. As evidenced in Table 2, all of the intercorrelations were low (.01 to .27), with the exceptions of the correlation between cash overs and cash shorts (.94). This high correlation caused an increase in the average intercorrelation for these two criteria. Thus cash overs and cash shorts received the largest weights in this composite.

Table 2

Intercorrelation Matrix of the Five Criterion Used

|  | CO[a] | CS | NOS | SR | CC | Avg. Intercor. | Inverse Avg. Intercor. |
|---|---|---|---|---|---|---|---|
| CO | 1.0 | .94 | .24 | -.01 | -.03 | .31 | 3.23 |
| CS | .94 | 1.0 | .27 | -.03 | -.02 | .32 | 3.13 |
| NOS | .24 | .27 | 1.0 | .18 | -.03 | .18 | 5.56 |
| SR | -.01 | -.03 | .18 | 1.0 | .03 | .06 | 16.67 |
| CC | -.04 | -.02 | -.03 | .03 | 1.0 | .03 | 33.33 |

[a] CO = cash overs; CS = cash shorts; NOS = number of cash overs and cash shorts; SR = supervisor ratings; CC = customer complaints.

The weights for the inverse of the average intercorrelation are also presented in Table 2. The inverse values were derived by dividing the average intercorrelation into one. This mathematical manipulation gives customer complaints and supervisor ratings, the criteria with the lowest average intercorrelations, the largest weight. Cash overs and cash shorts are weighted the lowest in this composite.

The weights for the expert bid system were obtained from subject matter experts. Six supervisors of tellers, with an average of 15 years in the banking industry and seven years directly related to supervising tellers, were asked to rate the importance of each criterion. Each supervisor was given 100 points to be distributed among the criteria using the criterion's relevance to change making ability as the reason for the amount of weights assigned. The weights from the six participants were then added together and averaged to obtain the expert bid for each criterion. The experts judged the number of cash overs and cash shorts as the most relevant criterion, then cash shorts and cash shorts and cash overs. Supervisor ratings and customer complaints were weighted as the least relevant.

The formulas used for each composite is summarized in Table 3. The weighting schemes consisted of a composite weight multiplied by each criterion. Each criterion was then added up with the other weighted criteria to form a composite score. This composite score served as the composite criterion for each respective weighting scheme. The next segment examines how the various composites relate to a predictor in a validation model.

Table 4 presents the validity coefficients between the TELLER and the composites which were described in the previous paragraphs. There is a wide range of differences

Table 3

*Weights Used for Composite Scores*

1. Equal unit weighting method:

   Composite = (1)CO[a] + (1)CS + (1)NOS + (1)SR + (1)CC

2. Average intercorrelation weighting method:

   Composite = (.31CO + (.32)CS + (.18)NOS + (.06)SR + (.03)CC

3. Inverse of the average intercorrelation weighting method:

   Composite = (3.23)CO + (3.13)CS + (5.56)NOS + (16.67)SR + (33.33)CC

4. Expert bid weighting method:

   Composite = (.20)CC + (.21)CS + (.32)NOS + (.13)SR + (.14)CC

[a] CO = cash overs; CS = cash shorts; NOS = number of cash overs and cash shorts; SR = supervisor ratings; CC = customer complaints.

Table 4

*Validity Coefficients for Composite Scores Using the TELLER as a Predictor*

| Composite | TELLER |
|---|---|
| Equal Units | -.16* |
| Avg. Intercor. | -.03 |
| Inverse of Avg. Intercor. | -.22** |
| Expert Bid | -.13 |

*Note*. *N* = 157; $^{*}p < .05$; $^{**}p < .005$.

in the use of the four methods as composite criterion for the TELLER. The lower the score on each criterion composite, the better the performance. The higher the score on the TELLER, the better the performance. This accounts for the inverse relationship of the TELLER to the various composites.

Two coefficients achieved statistical significance: the equal units composite and the inverse of the average intercorrelation composite. The expert bid composite was not statistically significant. The average intercorrelation method was also not significant.

An opportunity to gain insight into the relationships described in Table 4 is to determine what commonalties are shared by each composite score. This was achieved by inspecting the percentage of shared variance among the five separate composite scores. The coefficients of determination are shown in Table 5. The inverse of the average intercorrelation shared the least variance with the other three composites: average intercorrelation (.08), expert bid (.28), and equal unit (.45). The expert bid composite shared a large amount of variance with the equal unit method (.94) and the average intercorrelation method (.85).

One additional analysis was included in order to develop an understanding of the effect of using a composite criterion and the relationship between the TELLER and the

Table 5

Coefficients of Determination ($r^2$) Among the Four Composite Scores

| | Equal Unit | Avg. Intercor. | Inverse Avg. Intercor. | Expert Bid |
|---|---|---|---|---|
| Equal Unit | 1.0 | .79 | .45 | .94 |
| Avg. Intercor. | | 1.0 | .08 | .85 |
| Inverse Avg. Intercor. | | | 1.0 | .28 |
| Expert Bid | | | | 1.0 |

criterion composites. This analysis examined the relationship between the TELLER and each criterion considered separately. The validity coefficients for the criteria using the TELLER as a predictor are presented in Table 6.

The TELLER exhibits varying levels of effectiveness in predicting the five criteria independently. The TELLER proves to be the most effective at predicting the criterion supervisor ratings (-.27). This correlation is also the only one to attain statistical significance. The other coefficients in Table 6, although not significant, indicate low (-.14 and -.11) to basically zero (.05 and .02) levels of relationship with the predictor.

Table 6

Validity Coefficients for Multiple Criteria Using the TELLER as a Predictor

| Criterion | TELLER |
|---|---|
| Cash overs | .02 |
| Cash shorts | .05 |
| Number of overs and shorts | -.14 |
| Supervisor ratings | -.27* |
| Customer complaints | -.11 |

Note. $N$ = 157, *$p$ < .005.

# CHAPTER IV

## DISCUSSION

Several interesting issues emerged from the results found in this study. The results suggested that the development of the criterion in the validation process deserves considerable attention. The results indicated that the different composites could lead to different conclusions about the effectiveness of the predictor. Evidence was also found which suggested that the effect of combining or leaving separate multiple criteria can impact the understanding of the relationship of the predictor with the criteria.

One conclusion demonstrated in this investigation is that this researcher's decisions made concerning the criteria consisted primarily of judgments and individual estimations. While many standards and guidelines are available, the proper selection and treatment of the criterion is the responsibility of the person or people evaluating the test. Even in the use of different weighting techniques, where relatively objective methods exist, the choice of a weighting scheme is left up to the discretion of the test evaluator. This study is evidence that such discretion can lead to many different conclusions about the quality of a chosen criterion and more importantly, as an end result, various conclusions about the effectiveness of the test being validated.

Even though the criteria for this study were chosen prior to the research, the criteria should still be scrutinized for effective development. An investigation into the development of the criteria used in this study revealed very little opportunity for infallible judgments about the quality of the criteria. In the methods section, the five criteria used in this study were briefly inspected for adherence to the many standards of criterion development. The following section is a more in-depth examination of the three basic requirements identified in the introduction: reliability, validity, and practicality.

## Reliability

Reliability of the criterion is concerned with a measure's stability and consistency. One method for determining a measure's reliability is through the use of a reliability coefficient. Four of the criteria, cash overs, cash shorts, number of cash overs and cash shorts, and customer complaints are performance criteria which do not lend themselves to the available statistical procedures for determining a reliability coefficient.

Since the performance measures are merely recorded behavior, the reliability of such criteria is considered to be one. Caution should be taken when using criteria for which reliability coefficients cannot be derived and are assumed to be perfect. As measurement theory suggests, it is unlikely that perfectly reliable measures exist. The four performance criteria are subject to the same sources of unreliability as criteria for which a coefficient can be determined.

There are many possible sources of unreliability for the four performance criteria used in this study. Possible sources include but are not limited to: (a) the consistency with which the actual performance was recorded, (b) the range of opportunity each subject had to demonstrate the desired performance, (c) the clarity of the performance desired, and (d) outside assistance in achieving the performance.

The fifth criterion, supervisor ratings, deserves special consideration because of the frequent use of ratings as criterion and due to the nature of ratings in general. Because of the heavy reliance upon ratings as a criterion for test validation purposes, much attention has been given to determining the reliability of ratings (Guion, 1987; Nagle, 1953).

Several methods exist for deriving a reliability coefficient when multiple raters or multiple ratings are used to evaluate an individual's ability. Unfortunately, for the purpose of this study, the requirements of the rating consisted of a single rater, assessing the subject's performance only one time. This type of performance measure, a single rating on a single occasion, does not lend itself to the available statistical procedures for deriving a reliability coefficient. So, like the four other criteria, supervisor ratings can only be assumed to have a reliability of one.

Whether ratings are assumed to be perfectly reliable as in this case, or when a coefficient can be derived, ratings are subject to many sources of unreliability. Nagle (1953) identifies several sources of unreliability which are specific to the use of ratings: (a) the competency of the raters, (b) the simplicity of the performance to be rated, (c) the degree to which the performance is easily observable, (d) the opportunity the raters have to observe the performance, and (e) the degree to which the rating task is defined.

The exact reliability of cash overs, cash shorts, number of cash overs and cash shorts, supervisor ratings, and customer complaints as used in this study is unknown. An important aspect of using objective criterion measures as criteria, or as predictors, is the conscientious administration and monitoring of the collection of the measures. Though no reliability coefficient can be determined for the five criteria used in this study, careful consideration of the sources of unreliability discussed above should help guard against unnecessarily low reliability in the criterion.

Validity

Validity can be defined as the theoretical correlation between the performance criterion and the ultimate criterion. The validity of a criterion focuses on how representative and relevant the criterion is of the desired performance. Similar to a criterion for which no reliability coefficient can be determined, the validity of a criterion cannot be derived mathematically. Estimations of the validity of a chosen criterion are made through the use of judgments and interpretations.

The concern about validity for this study centers on how relevant and representative cash overs, cash shorts, number of cash overs and cash shorts, supervisor ratings and customer complaints are with respect to an individual's ability to handle money and make change effectively. One method for insuring valid criteria is through the use of a job analysis. A job analysis should be performed prior to the choice of the criteria. Job analysis methods identify measures which are critical to effective performance in the job. Because the criteria were already chosen for this study, a complete job analysis was not feasible nor was one available.

One facet of job analysis was utilized as part of the expert bid method of weighting a composite. Expert judgments proved to be helpful in determining the relevance of the criteria. Bank teller supervisors were asked to assign weights to the five criteria depending on each criterion's relevance to a teller's ability to handle money and make change effectively. Using a total scale of 100 points, the experts perceived numbers of cash overs and shorts as the most relevant (32 points), then cash shorts (21 points) and cash overs (20 points). Supervisor ratings (13 points) and customer complaints (14 points) were considered the least relevant. The teller supervisors did not have the opportunity to include additional criteria which they may have perceived as more relevant.

Valid criterion selection necessitates a thorough understanding of what performance is going to be measured. The use of an ultimate criterion or conceptual criterion is helpful in generating many of the possible ways in which the criterion can be conceptualized. An important step in criterion validity is moving from the ultimate criterion to the performance criterion. The more effectively one moves from the ultimate criterion to a criterion which can actually be measured, the higher the criterion's validity will be.

### Practicality

The importance of moving from an ultimate criterion to a performance criterion successfully has significance for an additional concern of the criterion--practicality. The criterion's practicality, the ability to gather the criterion data easily and inexpensively, is an important consideration in the selection and use of most criteria. Often criteria can be conceptualized that would, for practical purposes, be too difficult or too expensive to collect.

Cash overs, cash shorts, number of cash overs and cash shorts, supervisor's ratings, and customer complaints are all judged as being very practical criteria for data collection purposes. With the exception of supervisor's ratings, the four other criteria are performance variables which are included in a bank teller's normal record keeping procedures. Supervisor's ratings required a minimal amount of effort by the teller's supervisor.

The practicality of a criterion is often invoked as a major factor in the use of a particular criterion. Availability and expediency often characterize criteria used in validation studies. The importance of practicality in the criterion is based on the discretion of the individual choosing the measures. No linear relationship exists between the practicality of the criterion and the judged quality of the criterion. There are no specific rules that would suggest that the more impractical the criterion, the worse the criterion is, or that the more practical the criterion, the better the criterion is.

The second part of this study investigated the use of multiple criteria versus a composite criterion in a validation procedure. Because of the investigatory nature of this project, both the multiple criteria approach and several composite criterion methods were employed.

## Composite Criterion

This was not a validation study. It was an investigation into the use of the criterion. The use of the predictor was designed to assist in understanding the effect of differential treatment of the criteria used for this study. An effective and efficient manner of examining the criteria was through the use of a traditional validation model in which predictor-criterion relationships are the primary focus.

An understanding of the TELLER was secondary to the purpose of this investigation. Insight into the predictive ability of the TELLER was gained as a consequence of the examination of the criteria. The primary function of the TELLER was to serve as a tool for interpreting the effects of the manipulation of multiple criteria.

The results indicated certain criteria or criterion composites show a higher correlation with the predictor when compared to others. It is important to note that the predictor's relationship to the criterion is not the only measure of the quality of the criterion used. The predictive efficiency of the TELLER for any given criterion is not the best measure of the goodness of the criterion. Judgment about the development and the quality of the criterion should consider more than just the discriminative ability of the criterion. Many of the important requirements of the criterion were addressed in the above discussion.

The next segment examines two important questions regarding the use of criterion composites. Do the different weighting methods generate composites which differ from one another? If so, what effect do the different composites have on determining the validity of the test? First, attention is focused on whether the various strategies for combining the criteria actually result in composites that are different from one another.

Evidence is provided which suggests that the criterion composites do represent different concepts. In seeking to understand if the composites represent different concepts, information was found which offers insight into why the composites elicit such disparate validity coefficients when correlated with the TELLER.

Two basic notions can be used to account for the differences found when using criterion composites: one is conceptual, the other is statistical. The premise of using a composite is to combine multiple criteria in an objective fashion which allows for uncomplicated decision making.

Composite criterion methods are conceived to emphasize a certain rationale for combining the criteria.

The four composites developed for this study each use a different justification for the combination of the criteria. The equal units composite assumes each criterion is equally important. The average intercorrelation composite extracts any general factor that the criteria might share. The inverse of the average intercorrelation composite draws on the unique contributions of each criterion. The expert bid composite is based on the order of relevance of each criterion.

The way each composite is conceptualized influences the derivation of the weights used for combining the criteria. It has been suggested that the four composites are conceptually different. Possibly, the reason the composites prove to be different, at least in empirical terms, is the different statistical procedures used to determine them.

An illustration of the effect that the different statistical procedures have on the outcome of the composite is the use of the average intercorrelation method when compared to the inverse of the average intercorrelation method. Conceptually, these two methods have exact opposite justifications for weighting a composite. The average intercorrelation method gives the greatest weight to the criterion which contributes the most to a general factor.

The inverse of the average intercorrelation gives the largest weight to the criterion which contributes the most uniqueness to the composite, or the criterion measure which contributes the least to a general factor. The large discrepancy between the validity coefficients of the two methods when correlated with the TELLER (average intercorrelation .03 and inverse of the average intercorrelation -.22) suggests the composites are getting at the separate conceptual facets upon which they are based.

Up to this point, evidence that the various composites actually generate criterion composites which are different from one another has focused on the relationship of the composite to a predictor outside the composite. Additional evidence of the difference is the obviously distinctive weighting schemes used for each composite. A third aspect which was helpful in determining if the weighting methods were truly distinguishable from one another, is an examination of the proportion of accountable variation ($\underline{r}^2$) found among the four composite scores. This analysis looks at the relationship of the composites with one another.

In a predictor-criterion model the square of $\underline{r}$ indicates the percentage of criterion variance accounted for, given a knowledge of the predictor (Cascio, 1987). The statistic $\underline{r}^2$ is known as the coefficient of determination. Use of the coefficient of determination allows the researcher to have an idea of how effective the test is at predicting the criterion. For purposes of this discussion, $\underline{r}^2$ was used to determine the similarity or difference of the four criterion composites. If $\underline{r}^2$ is large the two composites are similar. That is, both composites are accounting for the same variance. If $\underline{r}^2$ is small, the two composites are different and the composites are accounting for different variance.

Evidence of the differential effects of the various weighting techniques would be found if the composites that are the most similar to one another have composite scores which have a large $\underline{r}^2$. Conversely, the composites which are the least similar to one another would have composite scores which have a small $\underline{r}^2$. An analysis of the coefficients of determination among the four composite scores offered support for this supposition.

The average correlation composite and the inverse of the average intercorrelation composite, which are considered dissimilar, also indicated the smallest $\underline{r}^2$ (.08) among the four composite scores. Such a small $\underline{r}^2$ is evidence that the two composites are unique to one another. The methods from which the two composites were derived result in criterion composites which were accounting for different portions of variance.

The expert bid method and the equal unit method are composites that use similar weights and have comparable validity coefficients (-.13 for the expert bid composite and -.16 for the equal unite composite). These two composites exhibited the largest $\underline{r}^2$ (.94). Such a large $\underline{r}^2$ is evidence that the two composites are similar to one another and are accounting for the same variation. The expert bid composite and the equal unit composite are so similar that one could be used in the place of the other with little or no information lost.

The average intercorrelation method and the inverse of the average intercorrelation method are conceptualized in opposite terms. They exhibited the smallest coefficient of determination. The two composites are the most different with respect to the weights used for combining the criteria, and they had the largest discrepancy with respect to their validity coefficients when correlated with the TELLER.

The expert bid weighting method and the equal unit method used similar weights to combine the criteria. The composites are predicted with comparable efficiency by the TELLER and the composite scores exhibited the largest coefficient of determination. These findings serve as empirical evidence of the capability that the various weighting methods possess in generating criterion composites that are different from one another, both conceptually and statistically.

Evidence for the ability of the various weighting schemes to generate different types of criterion composites has implications for the choice of the composite criterion used. Knowledge of the conceptual and statistical nature of the various weighting methods is beneficial to the effective use of a particular composite criterion in the validation process. Separate criterion composites have the potential to elicit different conclusions with regard to effectiveness of the test. The following discussion focuses on the effect of using several composites for determining the validity of the TELLER.

A notable consequence of using several criterion composites for a single validation study is the possibility that the composites could lead to different conclusions about the validity of the test. This study exhibited a wide range of results from the use of the four composites. The predictive efficiency of the TELLER for the four composites ranged from -.03 to -.22. Any one of these composites considered in isolation would lead to very different conclusions about the predictive ability of the test. The following section examines two of the composites and proposes some plausible conclusions from their use in a validation procedure.

The different correlations from the individual composites and the TELLER, which are also considered validity coefficients, have many possible implications. One implication is the possibility of a false negative. That is, a test evaluator using one composite for validation purposes might conclude falsely that the test has no predictive value.

For example, if the average intercorrelation composite was employed as the only composite, the test would be considered invalid. When in reality, the weak relationship is a product of the use of the average intercorrelation composite, as opposed to the potential predictive ability of the test.

Another possible implication is the use of biased reporting in favor of a specific point of view, especially in a situation in which the test developer performs the validation study. This person might be prone, for certain reasons (marketing is often a key one) to report only the validity coefficient which offers support for the predictive efficiency of the test.

In the case of the TELLER, the test evaluator would report only the correlation between the TELLER and the inverse of the average intercorrelation composite. Thus, implying that the TELLER has a (relatively) effective and significant predictive ability with a given criterion. When actually the significant relationship exists only as a consequence of the use of a specific composite. There is nothing particularly deceptive about reporting the strongest relationship as the validity of the test.

An important aspect of evaluating a test is to determine the test's strengths and its weaknesses. A potential problem occurs when the test evaluator fails to explore the full potential of the test. It is valuable to investigate both the strong and weak relationships found between the test and the various criteria, or criterion used. Then the test developer or evaluator gains insight not only into what the test is most effective at predicting, but also into what the test is least effective at predicting.

Information has been presented about the nature of the composites used in this study. It was shown that various weighting methods generate different forms of composites. The TELLER proved to have different levels of predictability with regard to the four criterion composites. The relationship of the TELLER to the composites was derived from both conceptual and statistical formulations. It was also shown that the use of different composites can lead to different conclusions about the validity of the test.

## Multiple Criteria

It would be easy to stop at this point and recommend that there is a thorough understanding of the criteria used for this study. An argument for the use of a multiple criteria approach suggests the criteria should be examined further. A composite is a rationally justified, statistical combination of multiple criteria. The composites obtained are useful in the decision making process. The knowledge of the relationship of the predictor to each multiple criteria considered separately adds information that could not be obtained from the use of the composite criterion approach.

Many benefits are derived from examining the criteria separately in a validation process. Because the criteria are not combined, the relationship of the predictor to each criterion can be examined in detail. The multiple criteria approach assists in a psychological and behavioral understanding of the relationship between the predictor and criterion.

As an example, the TELLER appears to predict the criterion supervisor ratings the most effectively (-.27). What this correlation represents would not even be considered by those who use a composite criterion for the validation process. Using the multiple criteria approach enables the researcher to gain an understanding into why the relationship between the TELLER and supervisor ratings is the strongest of the five criteria used.

The high correlation between the TELLER and supervisor ratings is especially interesting for two reasons. One, the TELLER was designed to predict a fairly specific behavior, a person's ability to handle money and make change effectively. Two, the nature of ratings do not lend themselves to predicting specific behaviors.

A possible explanation for the relationship stems from an understanding of what the predictor and the criterion represent. In early work with the TELLER, it was found that the test correlated with a proven measure of general mental abilities; the Wonderlic at .66 (Helm, 1990). This correlation suggests that the TELLER has some similarities to a general mental abilities test. The study using the TELLER and the Wonderlic found that people who performed well on the Wonderlic also tended to perform well on the TELLER. It is tenable that the TELLER is a better predictor of a measure of general mental abilities than the specific behavior of change making ability.

Additionally, it is possible that the criterion supervisor ratings is better representative of the supervisor's judgment of an individual's general mental abilities rather than a rating of a person's change making abilities. Teller supervisors are probably not as skilled at identifying tellers who are effective change makers as they are at identifying people who would be considered intelligent. Therefore, the criterion supervisor ratings may actually be the supervisor's perception of the teller's general mental abilities.

The TELLER's relationship to the Wonderlic suggests that the TELLER has the capability to differentiate individuals who are high in general mental abilities from those who are low in general mental abilities. That is, the TELLER separates the intelligent people from the less intelligent people. When asked to rate tellers on the ability to make change, the supervisors may actually be judging individuals on the basis of their general mental abilities, or how intelligent they are.

Although this explanation has not been investigated extensively, it does serve as a logical means for understanding the correlation found between the TELLER and supervisor ratings. More importantly, it illustrates the value of using a multiple criteria approach in conjunction with a composite criterion approach. Such an explanation would not be available if only the composite criterion approach was employed for validation purposes.

Knowledge of the relationship of the TELLER to the individual criteria also proved to be helpful in understanding the predictive efficiency of the TELLER to the four composites. Two of the composites are examined: the inverse of the average intercorrelation method and the average intercorrelation method. The inverse of the average intercorrelation method showed the highest relationship to the TELLER. The average intercorrelation method showed the lowest relationship to the TELLER.

Upon examining the multiple criteria validity coefficients, it was discovered that the inverse of the average intercorrelation composite weighted most heavily the criteria which were related the highest to the predictor. Customer complaints, supervisor ratings, and number of cash overs and cash shorts showed the strongest individual relationships to the TELLER and were weighted the most heavily. The criteria related the poorest to the predictor were weighted the lowest. Cash overs and cash shorts showed the weakest individual relationships to the TELLER and were weighted the least. The inverse of the average intercorrelation method produced a statistical weighting scheme which emphasized the stronger individual correlations and de-emphasized the weaker individual correlations in the formation of the composite.

The average intercorrelation method generated a composite that when correlated with the predictor produced a nonsignificant, relatively zero validity coefficient. The relationship between the average intercorrelation composite and the TELLER became clearer by using the multiple criteria approach and examining the weights given to the individual criteria.

The largest weights were given to the criteria with the poorest correlation with the predictor. Cash shorts and cash overs showed the weakest individual relationships to the TELLER and were weighted the most heavily. The smallest weights were given to the criteria with the strongest correlation with the predictor. Customer complaints, supervisor ratings, and number of cash overs and cash shorts showed the strongest individual relationships to the TELLER and were weighted the least heavily. The average intercorrelation method produced a statistical weighting scheme which de-emphasized the stronger individual correlations and emphasized the weaker individual correlations in the formation of the composite.

### Conclusion

The development and treatment of the criterion has important ramifications with respect to the use of the criterion in the validation process. Every time a test is validated many decisions must be made about the criterion. The inclusion or exclusion of certain criteria and whether to combine or leave separate multiple criteria are only two of the many concerns which must be addressed in choosing or evaluating criterion measures. The decisions made by the test developer about criterion measures can have a significant consequence on the outcome of the study.

Since so many validation studies have been performed through the years, it would seem that there would be an abundance of documentation which addresses the impact of the criterion in validation studies. A considerable amount of literature on the impact of the criterion does exist, but the literature focuses primarily on theoretical implications. A dilemma encountered in the use of criteria for real world applications is the deficiency of suggestions and recommendations for handling concerns about the criterion at the practitioner's level.

Without a well-grounded theoretical understanding of the criterion, little advancement can be made in terms of the criterion's practical applications. Yet, only a small amount of progress has been made about an understanding of the impact of the criterion on the validation process since the original investigation into the predictor criterion relationship. The knowledge that is available about the criterion is difficult to apply in the actual validation process.

Criterion research should now focus on how to present the available knowledge in terms that would be meaningful and applicable to the practitioner. This study served as an illustration of the problem of applying the theoretical suggestions and guidelines of the development and treatment of the criterion in an empirical study. After much effort was put into understanding the necessity of careful consideration of the criterion, the researcher had to resort to judgment and estimations about the use of the criterion in a practical setting. Further research is suggested in defining techniques for the development and treatment of the criterion.

# REFERENCES

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

Bass, B. M. (1962). Further evidence on the dynamic characteristics of criteria. Personnel Psychology, 15, 93-98.

Blum, M. L., & Naylor, J. C. (1968). Industrial psychology: Its theoretical and social foundations (Rev. ed.). New York: Harper & Row.

Brogden, H. E., & Taylor, E. K. (1950). The dollar criterion--applying the cost accounting concept to criterion construction. Personnel Psychology, 3, 133-167.

Cascio, W. F. (1987). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Cattell, R. B. (1957). Personality and motivation structure and measurement. New York: Harcourt, Brace, & World.

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

Edgerton, H. A., & Kolbe, L. E. (1936). The method of minimum variation for the combination of criteria. Psychometrika, 1, 183-187.

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

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco, CA: W. H. Freeman and Company.

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

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

Guion, R. M. (1987). Changing views for personnel research. Personnel Psychology, 40(2), 199-213.

Helm, K. G. (1989). TELLER. Unpublished test, Dallas, Texas.

Helm, K. G. (1990). [Comparison of the TELLER to the Wonderlic]. Unpublished raw data.

Helm, K. G. (1978). The development of a criterion composite for scales success. Unpublished dissertation, Texas Christian University.

James, L. R., & Ellison, R. L. (1973). Criterion composites for scientific creativity. Personnel Psychology, 26, 147-161.

Jenkins, J. G. (1946). Validity for what? Journal of Consulting Psychology, 10, 93-98.

Krug, R. E. (1961). Personnel selection. In B. von & H. Gilmer (Eds.), Industrial psychology. New York: McGraw-Hill.

Landy, F. J. (1989). Psychology of work behavior (4th ed.). Pacific Grove, CA: Brooks/Cole Publishing Co.

Muchinsky, P. M. (1983). Psychology applied to work: An introduction to industrial and organizational psychology. Homewood, IL: The Dorsey Press.

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

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

Thorndike, R. L. (1949). Personnel selection: Test and measurement techniques. New York: Wiley.

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

Uniform Guidelines on Employee Selection Procedures. (1978). Federal Register, 43(166), 38290-38309.

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

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

Wherry, R. J. (1957). The past and future of criterion evaluation. Personnel Psychology, 10, 1-5.